Stress Testing Classifiers around the Decision Boundary with Latent Diffusion

Inês Gomes[1,2,4] **André Restivo**[1,2] **Moisés Santos**[1] **Carlos Soares**[1,2,3] **Jan N. van Rijn**[4] **Luís F. Teixeira**[1,5]

[1]Faculty of Engineering, University of Porto
[2]Artificial Intelligence and Computer Science Laboratory, Portugal
[3]Fraunhofer Portugal AICOS, Portugal
[4]Leiden Institute of Advanced Computer Science (LIACS), Leiden University, the Netherlands
[5]INESC TEC, Portugal

Abstract Deep learning models can achieve high predictive performance and yet still fail in unexpected ways. Stress testing is a model auditing approach that identifies and reports model limitations in a data-driven way. One approach is to generate images near the decision boundary of a classifier, where predictions are most uncertain, thereby improving our understanding of its decision-making process. Existing methods train variational autoencoders or generative adversarial networks and are limited to small images and pairs of classes. To overcome these limitations, we propose an alternative architecture based on diffusion models. First, we introduce a new metric — Kullback–Leibler divergence to the decision boundary (KLDB) — which measures how close an image is to the decision boundary of any class subset, allowing users to decide which decision boundaries to explore. Then, we use KLDB to guide the sampling of a pretrained text-to-image latent diffusion model. On ImageNet subsets, our method reduces KLDB by up to 51% over prompt-only baselines, and by visual inspection, it generates ambiguous images that expose classifier limitations. This method enables stress testing on complex datasets without retraining diffusion models and supports auditing any decision boundary by selecting arbitrary class combinations.

1 Introduction

Understanding how deep learning models make decisions is important for interpretability and trust (Barredo Arrieta et al., 2020), especially in safety-critical domains like healthcare or autonomous systems. A recent approach to auditing these models is called *stress testing*, which identifies and reports model limitations in a data-driven way (Gomes et al., 2024). One line of work focuses on the generation of images near the decision boundary of classifiers to characterize their limitations and better understand their decision-making process. Existing methods like DeepDIG (Karimi and Derr, 2022), GASTeN (Cunha et al., 2023), and MIMICRY (Abdellatif et al., 2024) use variational autoencoders (VAEs) or generative adversarial networks (GANs) to create such samples. However, these approaches have several limitations. First, they are restricted to small or grayscale images, which limits scalability. Second, they require training a generative model from scratch, adding complexity and computational cost. Third, there is no standard way to measure proximity to the decision boundary, meaning that each method defines its own, making comparisons difficult. Finally, most existing methods focus on auditing binary classifiers or, in the multiclass setting, reduce the analysis to pairs of classes, which limits the exploration of more complex decision boundaries.

To overcome these limitations, we propose a new stress testing framework based on pretrained latent diffusion models. Our method combines text prompts with classifier guidance during denoising, requiring no generative model training. As a result, it supports full-colour, high-resolution

(a) reference (b) prompt only (c) prompt + classifier

Figure 1: Images from the classes “chihuahua” and “papillon” sampled from the reference (a) near (i) and far (ii) from the decision boundary; (b) Synthetic images generated using only prompt guidance; (c) Synthetic images with our method, combining prompt and classifier guidance.

images while reducing implementation complexity. We introduce a new metric, Kullback–Leibler divergence to the decision boundary (KLDB), which quantifies how close an image is to a classifier’s decision boundary for any selected subset of classes. Prompt guidance uses classifier-free guidance (Ho et al., 2021) to influence the generation within the desired class subset. Classifier guidance updates the latents to guide images toward regions with low KLDB, that is, near the decision boundary. The method works with any subset of classes and requires only a pretrained latent diffusion model.

