A Dataless Approach to Latent Evolutionary Images

Caio Santana, Arthur Buzelin, Pedro Bento, Yan Aquino,
Victoria Estanislau, Samira Malaquias, Wagner Meira Jr., Gisele L. Pappa
Dept. de Ciência da Computação, Universidade Federal de Minas Gerais (UFMG), Belo Horizonte, Brazil
{caiosantana, arthurbuzelin, pedro.bento, yanaquino, victoria.estanislau, samiramalaquias, meira, glpappa}@dcc.ufmg.br

***Abstract*—Building on our previous *Evolutionary Bias Identification via Embeddings* (EBIE) framework, we present *Latent Evolutionary Images* (LEI), a dataless framework for auditing black-box image classifiers by evolving interpretable synthetic images in the latent space of a generative model. Unlike traditional adversarial or interpretability approaches, LEI operates entirely without labeled training data or gradient access, relying instead on a genetic algorithm that explores the latent space of a pre-trained Variational Autoencoder (VAE). A frozen classifier evaluates the fitness of each generated image, steering evolution toward samples that maximize confidence in a chosen target class. We demonstrate that this process reliably uncovers latent directions associated with class semantics, even in the absence of explicit supervision. Experimental results show that LEI not only produces visually coherent representations aligned with the classifier's internal logic, but also reveals class-specific biases and facilitates the generation of effective latent adversarial examples[1].**

***Index Terms*—evolutionary algorithms, latent space optimization, variational autoencoder, black-box model auditing, image classification, image generation**

I. Introduction

As machine learning systems are increasingly deployed in safety-critical areas – such as medical diagnostics, autonomous driving, and content moderation [1]–[3], the need for tools that promote model transparency has grown accordingly. A central challenge in this effort is auditing the behavior of black-box image classifiers: models whose training data, architecture, or internal parameters are partially or entirely unknown. Understanding what such models "look for" when making decisions is crucial for revealing hidden biases, spurious correlations, and other potentially harmful behaviors.

In our previous work, the *Evolutionary Bias Identification via Embeddings* (EBIE) framework [4] proposed an evolutionary method for uncovering bias in text-based models. By interpreting the black-box model as a fitness function within an evolutionary algorithm, we evolved synthetic sentences that gradually came to resemble the model's original training data. Analysis of these sentences and their embedding trajectories revealed several hidden biases in the model's behavior.

However, that original approach was limited to textual data. While EBIE marked a significant step forward, this limitation restricts its applicability. Many real-world scenarios where bias detection is critical involve other formats such as images or videos. For instance, biased behavior can surface in face recognition systems, autonomous driving perception models, or medical image classifiers [5]–[7].

[1]https://github.com/caio-santt/LEI-Latent-Images-Evolution

To extend EBIE beyond text, we propose LEI, a methodology that adapts the same evolutionary search to the visual domain. Furthermore, instead of manipulating raw pixels – which often leads to unnatural or adversarial artifacts – we operate in the latent space of any given generative model. In this new pipeline, generic pre-trained Variational Autoencoders (VAEs) were used to decode latent vectors into a population of images, whose evolution is guided by the classifier's feedback. This allows us to explore and expose the model's internal decision boundaries in a more semantically meaningful way, even without opening the black box or having access to the training dataset.

II. Related Work

As previously mentioned, our earlier work [4] demonstrated how evolutionary search can be used to expose hidden biases in NLP models without access to their training data. Therefore, it laid the groundwork for applying similar black-box evolutionary strategies to other modalities, such as images. To contextualize our approach, we review a range of related methods covered as follows.

Black-box Model Auditing and Replication. Recent research has explored a range of strategies to understand and audit the behavior of black-box classifiers. Barbalau et al. [8] propose an evolutionary generative framework that trains surrogate models on synthetic inputs eliciting confident responses from a target model, while Lomurno et al. [9] introduce a knowledge distillation pipeline that leverages synthetic data to replicate classifier behavior while preserving data privacy.

Interpretability through Latent Manipulation. Beyond replication, several approaches have focused on enhancing interpretability through manipulation of latent representations in generative models. DISCOVER [10], for example, enables counterfactual explanations in medical imaging by learning disentangled latent spaces. Similarly, Voynov and Babenko [11] propose an unsupervised method to identify semantically meaningful directions within GAN latent spaces, allowing for controlled and interpretable image transformations.

