Teaching Humans Subtle Differences with *DIFF*usion

Anonymous Author(s)
Affiliation
Address
email

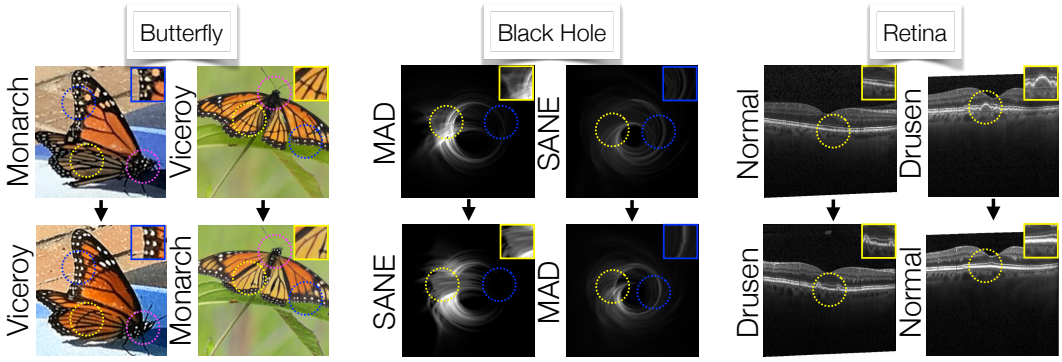


Figure 1: ***DIFF*usion Counterfactuals.** We illustrate the counterfactual results from our methods on the Butterfly dataset, the Black Hole dataset, and the Retina dataset. In the Butterfly dataset, the Viceroy has a cross-sectional line (**yellow**), a smaller head with less dots (**magenta**), and more "scaley" dots (**blue**), compared to the Monarch. In the Black Hole dataset, SANE has more uniform wisps (**yellow**) and less of a prominent photon ring (**blue**) as compared to MAD, with these distinguishing features discovered through our method rather than known a priori. In the Retina dataset, normal retinas lack the horizontal line bumps (**yellow**) present in retinas with drusen.

Abstract

Scientific expertise often requires recognizing subtle visual differences that remain challenging to articulate even for domain experts. We present a system that leverages generative models to automatically discover and visualize minimal discriminative features between categories while preserving instance identity. Our method generates counterfactual visualizations with subtle, targeted transformations between classes, performing well even in domains where data is sparse, examples are unpaired, and category boundaries resist verbal description. Experiments across six domains, including black hole simulations, butterfly taxonomy, and medical imaging, demonstrate accurate transitions with limited training data, highlighting both established discriminative features and novel subtle distinctions that measurably improved category differentiation. User studies confirm our generated counterfactuals significantly outperform traditional approaches in teaching people to correctly differentiate between fine-grained classes, showing the potential of generative models to advance human visual learning and scientific research.

1 Introduction

Generative models, especially large-scale image diffusion models, have transformed text-to-image creation, opening new ways to visualize concepts across various domains. While these models

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

excel in everyday contexts with clear category distinctions, a far more challenging frontier exists in scientific fields where visual differences between categories are so subtle that they often remain unknown and unidentified even to domain experts.

In specialized scientific domains, the complete set of visual features distinguishing between categories may be partially or entirely undiscovered. For example, astronomers studying black hole simulations have no established verbal characteristics to differentiate MAD from SANE models because these distinguishing features have not yet been comprehensively identified. Entomologists may differentiate Viceroy and Monarch butterflies through the Viceroy's characteristic cross-sectional black line, yet may miss other distinguishing features that could further help the differentiation. This represents the fundamental challenge for visual expertise training: how do we teach recognition of patterns we ourselves don't fully understand?

One of the most effective ways to reveal subtle category differences is to transform an image and rapidly flip between the original and its altered version. In scientific domains, using generative models for such targeted image editing faces three key challenges: (1) automatically identifying discriminative features that may not be known or easily articulated even by experts, (2) limiting changes exclusively to these category-defining features, and (3) preserving all other identity characteristics of the instance. We develop a system that combines state-of-the-art image editing techniques with visual algebraic conditioning guidance to address these challenges in data-scarce scientific domains. Our approach automatically identifies discriminative features through visual algebraic operations that extract category-specific information without requiring explicit articulation. By integrating inverted noise maps (z) to preserve identity features with conditioning vectors (c) that guide category transformations, our system achieves effective identity-preserving yet category-changing results, that isolate and visualize subtle differences between scientific categories.

Our approach overcomes limitations in current counterfactual visualization methods, which have traditionally been applied in domains where category distinctions are already well-understood and easily verbalized. Text-guided editing methods rely on linguistic descriptions, which can be too ambiguous to specify desired visual changes. Methods like Concept Sliders' [14] effectiveness, which is guided by the image distributions themselves, depend on paired examples in most cases-a constraint limiting their use in teaching scenarios. Visual counterfactual generation methods often rely on gradients from a classifier, a limitation when data is scarce. Classifier-free alternatives, like TIME [24], struggle with image quality and coherence for subtle differences.