We compare our method to a prompt-only baseline on ImageNet (Deng et al., 2009) subsets and observe that the generated images achieve KLDB values between 48% and 51% lower, indicating that the images generated by our method are closer to the decision boundary. Visual inspection confirms this finding, as many images display ambiguous features, hybrids, or occlusions that confuse the classifier, which can be used to stress test the classifier and highlight its limitations.

2 Stress Testing with Latent Diffusion Models

We propose a method to synthesize images near any decision boundary of a given classifier f_ϕ using a pretrained text-to-image latent diffusion model G_θ. Our framework works in two stages. First, given a subset of target classes $\mathcal{C}$, we define a textual prompt that lists the class labels, conditioning the diffusion model via classifier-free guidance (Ho et al., 2021). Then, to guide the sampling to the decision boundary between classes in $\mathcal{C}$, we introduce a new metric called Kullback–Leibler divergence to the decision boundary (KLDB). Existing metrics for boundary-based image generation (e.g., Cunha et al. (2023), Karimi and Derr (2022), Weiss et al. (2023)) are limited to binary classification or fixed class pairs, and do not support arbitrary subsets in multiclass settings.

KLDB, defined by Equation 1, addresses this limitation by measuring the pointwise Kullback–Leibler (KL) divergence between the classifier predicted probability distribution f_ϕ, and a target distribution B over the selected classes $\mathcal{C} \subseteq \{1, 2, \ldots, n\}$. For a calibrated classifier, the decision boundary between classes in $\mathcal{C}$ occurs where $f_\phi(x)$ is equiprobable, i.e., $f_\phi(x_i) = \frac{1}{|\mathcal{C}|}$. In this setting, the KL divergence penalises deviations from uniformity, reaching zero only at the boundary. Thus, KLDB is an unbounded metric: higher values indicate a greater distance from the decision boundary, while a value of zero means that the data point lies exactly on it.

$$\mathrm{KLDB}(B \,||\, f_\phi(x)) = \sum_{i=1}^{n} B_i \log\left(\frac{B_i}{softmax(f_\phi(x))}\right), \quad \text{where} \quad B_i = \begin{cases} \frac{1}{|\mathcal{C}|} & \text{if } i \in \mathcal{C} \\ 0 & \text{if } i \notin \mathcal{C} \end{cases} \tag{1}$$

Since KLDB is differentiable, we use it to define a contrastive guiding loss that is integrated into the diffusion sampling process, following approaches such as Sehwag et al. (2022) and Hemmat et al. (2024). At each denoising step t, we modify the latent variable update to include a gradient correction that minimizes the KLDB loss.

Equation 2 summarises our modified sampling strategy, where SchedulerStep$(\cdot)$ is the denoising update from a predefined diffusion scheduler; $\epsilon_\theta^{\text{CFG}}(z_t, t, \beta)$ is the classifier-free guided noise prediction where β controls the strength of the prompt conditioning and $\mathcal{C}$ is the textual prompt; and α is a hyperparameter controlling the strength of the gradient correction. The gradient ∇^* is normalized to maintain stability across timesteps, as in Sehwag et al. (2022). Finally, following the approach from Hemmat et al. (2024), the KLDB loss is computed by first decoding the latent variable z_t into the pixel space using the decoder D_θ, and then applying the classifier f_ϕ to the decoded image. This process requires the scheduler to be deterministic.

$$z_{t-1} = \text{SchedulerStep}\left(z_t, \epsilon_\theta^{\text{CFG}}(z_t, t, \beta)\right) + \alpha \cdot \nabla^*_{z_t} \mathcal{L}_{\text{KLDB}}(z_t, C) \tag{2}$$

3 Experimental Setup

We build on the Stable Diffusion v1.4 model (Rombach et al., 2022) with the deterministic scheduler k-LMS, adopting the sampling modifications proposed by Lin et al. (2024). The denoising process uses $T = 40$ steps, and classifier guidance is applied every five steps, following the Hemmat et al. (2024). For classifier feedback, we use a ResNet-50 (He et al., 2016) pretrained on ImageNet-1k (Deng et al., 2009). We perform hyperparameter optimisation on the classifier guidance strength $\alpha \in \{0, 0.05, 0.1, 0.15, 0.2, 0.5, 1\}$ and the prompt conditioning scale $\beta \in \{3, 6, 9\}$ via classifier-free guidance, following best practices in diffusion literature. $\alpha = 0$ serves as our prompt-only baseline.