Exploratory Search in Latent Space. Building on the potential of latent space representations, other works have applied evolutionary or heuristic search methods to actively explore and generate informative samples. GLASSE [12] evolves latent vectors in a Conditional GAN to produce adversarial examples without relying on gradient information. EvocraftGAN [13] combines quality-diversity search with

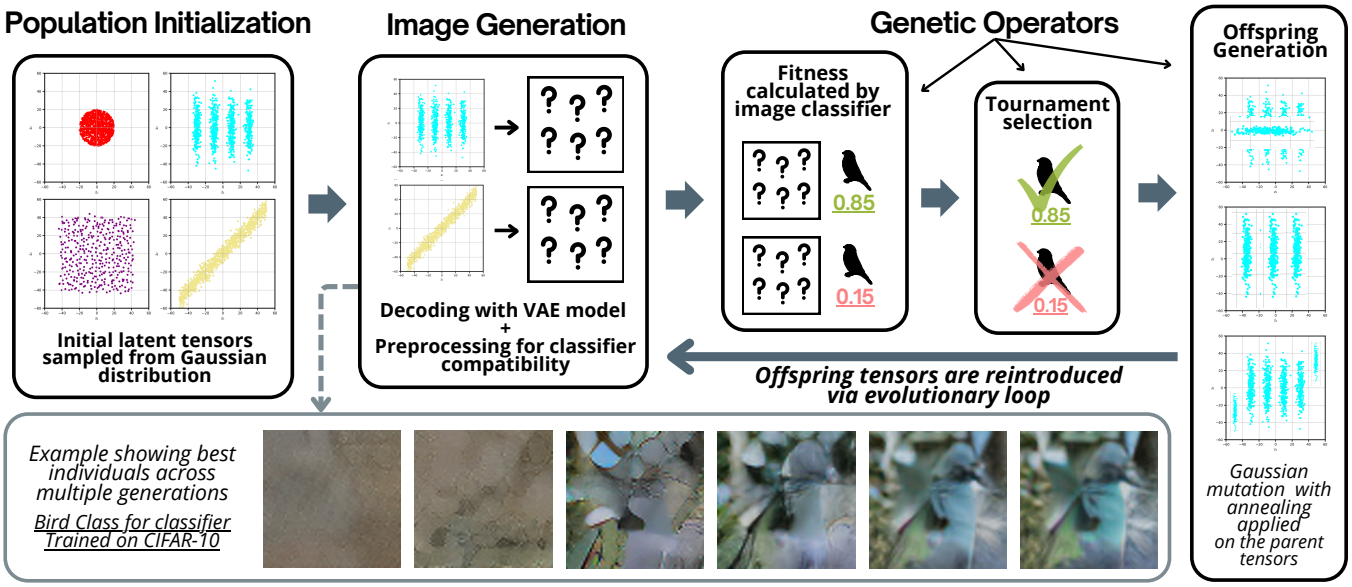


Fig. 1: Overview of the LEI pipeline described in Section III, followed by an example run generating a *bird* image.
Note: *This diagram displays a simplified, didactic, 2D visualization of the high-dimensional latent tensors used in the algorithm.*

CLIP-based scoring to uncover semantically rich and diverse image generations across the latent space.

Black-box Generative Optimization. In parallel, black-box optimization techniques have been employed to guide generative models without gradient. Ghanem et al. [14] evolve latent vectors using perceptual quality metrics such as NIQE to consistently enhance image fidelity from GANs. In a different modality, Iglesias et al. [15] apply local search to optimize negative prompts in text-to-image generation based on sentence similarity.

III. Methodology

Latent Evolutionary Images (LEI) extends the evolutionary core of EBIE and is a genetic algorithm [16] that evolves a population of images generated from latent vectors of a pre-trained Variational Autoencoder (VAE) [17]. Guided by the feedback of a frozen image classifier, the method steers the population toward figures that increasingly resemble a specific target class. LEI operates directly on compact latent representations, conducting the evolutionary search entirely within the VAE's latent space.

This section presents a detailed description of the algorithm and its underlying pipeline, illustrated in Figure 1. We begin by formalizing the representation of individuals in said latent space and describing the initialization of the population. We then introduce the fitness function, derived from an image classifier, along with the preprocessing steps necessary for compatibility. The genetic operators responsible for evolving the population are detailed next, including the mutation mechanism and crossover. We also explain the rationale behind the choice of the generative and classification models, outline the hyperparameters that govern the evolutionary process, and describe the steps taken to ensure reproducibility of all experiments.

A. Latent Representation and Population Initialization

Each individual in the evolutionary population is represented as a latent vector $z \in \mathbb{R}^{C \times H \times W}$, where $C = 4$, $H = 8$, and $W = 8$. These latent vectors correspond to compressed representations of 32×32 RGB images within the bottleneck space of a pre-trained Variational Autoencoder (VAE). Specifically, the experiments presented in this paper were conducted using the `stabilityai/sd-vae-ft-ema` model[2], which was trained on LAION-Aesthetics and LAION-Humans datasets and employs exponential moving average (EMA) weights.

