An Efficient Framework for Enhancing Discriminative Models via Diffusion Techniques

Chunxiao Li[1]*, Xiaoxiao Wang[2]*, Boming Miao[1], Chuanlong Xie[1], Zizhe Wang[3], Yao Zhu[3]†

[1]Beijing Normal University, Beijing, China
[2]University of Chinese Academy of Sciences, Beijing, China
[3]Tsinghua University, Beijing, China
chunxiaoli@mail.bnu.edu.cn, wangxiaoxiao23@mails.ucas.ac.cn, bomingmiao@mail.bnu.edu.cn, clxie@bnu.edu.cn, wangzz@act.buaa.edu.cn, ee_zhuy@zju.edu.cn

Abstract

Image classification serves as the cornerstone of computer vision, traditionally achieved through discriminative models based on deep neural networks. Recent advancements have introduced classification methods derived from generative models, which offer the advantage of zero-shot classification. However, these methods suffer from two main drawbacks: high computational overhead and inferior performance compared to discriminative models. Inspired by the coordinated cognitive processes of rapid-slow pathway interactions in the human brain during visual signal recognition, we propose the Diffusion-Based Discriminative Model Enhancement Framework (DBMEF). This framework seamlessly integrates discriminative and generative models in a training-free manner, leveraging discriminative models for initial predictions and endowing deep neural networks with rethinking capabilities via diffusion models. Consequently, DBMEF can effectively enhance the classification accuracy and generalization capability of discriminative models in a plug-and-play manner. We have conducted extensive experiments across 17 prevalent deep model architectures with different training methods, including both CNN-based models such as ResNet and Transformer-based models like ViT, to demonstrate the effectiveness of the proposed DBMEF. Specifically, the framework yields a 1.51% performance improvement for ResNet-50 on the ImageNet dataset and 3.02% on the ImageNet-A dataset. In conclusion, our research introduces a novel paradigm for image classification, demonstrating stable improvements across different datasets and neural networks.

Code — https://github.com/ChunXiaostudy/DBMEF
Extended version — http://arxiv.org/abs/2412.09063

1 Introduction

Image classification stands as a foundational problem in computer vision, typically tackled through either discriminative model-based (LeCun et al. 1998; He et al. 2016; Touvron et al. 2021) or generative model-based (Rish et al. 2001; Cheng and Greiner 2013; Han, Zheng, and Zhou 2022) approaches: the former directly models the posterior probability $p(y|\boldsymbol{x})$ for image classification, while the latter first learns the joint distribution $p(\boldsymbol{x}, y)$ of the data and then leverages Bayes' theorem to convert it to $p(y|\boldsymbol{x})$.

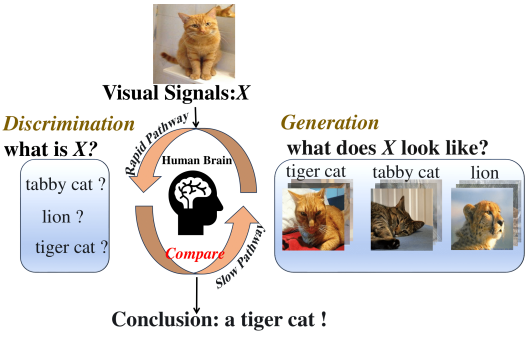

Figure 1: The process of the human brain handling visual signals is a dynamic interactive procedure. The rapid pathway transmits visual signals to the higher cortex to complete overall recognition, then assists the slow pathway in completing a "guess-verify-guess-verify" cognitive linkage. The rapid pathway of the brain can be regarded as a discriminative process, proposing several possible guesses regarding the visual signals. The slow pathway's verification of these guesses can be approximately considered as the reclassification process of a generative model under given conditions.

In research on applying discriminative models for image classification, numerous meticulously designed network architectures have been proposed, such as VGG (Simonyan and Zisserman 2014), ResNet (He et al. 2016), ViT (Dosovitskiy et al. 2020), DeiT (Touvron et al. 2021), and others. A well-trained ViT model can even achieve an astonishing accuracy of up to 88.59% on ImageNet. However, this approach provides predictions in a single-step process, which means it lacks the capability to re-evaluate and refine uncertain predictions as humans do. This limitation potentially hinders further performance improvements for the model.

The task of image classification based on generative models is challenging because it requires modeling the conditional likelihood of each label for the images. Early work has explored the potential of generative models based on

*These authors contributed equally.
†Corresponding author: Yao Zhu.

energy-based models (Zhao, Jacobsen, and Grathwohl 2020) or score-based models (Song et al. 2020; Zimmermann et al. 2021; Yoon, Hwang, and Lee 2021) for image classification tasks; however, due to the greater demand for data and resources, such methods are primarily applied to smaller datasets like CIFAR-10, and there remains a significant gap in their widespread application. The recent emergence of large-scale text-to-image diffusion models (Ho, Jain, and Abbeel 2020; Rombach et al. 2022) has greatly enhanced text-based image generation capabilities, inspiring researchers (Clark and Jaini 2024; Li et al. 2023a; Chen et al. 2023) to use density estimation from diffusion models for zero-shot classification without additional training. To be specific, they perform conditional denoising across various categories on the test image, selecting the category that yields the best denoising outcome as the image's label. Although these methods can be applied to higher-resolution images, they still suffer from two drawbacks: their performance significantly lags behind discriminative models, and the application of diffusion models in classification is time-intensive. Even with accelerated sampling methods, classifying a single image from ImageNet of 512×512 pixels may take up to two hours to complete.