Our evaluation focuses on one challenging ImageNet-1k subset with semantically similar classes: *chihuahua / papillon*. We generate 2,500 samples and the prompt is "chihuahua and papillon" to isolate the effect of classifier guidance. Evaluation metrics include: KLDB score to measure the proximity to the decision boundary; Shannon entropy of softmax predictions, to quantify the uncertainty of prediction (Weiss et al., 2023); FID (Heusel et al., 2017) to compare feature distributions between generated and reference images. Experiments were conducted on two NVIDIA GeForce RTX 2080 Ti GPUs and the proposed framework code is available[1] for reproducibility.

4 Results

In this section we present both quantitative and qualitative results. First, we evaluate how our framework can synthesize images close to the decision boundary while maintaining image quality. Then, we explore how the generated images can support model auditing by indicating classifier sensitivities and failure modes.

4.1 Impact of classifier guidance

We evaluate how classifier guidance strength α affects the synthetic images proximity to decision boundary and image quality. Table 1 reports, for each prompt strength β, the value of α with the lowest median KLDB score, compared to the baseline ($\alpha = 0$), that is, prompt-only generation.

We find that weaker prompt conditioning ($\beta = 3$) benefits from lower classifier guidance ($\alpha = 0.05$), while stronger prompts ($\beta = 6, 9$) achieve best results with more guidance ($\alpha = 0.10$). This suggests that weaker prompt conditioning allow classifier signals to guide the generation more easily. In all cases, classifier guidance significantly reduces KLDB scores by 48 to 51% compared to the prompt-only baseline. Entropy has similar results as images near the decision boundary exhibit higher uncertainty.

[1] `https://anonymous.4open.science/r/diffusion-boundary-4214/README.md`

β	α	FID ↓	M(KLDB) ↓	M(Entropy) ↑
3	0	62.83	2.48	0.08
	0.05	63.00	1.22	0.32
6	0	83.81	3.12	0.03
	0.10	82.15	1.59	0.16
9	0	91.37	3.16	0.03
	0.15	88.96	1.65	0.16

Table 1: Evaluation results for baseline ($\alpha = 0$) and best classifier guidance (α) for each $\beta \in \{3, 6, 9\}$, using the prompt "chihuahua and papillon". Arrows indicate the desirable score direction.

Finally, image quality remains stable across settings. The best α configurations maintain FID scores close to baseline values, indicating that guidance can increase decision-boundary alignment without degrading perceptual quality.

4.2 Model Auditing

Building on Gomes et al. (2024), we use synthetic image generation to uncover features that drive a classifier to be uncertainty of its decision. These insights can help diagnose model weaknesses and inform model cards (Mitchell et al., 2019).

Figure 1 compares real images, prompt-only generations, and our method. Generated images typically focus on the subject and simplify backgrounds, which may help approach the decision boundary but reduce visual diversity. Occlusions or missing parts in outputs may highlight the classifier's reliance on specific features, while unusual styles (e.g., grayscale, cartoon-like) reveal sensitivity to out-of-distribution inputs.

Our method also produces hybrid images that merge features from multiple classes. While effective in reducing classifier confidence, these images may lack realism and thus limit their diagnostic utility. Future work could incorporate a realism discriminator to improve image plausibility during sampling.

5 Conclusions

We present a classifier-guided diffusion method to generate synthetic images near a classifier's decision boundary. Prior approaches used VAEs or GANs were limited to binary and small-scale tasks. Our method scales to multiclass settings and large datasets using a pretrained latent diffusion model and a new metric, KLDB, which quantifies proximity to the decision boundary across any number of classes. We use KLDB to guide the sampling process through two hyperparameters: α for classifier guidance and β for prompt scaling.