To evaluate the generalizability of our approach, we also tested other generative models, such as `stabilityai/sdxl-vae`[3] and `stabilityai/sd-vae-ft-mse`[4], both producing qualitatively satisfactory results. The former was trained from scratch with larger batch sizes and improved high-frequency detail retention, while the latter continued training from the `sd-vae-ft-ema` checkpoint with a focus on minimizing mean squared reconstruction error.

Operating in the latent space offers several advantages: it restricts the search to plausible image representations, reduces the risk of unrealistic artifacts, and significantly lowers the dimensionality compared to pixel-level manipulations. These properties make the latent space particularly well-suited for the application of perturbations during genetic operations.

[2]https://huggingface.co/stabilityai/sd-vae-ft-ema
[3]https://huggingface.co/stabilityai/sdxl-vae
[4]https://huggingface.co/stabilityai/sd-vae-ft-mse

The initial population is sampled from a zero-centered multivariate Gaussian distribution with standard deviation $\sigma = 0.8$, i.e., $z \sim \mathcal{N}(0, \sigma^2 I)$. This initialization promotes latent diversity while maintaining plausibility. The population size is fixed at 250 individuals, the maximum value possible respecting computational cost. Visually, decoded images from the initial population appear as noisy and unstructured patterns, consistent with their random latent initialization.

B. Fitness Function and Classifier Integration

The fitness of each individual is determined by its alignment with a predefined target class, as assessed by a frozen image classifier. Each latent vector z is decoded into an image using the VAE decoder. The image is then resized and normalized according to CIFAR-10 statistics to match the input expectations of the classifier.

The primary classifier used in our experiments is a Vision Transformer (ViT), specifically the `nateraw/vit-base-patch16-224-cifar10`[5] model, which was pre-trained and fine-tuned on CIFAR-10. This dataset comprises ten classes: airplane, automobile, bird, cat, deer, dog, frog, horse, ship, and truck. The model outputs a logit vector over these ten classes, and the fitness is computed as the softmax probability assigned to the target class $c^\star$:

$$f(z) = \text{softmax}(\ell)_{c^\star}, \tag{1}$$

where ℓ is the logit output of the classifier. In our setup, the target class corresponds to *ship* (CIFAR-10 index 8), though the system supports arbitrary class selection.

Visually, as evolution progresses, the best individuals in the final generation tend to resemble the item associated with the chosen target class. Even when the resemblance is not overtly clear, it is typically possible to identify recurring color patterns and/or structural hints that align with the semantic characteristics of the target class. In our tests, for example, targeting the *bird* class resulted in images that regularly featured: rounded grayscale shapes, as well as red beak-like shapes and a green background, whereas the *ship* class often displayed hull-like shapes and dominant shades of blue and gray. Figures 1 and 2 illustrate most of these patterns.

It is important to emphasize that the algorithm is not limited to this setting. We tested it successfully with simpler datasets such as MNIST, using different classifiers trained on grayscale digits. These adaptations required resizing decoded images to different resolutions and adjusting the preprocessing pipeline to the classifier's expected input format.

More broadly, the method is classifier-agnostic: any image classifier can be used, provided that appropriate adjustments are made for image resolution, input normalization, and decoder compatibility. The computational cost of evaluation naturally scales with the complexity of the chosen classifier and image resolution.

[5]https://huggingface.co/nateraw/vit-base-patch16-224-cifar10

Fig. 2: Figures generated by LEI for *boat* and *bird* classes respectively, using a CIFAR-10 classifier.

C. Genetic Operators

The evolutionary dynamics of LEI are driven by mutation and crossover operators applied directly in the latent space. These operations alter the latent vectors of individuals across generations in order to explore the space and progressively improve fitness.

Mutation is the primary source of variation in our framework. For each selected individual, Gaussian noise is added to its latent vector z to produce a mutated offspring z', following:

$$z' = z + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma_t^2 I), \tag{2}$$

where σ_t is the standard deviation at generation t, modulated via an annealing schedule. Specifically, σ_t decays linearly over generations to allow for broader exploration at the beginning of evolution and more refined local search, exploitation, toward the end. This approach prevents premature convergence while maintaining diversity in early stages.

Crossover is implemented as a linear interpolation between two parent vectors z_1 and z_2:

$$z' = \alpha z_1 + (1 - \alpha) z_2, \tag{3}$$

where $\alpha \in [0, 1]$ is sampled uniformly at random for each crossover operation.

The population undergoes **tournament selection**, where individuals are sampled in groups and the fittest are chosen for reproduction. **Elitism** is employed by directly copying the top k individuals to the next generation unchanged, ensuring that high-quality solutions are preserved throughout the evolutionary process.

D. Hyperparameters and Evolutionary Configuration

The LEI algorithm relies on a compact yet effective set of hyperparameters to control the evolutionary process. These parameters govern aspects such as population diversity, convergence rate, and exploitation of high-fitness individuals.