In response to the shortcomings of existing image classification approaches, we pose a question: Can we leverage the strengths of both discriminative and generative models to better accomplish the task of image recognition? This paper answers this question by drawing inspiration from the human brain's process of image recognition. As shown in Fig. 1, when the brain receives visual signals, the image information of objects is rapidly transmitted to the higher cortex through a rapid pathway, where a guess is made (Sillito, Cudeiro, and Jones 2006). The result of the guess is then cross-verified with new inputs through feedback connections. The rapid pathway recognizes the object as a whole, and its results assist the slow pathway in recognizing local information of the object. Through such a repetitive process, the object is recognized. When faced with uncertain visual signals, the brain undergoes an alternation of information uploading and downloading, continuously engaging in a "guess-verify-guess-verify" cycle(Sillito, Cudeiro, and Jones 2006; Chen et al. 2014) until a definitive result is obtained. Therefore, in this paper, inspired by the brain's signal recognition process that intricately combines the rapid and slow pathways for synergistic effect, we propose the Diffusion-Based Discriminative Model Enhancement Framework (DBMEF). This approach utilizes diffusion models to endow discriminative models with the ability to re-evaluate uncertain predictions. Specifically, within this framework, discriminative models first make a preliminary prediction on the test input, akin to the brain's rapid pathway. If the uncertainty of the prediction is low, it is considered the final output. However, if the prediction's uncertainty is high, the test sample is fed into the diffusion model. Through the generative model's powerful image understanding capabilities, the test sample undergoes a rethinking process, mirroring the brain's slow pathway. This methodology not only simulates the human brain's process of dealing with visual signals but also significantly enhances the classification accuracy of various discriminative models (e.g., ResNet, VGG, DeiT, ViT).

The contribution can be summarized as follows:

- We propose a plug-and-play framework called Diffusion-Based Discriminative Model Enhancement Framework (DBMEF), which is inspired by the coordinated cognitive processes of rapid-slow pathway interactions in the human brain during visual signal recognition.
- Experimental results have demonstrated that DBMEF can significantly enhance the classification accuracy and generalization capability of deep neural networks. The framework notably increases the performance of ResNet-50 by 1.51% on the ImageNet dataset, and by 3.02% on the ImageNet-A dataset.
- We provide an extensive and detailed ablation study on the proposed framework, uncovering that an overly intensified negative control factor λ adversely impacts the efficacy of the framework and our proposed method outperforms other diffusion-classifiers with significantly fewer time-step samplings.

2 Related Work

Diffusion models(Ho, Jain, and Abbeel 2020; Song, Meng, and Ermon 2020; Rombach et al. 2022; Karras et al. 2022) have garnered significant interest for their exceptional generative abilities, especially in image generation, where they have outperformed GAN models(Goodfellow et al. 2020) in producing high-quality images. In 2021, OpenAI introduced Classifier Guidance Diffusion(Dhariwal and Nichol 2021), a technique that facilitates gradient adjustment of images in the generation phase by diffusion models, thus permitting conditional generation based on specified categories. Following this, in 2022, Google unveiled Classifier-Free Guidance Diffusion(Ho and Salimans 2022), a method that eliminates the need for training a separate explicit classifier and instead integrates conditional guidance within the training phase, resulting in enhanced generative capabilities. The latent diffusion model(Rombach et al. 2022) maps images to a latent space for both the noising and denoising processes, significantly reducing training expenses and expanding the potential applications and generative capabilities of diffusion models. This approach has laid a crucial foundation for the development of Stable Diffusion, showcasing the versatility and adaptability of diffusion models across various domains (Xu et al. 2023; Tang et al. 2024; Zhao et al. 2023).

In this context, utilizing diffusion models for image classification introduces a novel approach. SBGC(Zimmermann et al. 2021) employs score-based generative models to estimate the conditional likelihood $P(x|y)$, which is then utilized for image classification(Li et al. 2023a; Clark and Jaini 2024; Chen et al. 2023) and further enhanced by integrating diffusion models into these task. RDC (Chen et al. 2023) leverages diffusion models to evaluate its adversarial robustness in comparison to traditional discriminative classifiers. Li et al. (Li et al. 2023a) approximates the log probability $\log p_\Theta(x|c)$ through EBLO and identifies the optimal condition c by analyzing noise predictions in the denoising phase, demonstrating notable zero-shot classification ca-

pabilities. Furthermore, a diffusion-based classifier applied to the DiT-XL/2 model (Peebles and Xie 2023), trained on the ImageNet dataset for supervised classification, achieved a notable accuracy of 79.1% on the ImageNet validation set. Nevertheless, this performance level does not yet meet the benchmarks established by leading contemporary supervised deep networks on ImageNet, such as ViT-Base, DeiT-Small, and DeiT-Base, all of which can experience significant improvements in classification performance through our methodology. In (Clark and Jaini 2024), the authors similarly utilize a diffusion model as a classifier. The distinction between these two papers lies that (Clark and Jaini 2024) introduces learnable time-step sampling weights and shifts the loss function focus from measuring the discrepancy between predicted noise and added noise, to quantifying the difference between denoised image and original image. Despite accelerating the sampling process(Li et al. 2023a; Clark and Jaini 2024), the inference duration for a single image from ImageNet can extend to two hours on a GeForce RTX 4090 GPU, which significantly limits the practical application of diffusion models in real-world scenarios despite their innovative use.

In our work, we challenge the conventional belief that "discriminative models typically perform better in image classification tasks" by synergizing diffusion models with discriminative models through confidence protector and posterior probability adjustment. This research significantly enhances the accuracy of discriminative models across various architectures, while reducing the inference time to merely 1% of that of existing diffusion-based classifiers.

3 Framework

In this section, we introduce the Diffusion-Based Discriminative Model Enhancement Framework (DBMEF). We first provide an overview of the diffusion model in Sec. 3.1, and we introduce the overall process of DBMEF in Sec. 3.2. In Sec. 3.3, we introduce the confidence protector and diffusion model classifier. In Sec. 3.4, we present two effective methods to further improve the performance. The Pytorch-like pseudo algorithm of DBMEF is given in Appendix. A.