Through experiments across six domains, we demonstrate our approach's effectiveness in highlighting visual differences between categories. For instance, in black hole simulations, where distinguishing characteristics between MAD and SANE models remain largely unknown, our counterfactual visualizations emphasize distinct visual patterns in the image distribution. The transformations draw attention to variations in the uniformity of wisps and prominence of the photon ring, which are features that black hole experts themselves had not identified.

User studies confirm the effectiveness of our approach: participants who trained with our counterfactual visualizations demonstrated significantly better category differentiation performance than those using traditional approaches with unpaired images. This validates that our method highlights meaningful visual patterns that can be used to build expertise, even when those subtle patterns have not yet been explicitly identified or understood.

2 Related Work

Visual Counterfactual Explanations. A counterfactual image shows how an input would appear if altered to switch its class, enhancing interpretability. Counterfactual inference crafts images that not only differ in classification but also clarify the visual features defining each distribution. Approaches for visual counterfactual explanations (VCEs) make use of generative model edits, with VAEs [41], GANs [29], and more recently, diffusion-based methods [22–24, 3, 47, 12]. Most diffusion-based approaches adapt classifier guidance [10] to steer the generative process of counterfactuals, requiring access to the classifier and test-time optimization to produce counterfactual images. However, generating counterfactuals this way can be challenging, as the optimization problem closely resembles that of adversarial examples. TIME [24] proposes an alternative approach by using Textual Inversion [13] to encode class and dataset contexts into a set of text embeddings, providing a black-box framework for counterfactual explanations. While this removes the need for direct classifier access, Textual

Inversion is primarily designed for personalization, focusing on regenerating concepts in novel scenes rather than preserving image structure-an essential aspect of counterfactual generation.

Image Editing. Recent advances in text-to-image diffusion models [39, 42, 44, 36, 28] have enabled test-time controls for image editing, ranging from semantic modifications to attention-based edits and latent space manipulation. Early approaches, such as SDEdit [34], applied noise to an image and then denoised it using a new prompt, but this often resulted in significant structural changes. Later methods refined direct prompt modifications by incorporating cross-attention manipulations or masking to better preserve image structure [17, 38, 4, 50, 9]. Brooks et al. [5] use controlled edits from these methods to train a new diffusion model based on instruction-driven prompts. However, these approaches are limited to text-driven modifications, which restrict the flexibility of edits beyond what can be described with text. Unlike single-image editing methods, Concept Sliders [14] introduce a different approach by optimizing a global semantic direction across the diffusion model. While text pairs can guide their optimization, they also propose visual sliders based on image pairs. However, the visual slider approach struggles with unpaired data.

Diffusion Models with Image Prompts. Text-to-image diffusion models generate images from text prompts, but text often falls short in capturing nuanced concepts. Image prompts offer a richer alternative, conveying nuanced details more effectively, as "a picture is worth a thousand words." DALL-E 2 [39] pioneered this by conditioning a diffusion decoder on CLIP image embeddings, aided by a diffusion prior for text mapping. Later works offer different architectures [40] or adapt text-to-image models for image prompts [56, 2, 28, 15].

Diffusion Inversion. Editing a real image typically requires first obtaining a latent representation that can be fed into the model for reconstruction. This latent representation can then be modified, either directly or by altering the generative process, to produce the desired edit. Most diffusion-based inversion methods rely on the DDIM [48] sampling scheme, which provides a deterministic mapping from a noise map to a generated image [35, 52, 38]. However, this approach introduces small errors at each diffusion step, which can accumulate into significant deviations, particularly when using classifier-free guidance [18]. Instead of predicting an initial noise map that reconstructs the image through deterministic sampling, an alternative approach considers DDPM [19] sampling and inverts the image into intermediate noise maps [54]. Building on this, Huberman-Spiegelglas et al. [21] proposed an inversion technique for the DDPM sampler, along with an edit-friendly noise space better suited for editing applications. We use this technique while conditioning on image prompts.

Machine Teaching. Machine teaching optimizes human learning via computational models. Early work framed this as an optimization task, minimizing example sets for efficient teaching [58]. Generally, the field of machine learning for discovery has machine teaching as a goal [26, 6]. Recent advances leverage generative models and LLMs for cross-modal discovery, synthesizing representations for conceptual learning [7], decoding structures in mathematics, or programs for scientific discovery [33, 43]. Parallel efforts amplify subtle signals for perception: language models detect fine-grained textual differences [11], while video motion magnification enhances visual cues [31, 55, 37]. These methods, though effective for fine-grained discrimination, typically require aligned, abundant data and focus on single modalities. Our work extends these efforts, using diffusion models to generate visual counterfactuals for nuanced category learning.