The population size is fixed at 250 individuals per generation, and evolution proceeds for 350 generations. A mutation probability of 0.95 is applied, while crossover probability is 0.05. Elitism is enforced by directly copying the single best

(a) Gen 1; Fitness: 0.142 (b) Gen 50; Fitness: 0.987 (c) Gen 150; Fitness: 0.994 (d) Gen 350; Fitness: 0.996

Fig. 3: Evolution of the best LEI individual for the target class *ship* across generations, transitioning from near-random noise to a visually ship-shaped outline, displaying the genetic algorithm's capacity to generate semantically meaningful images.

individual from each generation to the next, preserving top-performing solutions.

Interestingly, turning off elitism resulted in unstable convergence and poor final performance, highlighting the importance of preserving top individuals. Setting elitism to one was sufficient to ensure consistent improvement and yielded better results than higher elitism values, which tended to reduce diversity prematurely.

The mutation intensity is modulated through an annealing schedule: the perturbation standard deviation σ_t starts at a maximum value of 0.75 and decays linearly to 10% of that value over the course of evolution. This schedule encourages broad exploration early on and finer local search, exploitation, in later generations.

All hyperparameters were initially selected based on the guidelines from [4], then adjusted through preliminary testing. The final configuration follows standard practices in evolutionary computation, aiming to balance computational cost and search effectiveness.

E. Reproducibility and Considerations

To ensure that results are reliable and experiments can be replicated, the entire LEI pipeline is configured with a retrievable random seed. This deterministic setup guarantees that population initialization, selection, mutation, and evaluation produce identical outputs across runs, provided the same computational environment is used.

In addition to the algorithm hyperparameter current configuration, the framework stores key outputs for post-analysis, including all final populations, intermediate best individuals, and corresponding decoded images and fitness scores. This enabled both qualitative inspection of evolutionary trajectories and quantitative assessment of performance.

To further facilitate reproducibility, the full source code of the LEI framework has been publicly released; the GitHub repository URL is provided as a footnote on the first page of this paper.

IV. Experiments

In this section, we analyze the results of our algorithm from multiple perspectives. First, we present a qualitative inspection of the images generated by LEI, highlighting how visual features evolve over time and how human observers can interpret these features to understand why certain samples are classified in specific ways. This sheds light on the internal logic of the classifier being audited.

Next, we investigate the embeddings corresponding to the generated images. By analyzing how these latent vectors shift during evolution, we aim to understand the classifier's sensitivity to specific regions of the latent space, revealing patterns in how the model encodes semantic concepts and potentially identifying biases or blind spots.

Finally, we explore how insights gained from both visual and embedding analyses can be leveraged to craft adversarial examples. Specifically, we demonstrate that LEI can be used not only for interpretability but also to expose vulnerabilities in the classifier, generating inputs that trigger high-confidence predictions despite being perceptually ambiguous or semantically inconsistent. This highlights the dual role of our method: both as a diagnostic tool for model auditing and as a mechanism for stress-testing the robustness of black-box classifiers.

A. Image analysis

One of the central contributions of LEI is its capacity to produce interpretable synthetic images that reflect the classifier's internal representation of a given class. Figure 3 shows a selection of individuals decoded from latent vectors at different stages of the evolutionary process targeting the 'ship' class.

At generation 1 (Figure 3a), the best individual is still indistinguishable from structured noise. By generation 50 (Figure 3b), however, LEI has already produced an image that the classifier labels as *ship* with **98.7%** confidence, even though human observers perceive little more than a muted beige-gray canvas featuring faint angular patches and a darker smudge near the centre. This mismatch is our first indication of bias: the model seems to equate "dull metallic textures and oblique edge fragments" with the presence of a ship, a shortcut likely inherited from training images where steel hulls dominate the scene rather than the vessel's overall outline.