3.1 Preliminary Knowledge on Diffusion Models

The diffusion model (Ho, Jain, and Abbeel 2020; Song, Meng, and Ermon 2020; Rombach et al. 2022) consists of two processes: forward diffusion with adding noise and inverse diffusion with denoising. In the forward process, a small Gaussian noise is gradually added to a true data distribution $\boldsymbol{x}_0 \sim q(\boldsymbol{x})$. The distribution eventually converges to an isotropic Gaussian distribution. The mean and variance of the added noise are determined by parameter β_t, then we get:

$$\begin{aligned} q\left(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1}\right) &= \mathcal{N}\left(\boldsymbol{x}_t; \sqrt{1-\beta_t}\boldsymbol{x}_{t-1}, \beta_t I\right) \\ &\text{s.t. } 0 < \beta_t < 1. \end{aligned} \tag{1}$$

In the reverse process, the original data is recovered from Gaussian noise $\boldsymbol{x}_T \sim \mathcal{N}(0, \mathbf{I})$. In (Ho, Jain, and Abbeel 2020), the posterior probability P_θ is approximated using a time-conditioned deep model, yielding:

$$\begin{aligned} P_\theta\left(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t\right) &= \mathcal{N}\left(\mu_\theta\left(\boldsymbol{x}_t, t\right), \Sigma_\theta\left(\boldsymbol{x}_t, t\right)\right), \\ P_\theta(\boldsymbol{x}_{0:T}) &= p(\boldsymbol{x}_T) \prod_{t=1}^{T} P_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t). \end{aligned} \tag{2}$$

The conditional diffusion model (Ho and Salimans 2022) utilize the ELBO as an approximation to the class-conditional log-likelihood as:

$$\log P_\theta(\boldsymbol{x} \mid y) \geq \mathbb{E}_{q(\boldsymbol{x}_{0:T})}\left[\log \frac{P_\theta(\boldsymbol{x}_{0:T}, y)}{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)}\right]. \tag{3}$$

By reparameterizing the right half of Eq. (3), similar to (Ho, Jain, and Abbeel 2020), we can obtain the following equivalent relationship:

$$\begin{aligned} &\mathbb{E}_{q(\boldsymbol{x}_{0:T})}\left[\log \frac{P_\theta(\boldsymbol{x}_{0:T}, y)}{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)}\right] \iff \\ &-\mathbb{E}_{t\sim[1,T],\boldsymbol{x},\varepsilon_t}\left[\left\|\varepsilon_t - \varepsilon_\theta\left(\boldsymbol{x}_t, t, y\right)\right\|^2\right]. \end{aligned} \tag{4}$$

3.2 DBMEF Overview

Fig. 2 illustrates the overall architecture of DBMEF, which comprises two critical components: confidence protector and diffusion model classifier. We further introduce two strategies to enhance the framework's performance: combining positive and negative text conditions and a voting mechanism. The DBMEF process is outlined as follows:

- An input image $\boldsymbol{x}$ is first processed through a deep discriminative model to extract its top-k labels. These labels are subsequently inputted into a confidence protector that employs a protection threshold to ascertain the necessity of further re-evaluation via a diffusion model.
- If the re-evaluation is required, both positive and negative text conditions are generated based on the top-k labels. Thereafter, the image $\boldsymbol{x}$, in conjunction with these text conditions, is introduced into a diffusion model classifier.
- Within this classifier, denoising is conducted under the governance of text condition controls. The denoising outcome is modulated by a negative control factor λ, which amalgamates the results from both positive and negative sets of text conditions to identify the label demonstrating the best predicted noise.
- Based on practical considerations, DBMEF can also flexibly extend from single-sample prediction of noise to multiple-sample prediction, further enhancing its effectiveness through consensus voting on the predictions.

3.3 Confidence Protector and Diffusion Classifier

The utility of the Confidence Protector is to emulate the human brain's determination of confidence when recognizing visual signals. Images that do not evoke enough confidence necessitate rethinking, while those recognized with sufficient assurance bypass the "guess and verify" process.