Experiments on subsets of ImageNet-1k show that our method effectively reduces KLDB compared to prompt-only baselines, producing images closer to the decision boundary. We also observed that low α values help find the boundary, while high β values improve prompt fidelity but reduce diversity. Visual inspection revealed that many boundary images exhibit ambiguous or mixed-class features, highlighting model uncertainty and potential limitations.

A limitation of this work is our current setup, since we use a single classifier, dataset, and diffusion model. Future work should explore how well our method generalizes across models and domains. Nonetheless, our approach provides a scalable and data-driven way to study decision boundaries, usable in combination with any available diffusion model, which can support the analysis of model behaviours and their robustness.

Acknowledgements. This work was partially funded by projects AISym4Med, CRAI, LIACC and the Leiden University SAILS programme. European Union under the Horizon Europe Framework Programme Grant Agreement Nº: 101095387; Agenda "Center for Responsible AI", nr. C645008882-00000055, investment project nr. 62, financed by the Recovery and Resilience Plan (PRR) and by European Union — NextGeneration EU. Funded by the European Union — NextGenerationEU; UID/00027 of the LIACC — Artificial Intelligence and Computer Science Laboratory — funded by Fundação para a Ciência e a Tecnologia, I.P./ MCTES through the national funds. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Commission. Neither the European Union nor the European Commission can be held responsible for them.

References

Abdellatif, A., Chen, X., Riccio, V., and Stocco, A. (2024). Deep Learning System Boundary Testing through Latent Space Style Mixing. *CoRR*.

Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., and Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58:82–115.

Cunha, L., Soares, C., Restivo, A., and Teixeira, L. F. (2023). GASTeN: Generative Adversarial Stress Test Networks. In *Advances in Intelligent Data Analysis XXI: 21st International Symposium on Intelligent Data Analysis*, pages 91–102.

Deng, J., Dong, W., Socher, R., Li, L.-j., Li, K., and Fei-fei, L. (2009). ImageNet : A Large-Scale Hierarchical Image Database. *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255.

Gomes, I., Teixeira, L. F., van Rijn, J. N., Soares, C., Restivo, A., Cunha, L., and Santos, M. (2024). Finding Patterns in Ambiguity: Interpretable Stress Testing in the Decision Boundary. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 8316–8321.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778.

Hemmat, R. A., Pezeshki, M., Bordes, F., Drozdzal, M., and Romero-Soriano, A. (2024). Feedback-guided Data Synthesis for Imbalanced Classification. *Transactions on Machine Learning Research*.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. (2017). GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. In *Advances in Neural Information Processing Systems*, volume 30.

Ho, J., Research, G., and Salimans, T. (2021). Classifier-Free Diffusion Guidance. In *Workshop on Deep Generative Models and Downstream Applications, NeurIPS*.

Karimi, H. and Derr, T. (2022). Decision Boundaries of Deep Neural Networks. In *21st IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 1085–1092.

Lin, S., Liu, B., Li, J., and Yang, X. (2024). Common Diffusion Noise Schedules and Sample Steps are Flawed. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 5404–5411.

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., and Gebru, T. (2019). Model Cards for Model Reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, pages 220–229.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. (2022). High-Resolution Image Synthesis with Latent Diffusion Models. In *IEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10674–10685.

Sehwag, V., Hazirbas, C., Gordo, A., Ozgenel, F., and Canton-Ferrer, C. (2022). Generating High Fidelity Data from Low-density Regions using Diffusion Models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11482–11491.

Weiss, M., Gómez, A. G., and Tonella, P. (2023). Generating and detecting true ambiguity: a forgotten danger in DNN supervision testing. *Empirical Software Engineering*, 28(6):146.