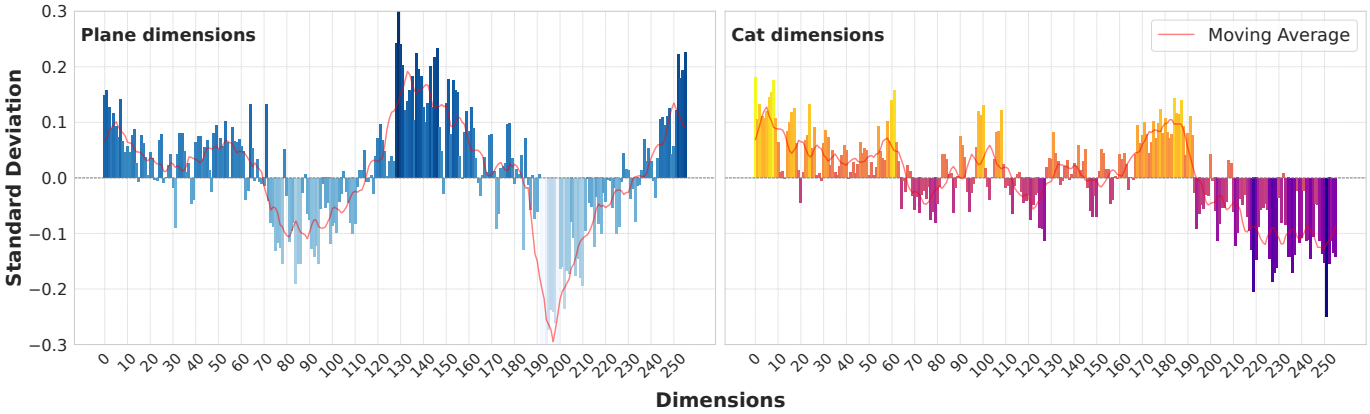


Fig. 4: Mean per-dimension variation of latent embeddings for plane and cat classes.

The next 300 generations refine these low-level cues into a coherent silhouette (Figures 3c–3d). By generation 350 the image shows an elongated hull, deck-like layers, and muted marine colours that humans readily interpret as a ship. Crucially, the *waterline* and *sky–sea colour gradient* remain exaggerated, suggesting that the classifier's notion of "shipness" relies more on the co-occurrence of blue hues and horizontal boundaries than on actual hull geometry.

This observation raises a question: if the model assigns such high confidence to what still lacks clear semantic structure to human eyes, what exactly is it detecting? To understand this phenomenon, we identify what we term a turning point, the moment during evolution when the classifier abruptly shifts from low to high confidence, often in response to subtle, visually imperceptible changes.

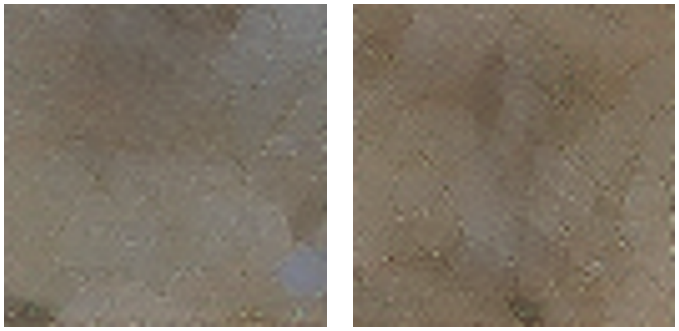
Figure 5 illustrates this turning point for the *airplane* class by comparing two consecutive generations of the same individual. Visually, the images from Generation 11 and Generation 12 appear nearly identical: both show abstract textures with muted tones and no recognizable shape. Gen 12 is perhaps slightly sharper or warmer in tone, but no distinct airplane features are discernible to the human eye. Yet, this small latent modification causes the classifier's confidence to leap from just **19.6%** to **81.9%**, a dramatic increase triggered by a seemingly subtle latent shift. This sharp transition strongly suggests that the model is responding not to global structure or class-defining shapes, but rather to minute texture cues, color blobs, or frequency patterns – features it learned during training that correlate with the label "airplane."

This insight reinforces a critical point: the classifier's internal notion of class identity does not always align with human perception. It is often dominated by superficial, low-level features that act as shortcuts or indirect cues for semantic understanding. LEI helps uncover these shortcuts by evolving latent vectors that lead to confident predictions, even when the resulting images lack clear semantic content.

To better understand the mechanics of these jumps, we now shift our focus from the visual domain to the latent space. Since images that look almost identical can produce radically different outputs, it becomes essential to analyze how small perturbations in latent vectors propagate through the classifier. This analysis offers a window into the classifier's implicit feature space and highlights which latent directions are most sensitive to class activation, providing a powerful auditing mechanism for black-box models.

B. Embedding Analysis

Now that we have identified a bias in the model toward specific latent representations, we can begin to unpack how small changes, like those shown in Figure 5, can lead to such drastic shifts in classification confidence. To better understand this behavior, we ran LEI across 1000 independent runs with varying random seeds, collecting latent embeddings from both the initial generation (before convergence) and the final generation (after convergence at generation 350). We chose 1 000 runs because it drives the standard error of the mean latent embedding to below 0.05 per dimension, ensuring that our aggregated results reflect genuine algorithmic behavior rather than sampling noise.



(a) Gen 11; Fitness **0.196** (b) Gen 12; Fitness **0.819**

Fig. 5: Classifier turning point in latent evolution: confidence rises from 19.6% to 81.9% between only two consecutive generations, while targeting the *airplane* class.

This extensive sampling allowed us to compute the median difference between initial and final embeddings across runs. For this analysis, we selected two target classes – *airplane*

and *cat* – to study how different semantic categories behave in latent space, and to later support both the interpretation of our airplane example and our adversarial attack strategy.