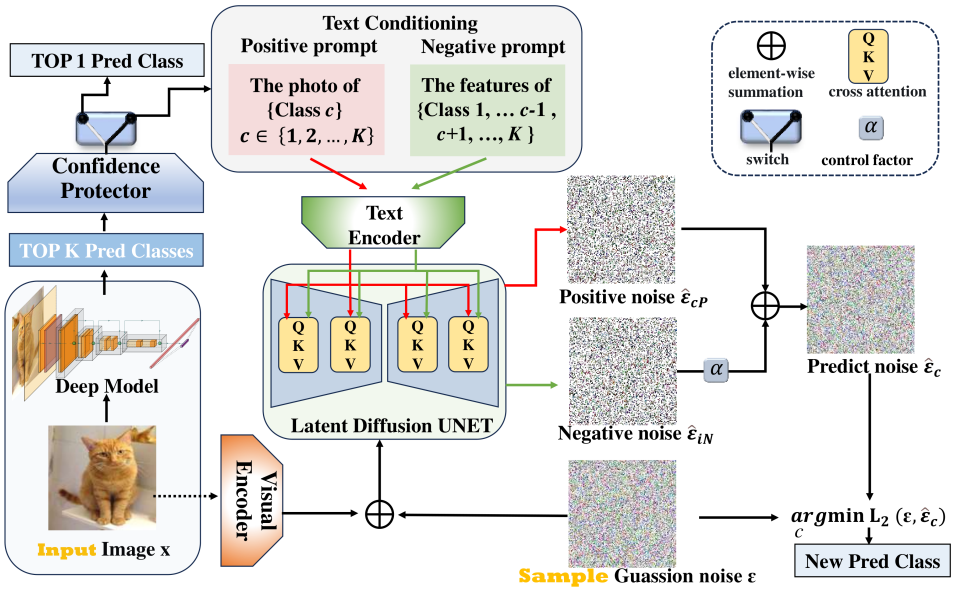

Figure 2: An overview of the Diffusion-Based Discriminative Model Enhancement Framework. For an input image $\boldsymbol{x}$, it first passes through a deep neural network to obtain its top-k labels. Then, a confidence protector determines the need for further analysis through a diffusion model. If required, positive and negative text conditions, derived from the top-k labels, are generated. These conditions, alongside $\boldsymbol{x}$, are then fed into the diffusion model. The label with the best denoising outcome is selected as the new predicted label.

For the protection threshold ($Prot$) within the Confidence Protector, we conduct a hypothesis test on the following issue: for an input image $\boldsymbol{x}$, is the output after processing by $f(\cdot)$ reliable?

$$H_0 : f(\boldsymbol{x}) \text{ is reliable} \quad \text{vs} \quad H_1 : f(\boldsymbol{x}) \text{ is not reliable} \tag{5}$$

Here the null hypothesis H_0 implies that the test input $\boldsymbol{x}$ can be reliably classified.

Regarding the determination of whether to reject the null hypothesis (H_0), we propose to derive a test statistic based on correctly classified samples within the training set where the discriminative model provides reliable predictions, as suggested in (Zhu et al. 2022). Given an $f(\cdot)$, an input image $\boldsymbol{x}_k$, assuming there are C categories, we can formulate the maximum probability of the model output as:

$$S(\boldsymbol{x}_k) = \max_{j \in [C]} \frac{\exp\left(f^j(\boldsymbol{x}_k)\right)}{\sum_{i=1}^{C} \exp\left(f^i(\boldsymbol{x}_k)\right)}. \tag{6}$$

Then, we can obtain a set of the maximum probability for the correctly classified images $\boldsymbol{S}$: $\{S(\boldsymbol{x}_1), S(\boldsymbol{x}_2), \ldots, S(\boldsymbol{x}_N)\}$, where $S(\boldsymbol{x}_i)$ represents the maximum probability of the model output for the i-th correctly classified image. We regard the lower α percentile of $\boldsymbol{S}$ as S_α as our test statistic and construct the rejection region for the null hypothesis H_0 as:

$$\mathcal{R} = \{\mathbf{x} : S(\boldsymbol{x}) \leq S_\alpha\}, \tag{7}$$

For any test input $\boldsymbol{x}$, the null hypothesis is rejected if $\boldsymbol{x} \in \mathcal{R}$. It is noteworthy that the α represents the significance level and also represents the maximum probability of the test making a Type I error which means the mistaken rejection of an actually true null hypothesis H_0. Protecting the null hypothesis H_0 is crucial in this hypothesis test, this leads to the introduction of $Prot$:

$$Prot = 1 - \alpha, \tag{8}$$

as a crucial parameter for the Confidence Protector.

Therefore, in order to reduce the probability of Type I errors, for preliminary discriminative models already exhibiting high accuracy, a higher $Prot$ is required while for a weaker discriminative model the $Prot$ should be appropriately smaller. In general, the range of $Prot$ is the open interval $(0, 1)$. When $Prot$ is 0, it means that all images need to undergo DBMEF. Conversely, when $Prot = 1$, it signifies that no images require re-thinking via DBMEF.

The diffusion model classifier transforms the conditional denoising outcomes of the diffusion model into an estimation of the posterior probability. In the Diffusion Model Classifier, the test dataset is set to $\left\{\left(\boldsymbol{x}^1, y^1\right), \ldots, \left(\boldsymbol{x}^n, y^n\right)\right\}$, for a given $\boldsymbol{x}$, $y \in \{C_1, C_2, \ldots, C_K\}$ represents the top-K possible labels output by the preliminary discriminative model. We transform the posterior probability $P(y = C_i \mid \boldsymbol{x})$ using Bayes' theorem and the reclassification results of the diffusion model are defined as follows:

$$\begin{aligned} \hat{y} &= \underset{C_i}{\operatorname{argmax}} P_\theta\left(y = C_i \mid \boldsymbol{x}\right) \\ &= \underset{C_i}{\operatorname{argmax}} \frac{P_\theta\left(\boldsymbol{x} \mid y = C_i\right)}{\sum_j P\left(y = C_j\right) P_\theta\left(\boldsymbol{x} \mid y = C_j\right)} P\left(y = C_i\right), \end{aligned} \tag{9}$$

To simplify the problem, we assume equiprobable classes and have $P(y = C_i) = \frac{1}{K}$. Thus,

$$\hat{y} = \underset{C_i}{\operatorname{argmax}} \log P_\theta(\boldsymbol{x} \mid y = C_i). \tag{10}$$

Then we use the reparameterized ELBO in Eq. (4) to replace $P_\theta(\boldsymbol{x} \mid y = C_i)$ in Eq. (10) as:

$$\begin{aligned} \hat{y} &\approx \underset{C_i}{\operatorname{argmax}} \mathbb{E}_{q(\boldsymbol{x}_{0:T})} \left[\log \frac{P_\theta(\boldsymbol{x}_{0:T}, y = C_i)}{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)} \right] \\ &= \underset{C_i}{\operatorname{argmin}} \mathbb{E}_{t\sim[1,T],\boldsymbol{x},\varepsilon_t} \left[\|\varepsilon_t - \varepsilon_\theta(\boldsymbol{x}_t, t, C_i)\|^2 \right]. \end{aligned} \tag{11}$$

the loss function is usually used to reparameterize the variational lower bound (VLB) as follows:

$$L_{VLB} = E_{q(\boldsymbol{x}_0:T)} \left[\log \frac{q(\boldsymbol{x}_{1:T} \mid \boldsymbol{x}_0)}{P_\theta(\boldsymbol{x}_0 : T)} \right], \tag{12}$$

$$L_{simple} = E_{t\sim[1,T],\boldsymbol{x},\varepsilon_t} \left[\|\varepsilon_t - \varepsilon_\theta(\boldsymbol{x}_t, t, C_i)\|^2 \right]. \tag{13}$$

Therefore, $\hat{y}$ can be written as:

$$\begin{aligned} \hat{y} &\approx \underset{C_i}{\operatorname{argmin}} L_{simple}(\boldsymbol{x}, y_{c_i}) \\ &= \underset{C_i}{\operatorname{argmin}} E_{t\sim[1,T],\boldsymbol{x},\varepsilon_t} \left[\|\varepsilon_t - \varepsilon_\theta(\boldsymbol{x}_t, t, C_i)\|^2 \right]. \end{aligned} \tag{14}$$

3.4 Negative Combination and Voting

The negative text conditions represent a reverse selection from the perspective of elimination. For each of the k alternative classes $\{C_1, \ldots, C_k\}$ for image $\boldsymbol{x}$, we construct the negative text conditions N_i for the i-th alternative class C_i as follows: $N_i = \{\text{The features of } C_1, \cdots, C_{i-1}, C_{i+1}, \cdots, C_k\}$. Under these negative text conditions, the predictive results of the diffusion model classifier in Eq. (14) can be reformulated as:

$$\hat{y} = \underset{N_i}{\operatorname{argmax}} \mathbb{E}_{t\sim[1,T],\boldsymbol{x},\varepsilon_t} \left[\|\varepsilon_t - \varepsilon_\theta(\boldsymbol{x}_t, t, N_i)\|^2 \right]. \tag{15}$$

Negative and positive text conditions serve as dual methods for articulating the same label. Therefore, combining both in a specific ratio is a natural approach. By integrating both, we anticipate an enhancement in label differentiation, which in turn is expected to elevate the model's precision. Following the (Dhariwal and Nichol 2021), after processing the positive and negative texts through the identical text-encoder, which generates distinct noises, the predicted noise for these divergent text conditions is merged in an element-wise summation at a predetermined ratio named the negative control factor λ ,as shown in the equation below:

$$Noise = Noise_{neg} + \lambda(Noise_{pos} - Noise_{neg}) \quad (\lambda \geq 1). \tag{16}$$

Moreover, voting ensemble (Dietterich 2000) is a common method to improve model accuracy. In our work, we have chosen a voting method that balances computational resources and performance, selecting the number of models participating in the vote as 5. Each model involved in the voting has the same hyperparameters and is a result of combining positive and negative text conditions within the DBMEF.

4 Experiments

This section provides extensive experiments to answer the following questions:

- Does our framework support various architectures of deep discriminative models with different training methods? See Sec. 4.1.
- Does our framework maintain its effectiveness under scenarios of distribution shifts and low-resolution conditions? See Sec. 4.2 and Sec. 4.3.
- What roles do the individual components of the framework play in enhancing model performance and how hyperparameters affect framework performance? See Appendix. B and Appendix. C.

4.1 Performance across Different Models

Baseline: We choose 17 discriminative models based on various training methods, datasets, and architectures. The selection comprises nine models employing supervised learning techniques, including DeiT-Base (DeiT-b), DeiT-Small (Deit-s), ViT-Base (ViT-b), ViT-Small (ViT-s), ResNet50, ResNet18, VGG16, MobileNetV3(Howard et al. 2019), and TinyNet(Han et al. 2020); six models that utilize self-supervised training methods and are fine-tuned on the linear layers with ImageNet2012 training set—specifically ViTb-MAE(He et al. 2022), ViTl-MAE, ViTs-DINOv2(Oquab et al. 2023), ViTb-DINOv2 and ResNet50-SimCLR(Chen et al. 2020), ResNet101-SimCLR; and two models using contrastive learning with fine-tuned linear layers: ViTb-CLIP(Radford et al. 2021), and ViTh-CLIP. The pretrained weights of nine supervised backbone models and ViTb-CLIP, ViTh-CLIP, as well as the image preprocessing methods used during inference, were sourced from the TIMM library [1]. The weights for the MAE model and the DINOv2 model are sourced from the official open-source repositories of Facebook and Google, respectively.

Set up details: Regarding the diffusion model, we adopt the most popular and mature architecture — Stable Diffusion V1-5. For the impact of other versions and types of diffusion models on DBMEF, please refer to Appendix. B. For DeiT-b, ViTh-CLIP, ViTb-CLIP, ViTb-MAE, ViTl-MAE, ViTb-DINOv2, we set $Prot$ at 0.99, while for other models, $Prot$ is set at 0.95. The time steps is set to 30, λ is fixed at 1.1, and the number of sub-models participating in voting is set to 5. In the hyperparameter experimentation section(see Appendix. B), we will investigate the impact of these key parameters on the potential for framework effectiveness enhancement. The evaluation experiments are conducted on the ImageNet2012-1k validation set. We fix the random seed and repeat the process five times, using the average as the reported result.