Figure 4 presents these results: the left panel shows the average dimensional shift for the *airplane* class, while the right panel shows the same for the *cat* class.

Starting with the *airplane*, we observe that certain latent dimensions strongly influence the classifier's output. For instance, in dimensions 120–150, higher embedding values are associated with increased model confidence, whereas in dimensions 180–210, the opposite trend appears: lower values correlate with higher scores. This indicates that the model is biased toward specific directions in latent space, treating some features as strong indicators of the class and others as suppressors.

A similar pattern emerges with the *cat* class. Here, the lower dimensions (0–60) tend to have a positive contribution when their values are high, while higher dimensions (190+) need to remain small for the model to assign a high confidence score. These consistent, directional shifts across runs reveal how the classifier relies on specific subsets of the latent space to make its decisions, reinforcing the idea that it has internalized shortcut heuristics that LEI is uniquely capable of exposing.

With this, we can return to the example presented in the image analysis (Figure 5), where a visually subtle change resulted in a dramatic increase in the model's confidence. When we inspect the latent embeddings of the two generations involved, a particularly revealing pattern emerges – the most significant positive shift occurs at dimension 254, with an increase of +0.34, while the largest negative shift happens at dimension 205, with a decrease of -0.40.

These shifts are far from random, they align precisely with the trends identified in our large-scale embedding analysis for the *airplane* class. As previously noted, increasing the value of dimension 254 boosts the model's confidence, while decreasing the value of dimension 205 has a similar positive effect on the predicted score. The fact that these are the dimensions that change most between Generation 11 and Generation 12 demonstrates how LEI captures the model's internal logic: rather than requiring drastic visual changes, the classifier is highly sensitive to specific latent activations that correspond to its learned heuristics.

This alignment between micro-scale changes (single-image evolution) and macro-scale trends (across 1000 runs) validates the interpretability power of our method. LEI not only reveals that the model is biased toward particular directions in latent space, but also pinpoints exactly which dimensions are responsible for triggering confident predictions, even when those changes are imperceptible to humans. This capability is critical for auditing black-box models and understanding how their internal representations connect to external decisions.

C. Adversarial Attack

Beyond helping us interpret what the model is doing, LEI can also be used to create adversarial examples – images that trick the model into making high-confidence predictions, even if the image looks strange or unclear to humans. Since LEI works in the latent space of a generative model, the changes it makes are small, structured, and stay within the kind of images the decoder can produce. This means the model can be strongly influenced by images that don't clearly look like the predicted class to us.

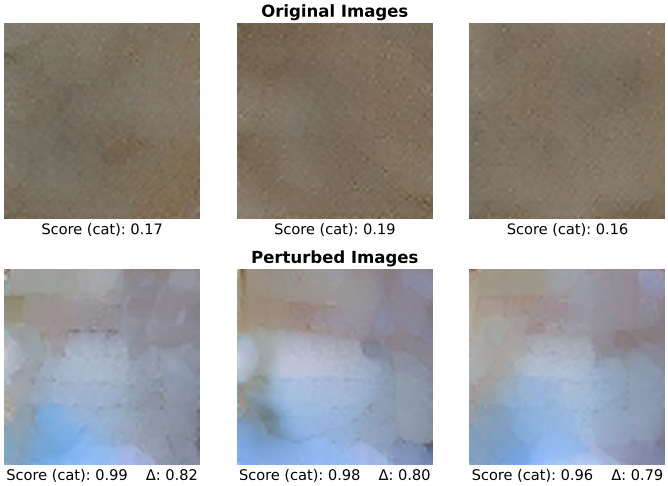


Fig. 6: Top 3 latent samples with the highest confidence increase after disturbance targeting the *cat* class. Each column shows the original random image (top) and its perturbed counterpart (bottom), along with their respective scores.

To test this adversarial potential, we calculated the average direction in latent space that LEI follows when evolving images toward a specific class, demonstrated in the last subsection. This "direction" works like a fingerprint: when we add it to a new latent vector, it tends to push the model to believe the image belongs to that class. By applying this shift to many random latent vectors, we can see whether it consistently fools the model, essentially checking if it works as a general adversarial attack in latent space.

We apply this learned shift, scaled by a fixed amount, to each latent vector and re-run the classifier. By comparing the model's confidence before and after the change, we measure how effective this attack is. This also gives us visual examples of when the attack works and when it doesn't, helping us better understand the model's vulnerabilities.

Figure 6 illustrates this phenomenon in practice, presenting three examples where the model's confidence increased the most after applying the learned perturbation. In some cases, the confidence increased by as much as 0.82. Yet, despite this dramatic shift, the resulting images remain nonrepresentational, abstract textures and color patterns with no discernible structure or semantic content. From a human perspective, the classifier's certainty appears unfathomable, as nothing in the perturbed outputs resembles a cat or any coherent object at all.