Results: Tab. 1 presents the performance enhancements of 17 deep discriminative models following their integration with DBMEF. The accuracy improvements on the ImageNet1k validation set ranged from 0.19% to 3.27%. Notably, to maintain a balance between accuracy and inference time, all models were configured to 30 timesteps, with-

[1] https://github.com/huggingface/pytorch-image-models

Training-method	Model	top1	w/o p	p+neg	p+pos	p+c	p+c+v	Δ
supervised learning on backbone model	**DeiT-b**	81.98%	66.37%	82.02%	82.41%	82.45%	**82.67%**	**0.69%**
	ViT-b	80.93%	64.98%	81.35%	81.40%	81.43%	**81.73%**	**0.80%**
	DeiT-s	79.86%	64.88%	80.18%	80.35%	80.42%	**80.78%**	**0.92%**
	ResNet50	76.15%	64.39%	76.68%	77.22%	77.40%	**77.66%**	**1.51%**
	ViT-s	75.99%	63.95%	76.72%	77.02%	77.10%	**77.64%**	**1.65%**
	VGG16	71.58%	63.98%	72.78%	73.39%	73.52%	**73.96%**	**2.38%**
	ResNet18	69.76%	63.44%	71.16%	71.85%	71.98%	**72.41%**	**2.65%**
	Mobilenetv3	67.64%	63.68%	69.55%	70.27%	70.40%	**70.79%**	**3.15%**
	Tinynet	66.96%	63.44%	68.87%	69.65%	69.72%	**70.23%**	**3.27%**
self-supervised learning + finetuning	**ViTb-MAE**	83.63%	65.97%	83.90%	83.95%	83.98%	**84.12%**	**0.49%**
	ViTl-MAE	85.92%	66.15%	86.09%	86.14%	86.21%	**86.29%**	**0.37%**
	ViTs-DINOv2	79.11%	65.14%	79.71%	79.95%	80.14%	**80.35%**	**1.24%**
	ViTb-DINOv2	82.01%	64.96%	82.21%	82.48%	82.52%	**82.69%**	**0.68%**
	ResNet50-SimCLR	76.31%	64.51%	77.35%	77.44%	77.53%	**77.71%**	**1.40%**
	ResNet101-SimCLR	78.22%	65.03%	78.41%	78.75%	78.82%	**79.03%**	**0.81%**
contrastive learning + finetuning	**ViTb-CLIP**	85.21%	65.66%	85.33%	85.36%	85.38%	**85.46%**	**0.25%**
	ViTh-CLIP	88.59%	65.81%	88.60%	88.61%	88.65%	**88.78%**	**0.19%**

Table 1: The method proposed in this article brings performance gains across different model architectures. We have compared the impact of various strategies for augmenting classification models with generative models. Here, top1 denotes the model's baseline accuracy. w/o p indicates using a strategy similar to that in (Li et al. 2023a), where the original task of selecting 1 out of 1000 is simplified to selecting 1 out of 5 without confidence protection. p+neg means the model has been processed by the confidence protector and only uses negative text as the text condition. p+pos indicates the model has been processed by the confidence protector and only uses positive text as the text condition. p+c represents the model being processed by the confidence protector and combining positive and negative text conditions. p+c+v signifies a voting process involving five p+c combinations, that is the most comprehensive DBMEF. Δ represents the improvement in the original top1 accuracy of the model achieved by applying DBMEF.

out optimizing other hyperparameters for optimal settings. Consequently, the full potential of the framework to boost the performance of deep discriminative models may not be entirely captured in these results. Nevertheless, the framework has demonstrated substantial improvements across the board, including for the highly advanced ViTh-CLIP model, which was derived from CLIP and trained and fine-tuned on LAION-2B and ImageNet-12K, further fine-tuned on ImageNet1k training set and the more modest TinyNet, underscoring its universal applicability.

The fourth column(w/o p) of Tab. 1 displays the results of directly reclassifying the top 5 labels from the discriminative model through the diffusion model without the protection step. This approach is similar to the traditional method of classification using diffusion models, simplifying the selection from 1 out of 1000 to 1 out of 5. Although its accuracy is higher than the performance in the previous works (Li et al. 2023a; Clark and Jaini 2024) on ImageNet, it is still significantly lower than the accuracy of the discriminative models themselves. This demonstrates that the unprotected method leads to inferior outcomes because re-evaluating all test input images might misclassify some of the results that the preliminary discriminative model had accurately predicted. This action lowers the overall performance of image recognition, thereby illustrating the essential role of confidence protector operations. More intuitive visualization results are presented in Appendix C.

The fifth column(p + neg) of Tab. 1 shows the results when applying both the Confidence Protector and solely the negative text condition, achieving classification accuracy that surpass the baseline top1 performance. However, its improvement is not as significant as when the Confidence Protector and only positive text condition are applied. The accuracy in sixth column(p+pos) far exceeds that of without protector (w/o p) in the third column, highlighting the crucial role of the Confidence Protector.

In the seventh column(p+c) of Tab. 1, the effect achieved by combining both the positive and negative text conditions using λ surpasses that of using either individual text condition alone, thereby demonstrating the effectiveness of employing λ to integrate these conditions. Moreover, as shown in the eighth column(p + c + v) of Tab. 1, the classification performance can be further enhanced by applying a voting mechanism to p+c. Ultimately, without additional hyperparameter tuning and only choosing small timesteps, DBMEF enhances the performance of these deep models by 0.19%-3.01%, fully demonstrating the excellent performance and potential of our framework.

4.2 Performance against Distribution Shifts

Baseline: We use the accuracy of the aforementioned four pre-trained deep discriminative models—ViT-Base, DeiT-Small, ResNet50, VGG16 on ImageNet-S, ImageNet-A, ImageNet-V2, ImageNet-E (background), and ImageNet-E

(position) as the baseline.