Both the original and modified images lack any obvious shape or structure that we would associate with the class. Still, the model becomes almost certain they represent a cat. This

shows how easily the model can be influenced by patterns that don't make sense to humans, highlighting the gap between what the model sees and what we understand.

Fig. 7: Latent adversarial shift on a structured image.
Left: original Airbus A380 photograph.
Centre: VAE reconstruction, labelled *airplane* by classifier.
Right: Adding "cat-direction" to tensor, now labelled as *cat*.

These results suggest that the learned perturbation doesn't create clearer or more meaningful visual features. Instead, it seems to activate hidden signals that the model has learned to associate with the target class – signals that are probably statistical rather than visual or semantic. The model responds with high confidence even when there are no recognizable features in the image. This shows that this specific classifier can be manipulated through changes in latent space that don't make sense to humans.

To see if this effect goes beyond random noise, we ran a specific test using a latent code that decodes into a clear, structured image of an airplane. This test is shown in Figure 7.

Visually, the image displays recognizable aircraft features, such as a central fuselage and wing-like extensions, and is assigned a high score for the *airplane* class. We then apply the same class-specific perturbation computed for the *cat* class to this latent code and decode the result. The perturbed image retains much of the original airplane-like structure: its silhouette, layout, and spatial features are still largely intact. However, the classifier's output undergoes a dramatic shift: confidence in the airplane class collapses, and the new image is classified with high confidence as a *cat*.

This result supports what we saw earlier: LEI's perturbation affects how the model "sees" the image internally, not just how it looks to us. More importantly, it shows that this effect works even on clear and structured images, not just on noisy or abstract ones. In this case, the model was convinced that an airplane was a cat, not because the image changed in any meaningful way to human eyes, but because a subtle shift in the latent space triggered internal features the model strongly associates with the *cat* class.

Rather than simply generating adversarial noise, LEI learns how the model makes decisions – what latent directions it favors, which features it treats as meaningful, and how those biases affect predictions. It then uses this knowledge to expose and exploit the model's internal shortcuts. These findings show that LEI is not just a generative tool, but a diagnostic method capable of revealing latent biases and fragilities in black-box models. By navigating the model's own internal logic, LEI helps us better understand, audit, and ultimately improve the trustworthiness of machine learning systems.

V. Discussion

LEI directly extends our EBIE lineage: it is a genetic algorithm that evolves latent vectors from a pre-trained VAE to generate synthetic images increasingly aligned with a classifier's internal representation. Operating without labels, gradients, or training data, it offers a flexible and dataless method for probing black-box classifiers through semantically meaningful generations.

It is a demonstration that evolutionary optimization in latent space can effectively probe black-box classifiers, even without labeled data or gradient access. By evolving samples that maximize the classifier's confidence in a target class, the method uncovers semantic cues the model relies on, and produces interpretable outputs aligned with its internal logic.

Our analysis reveals that evolution concentrates on a sparse subset of latent dimensions, suggesting that class-specific information is encoded along well-defined axes in the VAE's bottleneck. These findings reinforce the utility of latent space navigation for interpretability.

Beyond interpretability, LEI exposes vulnerabilities: learned latent perturbations can significantly boost class confidence without corresponding semantic content, highlighting risks of overreliance on superficial patterns. In some cases, even capable of changing classification of a previously very high confidence score, as shown through adversarial attack.

Limitations. LEI depends on the structure and quality of the chosen generative model. Poorly trained VAEs or latent spaces with limited semantic coherence may hinder evolution. Moreover, both genetic algorithms and image generation are computationally intensive processes. Scaling to higher resolutions or more complex classifiers can significantly increase runtime and resource demands.

Future work. Possible extensions include testing with richer generative models (e.g., diffusion-based), integrating concept attribution into the search, and applying LEI to real-world auditing tasks like fairness or safety analysis. Additionally, the general framework could be adapted to other data modalities – such as audio, video, or tabular representations – given appropriate generative and classification models.

Overall, LEI offers a simple, model-agnostic approach for exploring classifier behavior through interpretable, dataless latent evolution.