Set up details: We utilize the complete Diffusion-Based Discriminative Model Enhancement Framework, which includes 5 submodels for voting formed by positive and negative text condition combinations. The diffusion model selected in the framework is Stable Diffusion V1-5, with a $Prot$ 0.95 and time steps 30. λ is set to 1.1. The deep discriminative models chosen within the framework include ViT-b, Resnet50, VGG16, and DeiT-s. The pretrained weights are sourced from the TIMM library. We select the ImageNet-A (Hendrycks et al. 2021), ImageNet-V2(Recht et al. 2019), ImageNet-S(Gao et al. 2022), and Imagenet-E(Li et al. 2023b) datasets to assess the robustness of our proposed framework against distribution shifts. Additionally, the category labels of these datasets are subsets of the ImageNet1k labels, so no additional training is required.

Model	S	A	V2	E-bg	E-pos
ViT-b	29.79%	27.23%	77.23%	67.55%	75.39%
ViT-b*	**30.47%**	**27.41%**	**77.37%**	**68.61%**	**75.92%**
Resnet50	24.08%	0.00%	72.30%	62.22%	65.17%
Resnet50*	**25.82%**	**3.02%**	**72.98%**	**63.35%**	**65.31%**
VGG16	17.54%	2.57%	68.14%	56.71%	61.21%
VGG16*	**19.70%**	**3.84%**	**69.29%**	**57.92%**	**61.92%**
DeiT-s	29.42%	18.68%	76.42%	61.14%	70.29%
DeiT-s*	**31.01%**	**19.34%**	**77.02%**	**62.22%**	**70.74%**

Table 2: Results of the Distribution Shift Experiment. S, A and V2 denote ImageNet-S, ImageNet-A and ImageNet-V2 respectively. The asterisk (*) indicates enhanced classifier using our proposed DBMEF, while metrics without an asterisk represent vanilla classifier. The best results are highlighted in Bold.

As shown in Tab. 2, the DBMEF demonstrates stable improvement capability when facing different types of distribution shift datasets, including real images misclassified by ResNet50, new images with the same ImageNet1k labels, sketch images, and edited versions of original ImageNet images with altered backgrounds and object poses, with the most significant improvements observed for ImageNet-S and ImageNet-A. Notably, on ImageNet-A, DBMEF improved the performance of Resnet50 from 0.0% to 3.02%, indicating that DBMEF endowed the model with the ability to reconsider, thereby enhancing accuracy in more adversarial conditions.

4.3 Classification on Low-Resolution Datasets

Set up details: We selected the deep discriminative models ResNet18, ResNet34 and ResNet50. Additionally, we chose the CIFAR-10 and CIFAR-100 datasets for our experiments. These two datasets contain 10 and 100 categories, respectively, with each category's images having a resolution of 32 $\times$ 32 pixels(Krizhevsky, Hinton et al. 2009).

Baseline: We use the classification accuracy of ResNet18, ResNet34 and ResNet50, which have been fine-tuned 20 epochs on the CIFAR-10 and CIFAR-100 training sets with weights pre-trained on ImageNet, on their respective test sets as the baseline. The parameter settings for the DBMEF are identical to those described in Sec. 4.2.

DataSet	CIFAR10	CIFAR100
ResNet18	95.16%	80.94%
ResNet18*	**95.20%**	**81.22%**
ResNet34	95.41%	81.45%
ResNet34*	**95.44%**	**81.85%**
ResNet50	96.29%	83.85%
ResNet50*	**96.30%**	**84.46%**
VGG16	91.02%	70.49%
VGG16*	**91.68%**	**71.42%**

Table 3: Model performance on CIFAR10 and CIFAR100 datasets. The asterisk (*) indicates enhanced classifier using our proposed DBMEF, while metrics without an asterisk represent vanilla classifier. The best results are highlighted in bold.

In Tab. 3, despite the inherently high accuracy of ResNet18, ResNet34 and ResNet50 on two datasets, the application of our framework yielded further improvements. This demonstrates the effectiveness of strategic coordination between the protective mechanism and the diffusion classifier, even in scenarios where the baseline accuracy is already substantial. Moreover, for low-resolution images, transitioning the original images to a reduced-dimensional latent space (32 $\times$ 32) — as opposed to the larger latent space dimensions (64 $\times$ 64) utilized in the experiments on ImageNet — resulted in enhanced performance and increased processing speed concurrently. A more detailed discussion of inference speed is presented in the Appendix. D.

5 Conclusion

In modern image classification tasks, deep learning methods conventionally use either discriminative models or generative models independently. Our paper draws inspiration from the human brain's coordination of rapid and slow pathways in recognition tasks, proposing a novel framework—the Diffusion-Based Discriminative Model Enhancement Framework (DBMEF). This framework can effectively enhance the classification accuracy and generalization capability of discriminative models in a plug-and-play and training-free manner. We discovered that DBMEF exhibits strong universality, achieving stable performance improvements across 17 common deep discriminative models, including different network architectures and training methods. Additionally, DBMEF still achieves good improvement effects when facing data with distribution shifts and low-resolution data. Our work fills a gap within the present research field and aims to motivate researchers to further investigate the integration of diffusion models into more downstream applications, combining discriminative and generative modeling principles to fully harness their respective strengths.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (No. 12201048).

References

Chen, H.; Dong, Y.; Wang, Z.; Yang, X.; Duan, C.; Su, H.; and Zhu, J. 2023. Robust Classification via a Single Diffusion Model. *arXiv preprint arXiv:2305.15241*.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, 1597–1607. PMLR.

Chen, Y.; Akin, O.; Nern, A.; Tsui, C. K.; Pecot, M. Y.; and Zipursky, S. L. 2014. Cell-type-specific labeling of synapses in vivo through synaptic tagging with recombination. *Neuron*, 81(2): 280–293.

Cheng, J.; and Greiner, R. 2013. Comparing Bayesian network classifiers. *arXiv preprint arXiv:1301.6684*.

Clark, K.; and Jaini, P. 2024. Text-to-Image Diffusion Models are Zero Shot Classifiers. *Advances in Neural Information Processing Systems*, 36.

Dhariwal, P.; and Nichol, A. 2021. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34: 8780–8794.

Dietterich, T. G. 2000. Ensemble methods in machine learning. In *International workshop on multiple classifier systems*, 1–15. Springer.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Gao, S.; Li, Z.-Y.; Yang, M.-H.; Cheng, M.-M.; Han, J.; and Torr, P. 2022. Large-scale unsupervised semantic segmentation. *IEEE transactions on pattern analysis and machine intelligence*.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2020. Generative adversarial networks. *Communications of the ACM*, 63(11): 139–144.

Han, K.; Wang, Y.; Zhang, Q.; Zhang, W.; Xu, C.; and Zhang, T. 2020. Model rubik's cube: Twisting resolution, depth and width for tinynets. *Advances in Neural Information Processing Systems*, 33: 19353–19364.

Han, X.; Zheng, H.; and Zhou, M. 2022. Card: Classification and regression diffusion models. *Advances in Neural Information Processing Systems*, 35: 18100–18115.

He, K.; Chen, X.; Xie, S.; Li, Y.; Dollár, P.; and Girshick, R. 2022. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 16000–16009.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Hendrycks, D.; Zhao, K.; Basart, S.; Steinhardt, J.; and Song, D. 2021. Natural adversarial examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15262–15271.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33: 6840–6851.

Ho, J.; and Salimans, T. 2022. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*.

Howard, A.; Sandler, M.; Chu, G.; Chen, L.-C.; Chen, B.; Tan, M.; Wang, W.; Zhu, Y.; Pang, R.; Vasudevan, V.; et al. 2019. Searching for mobilenetv3. In *Proceedings of the IEEE/CVF international conference on computer vision*, 1314–1324.

Karras, T.; Aittala, M.; Aila, T.; and Laine, S. 2022. Elucidating the design space of diffusion-based generative models. *Advances in Neural Information Processing Systems*, 35: 26565–26577.

Krizhevsky, A.; Hinton, G.; et al. 2009. Learning multiple layers of features from tiny images. *Master's thesis, University of Tront*.

LeCun, Y.; Bottou, L.; Bengio, Y.; and Haffner, P. 1998. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11): 2278–2324.

Li, A. C.; Prabhudesai, M.; Duggal, S.; Brown, E. L.; and Pathak, D. 2023a. Your Diffusion Model is Secretly a Zero-Shot Classifier. In *ICML 2023 Workshop on Structured Probabilistic Inference & Generative Modeling*.

Li, X.; Chen, Y.; Zhu, Y.; Wang, S.; Zhang, R.; and Xue, H. 2023b. ImageNet-E: Benchmarking Neural Network Robustness via Attribute Editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 20371–20381.

Oquab, M.; Darcet, T.; Moutakanni, T.; Vo, H.; Szafraniec, M.; Khalidov, V.; Fernandez, P.; Haziza, D.; Massa, F.; El-Nouby, A.; et al. 2023. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*.

Peebles, W.; and Xie, S. 2023. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4195–4205.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Recht, B.; Roelofs, R.; Schmidt, L.; and Shankar, V. 2019. Do imagenet classifiers generalize to imagenet? In *International conference on machine learning*, 5389–5400. PMLR.

Rish, I.; et al. 2001. An empirical study of the naive Bayes classifier. In *IJCAI 2001 workshop on empirical methods in artificial intelligence*, volume 3, 41–46.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.

Sillito, A. M.; Cudeiro, J.; and Jones, H. E. 2006. Always returning: feedback and sensory processing in visual cortex and thalamus. *Trends in neurosciences*, 29(6): 307–316.

Simonyan, K.; and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Song, J.; Meng, C.; and Ermon, S. 2020. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*.

Song, Y.; Sohl-Dickstein, J.; Kingma, D. P.; Kumar, A.; Ermon, S.; and Poole, B. 2020. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*.

Tang, L.; Jia, M.; Wang, Q.; Phoo, C. P.; and Hariharan, B. 2024. Emergent correspondence from image diffusion. *Advances in Neural Information Processing Systems*, 36.

Touvron, H.; Cord, M.; Douze, M.; Massa, F.; Sablayrolles, A.; and Jégou, H. 2021. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, 10347–10357. PMLR.

Xu, J.; Liu, S.; Vahdat, A.; Byeon, W.; Wang, X.; and De Mello, S. 2023. Open-vocabulary panoptic segmentation with text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2955–2966.

Yoon, J.; Hwang, S. J.; and Lee, J. 2021. Adversarial purification with score-based generative models. In *International Conference on Machine Learning*, 12062–12072. PMLR.

Zhao, S.; Jacobsen, J.-H.; and Grathwohl, W. 2020. Joint energy-based models for semi-supervised classification. In *ICML 2020 Workshop on Uncertainty and Robustness in Deep Learning*, volume 1.

Zhao, W.; Rao, Y.; Liu, Z.; Liu, B.; Zhou, J.; and Lu, J. 2023. Unleashing text-to-image diffusion models for visual perception. *arXiv preprint arXiv:2303.02153*.

Zhu, Y.; Chen, Y.; Li, X.; Zhang, R.; Xue, H.; Tian, X.; Jiang, R.; Zheng, B.; and Chen, Y. 2022. Rethinking Out-of-Distribution Detection From a Human-Centric Perspective. *arXiv preprint arXiv:2211.16778*.

Zimmermann, R. S.; Schott, L.; Song, Y.; Dunn, B. A.; and Klindt, D. A. 2021. Score-Based Generative Classifiers. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*.