References

[1] H. Javed, S. El-Sappagh, and T. Abuhmed, "Robustness in deep learning models for medical diagnostics: security and adversarial challenges towards robust ai applications," *Artificial Intelligence Review*, vol. 58, no. 1, p. 12, 2024. [Online]. Available: https://doi.org/10.1007/s10462-024-11005-9

[2] P. S. Chib and P. Singh, "Recent advancements in end-to-end autonomous driving using deep learning: A survey," 2023. [Online]. Available: https://arxiv.org/abs/2307.04370

[3] K. Müller, H. Koelmann, M. Niemann, R. Plattfaut, and J. Becker, "Exploring audience's attitudes towards machine learning-based automation in comment moderation," in *Wirtschaftsinformatik 2022 Proceedings*, 17. Internationale Tagung Wirtschaftsinformatik, 2022, p. 1. [Online]. Available: https://aisel.aisnet.org/wi2022/human_rights/human_rights/1

[4] A. Buzelin, Y. Aquino, V. Estanislau, P. Bento, L. Dayrell, S. Malaquias, C. Santana, G. H. G. Evangelista, C. Grossi, P. B. Rigueira, L. G. Porfirio, M. S. Locatelli, W. Meira Jr., and G. L. Pappa, "Evolutionary bias identification with embeddings," in *Applications of Evolutionary Computation: 28th European Conference, EvoApplications 2025, Held as Part of EvoStar 2025, Trieste, Italy, April 23–25, 2025, Proceedings, Part II*. Berlin, Heidelberg: Springer-Verlag, 2025, p. 337–352. [Online]. Available: https://doi.org/10.1007/978-3-031-90065-5_21

[5] J. Buolamwini and T. Gebru, "Gender shades: Intersectional accuracy disparities in commercial gender classification," in *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, ser. Proceedings of Machine Learning Research, S. A. Friedler and C. Wilson, Eds., vol. 81. PMLR, 23–24 Feb 2018, pp. 77–91. [Online]. Available: https://proceedings.mlr.press/v81/buolamwini18a.html

[6] B. Wilson, J. Hoffman, and J. Morgenstern, "Predictive inequity in object detection," 2019. [Online]. Available: https://arxiv.org/abs/1902.11097

[7] A. J. Larrazabal, N. Nieto, V. Peterson, D. H. Milone, and E. Ferrante, "Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis," *Proceedings of the National Academy of Sciences*, vol. 117, no. 23, pp. 12 592–12 594, 2020. [Online]. Available: https://www.pnas.org/doi/abs/10.1073/pnas.1919012117

[8] A. Bărbălău, A. Cosma, R. T. Ionescu, and M. Popescu, "Black-box ripper: Copying black-box models using generative evolutionary algorithms," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. [Online]. Available: https://github.com/antoniobarbalau/black-box-ripper

[9] F. Lomurno and M. Matteucci, "Synthetic image learning: Preserving performance and preventing membership inference attacks," 2024. [Online]. Available: https://arxiv.org/abs/2407.15526

[10] O. Rotem, T. Schwartz, R. Maor, Y. Tauber, M. Tsarfati Shapiro, M. Meseguer, D. Gilboa, D. S. Seidman, and A. Zaritsky, "Visual interpretability of image-based classification models by generative latent space disentanglement applied to in vitro fertilization," *Nature Communications*, vol. 15, no. 1, p. 7390, 2024.

[11] A. Voynov and A. Babenko, "Unsupervised discovery of interpretable directions in the gan latent space," in *Proceedings of the 37th International Conference on Machine Learning (ICML)*, ser. Proceedings of Machine Learning Research, vol. 119. PMLR, 2020, pp. 9786–9796. [Online]. Available: http://proceedings.mlr.press/v119/voynov20a.html

[12] A. Ciortea, V. Puscas, V. Manta, and R. T. Ionescu, "Generating adversarial examples through latent space exploration of gans," in *Proceedings of the 18th International Conference on Computer Vision Theory and Applications (VISAPP)*. SciTePress, 2023, pp. 119–128.

[13] G. Bernat, A. Gaier, H. Touvron, M. Fontaine, and J. Lehman, "Evocraftgan: Quality-diversity search in the latent space of gans," 2024. [Online]. Available: https://arxiv.org/abs/2412.05842

[14] A. Ghanem and A. Tawfik, "A genetic algorithm for optimizing the latent space of a generative adversarial network for image generation," *Applied Sciences*, vol. 15, no. 10, p. 5228, 2025. [Online]. Available: https://www.mdpi.com/2076-3417/15/10/5228

[15] G. Iglesias, M. Zamorano, and F. Sarro, "Search-based negative prompt optimisation fornbsp;text-to-image generation," in *Artificial Intelligence in Music, Sound, Art and Design: 14th International Conference, EvoMUSART 2025, Held as Part of EvoStar 2025, Trieste, Italy, April 23–25, 2025, Proceedings*. Berlin, Heidelberg: Springer-Verlag, 2025, p. 94–110. [Online]. Available: https://doi.org/10.1007/978-3-031-90167-6_7

[16] A. Lambora, K. Gupta, and K. Chopra, "Genetic algorithm- a literature review," in *2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon)*, 2019, pp. 380–384.

[17] C. Doersch, "Tutorial on variational autoencoders," 2021. [Online]. Available: https://arxiv.org/abs/1606.05908