3 Method

We begin by introducing *DIFF*usion for counterfactual image generation, as illustrated in Figure 2. In Section 3.1, we provide the necessary background on diffusion models. In Section 3.2, we present our proposed method, outlining its design and implementation.

3.1 Diffusion Preliminaries

Diffusion models generate data by sampling from a distribution through iterative denoising of noisy intermediate vectors. A forward process is first applied, where noise is gradually added to a clean image x_0 over T steps. A noisy sample at timestep t can be expressed as

$$x_t = \sqrt{\bar{\alpha}_t} x_0 + \sqrt{1 - \bar{\alpha}_t}\epsilon, \quad t = 1, ..., T \tag{1}$$

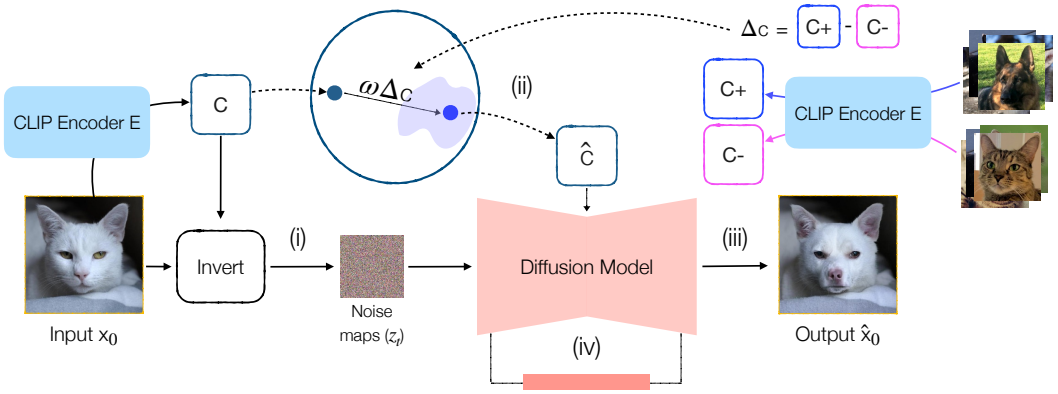


Figure 2: ***DIFF*usion method.** Our method consists of four parts. (i) Inverting the real image with DDPM-EF to obtain noise maps. (ii) Performing conditioning space arithmetic using positive and negative embeddings obtained from the training set. (iii) Generation via diffusion sampling, starting from the inverted noise conditioning on the manipulated conditioning vector $\hat{c}$. (iv) Optional domain tuning, in which we fine-tune the diffusion model for domain adaptation.

where $\epsilon \sim \mathcal{N}(0, \mathbf{I})$, α_t is a predetermined variance schedule, and $\bar{\alpha}_t = \prod_{i=1}^{T} \alpha_i$. The model learns to reverse the forward noising process, which can be expressed as an update step over x_t,

$$x_{t-1} = \mu_\theta(x_t, c) + \sigma_t z_t, \quad t = T, ..., 1 \tag{2}$$

where z_t are i.i.d standard normal vectors, σ_t is a variance schedule, and $\mu_\theta(x_t, c)$ is typically parameterized as:

$$\mu_\theta(x_t, c) = \frac{1}{\sqrt{\alpha_t}} \left(x_t - \frac{1 - \alpha_t}{\sqrt{1 - \bar{\alpha}_t}} \epsilon_\theta(x_t, t, c) \right) \tag{3}$$

Here $\epsilon_\theta(x_t, t, c)$ is the trained noise prediction network, and c is an optional conditioning context, such as an image prompt embedding.

3.2 DIFFusion

Given an input image x_0, our goal is to find a fine-grained, discriminative edit that changes a classifier's prediction. Let $\mathcal{R}_\theta(\mathbf{z}, c)$ be the recursive application of the denoising diffusion model from Equation 2. Our approach finds these edits by inverting the image x_0, into a sequence of noise maps, $\mathbf{z}$, and manipulating the CLIP embeddings of the original image, $c = E(x)$, into a resulting conditioning vector $\hat{c}$, before sampling the modified image. We generate the modified image $\hat{x}_0$ through:

$$\hat{x}_0 = \mathcal{R}_\theta(\mathbf{z}, \hat{c}) \tag{4}$$

Since the diffusion model must generate an image consistent with the original noise maps $\mathbf{z}$, and has a conditioning vector $\hat{c}$ that steers from the source towards the target class, the resulting samples maintain the identity of the original image, but with subtle modifications such that the class label flips.

Inversion. We are interested in extracting noise vectors $\mathbf{z}$, such that, if used in Equation 2, would recover the original image x_0. Note that any sequence of $T + 1$ images $x_0, ..., x_T$ can be used to extract consistent noise maps for reconstruction by isolating z_t from Equation 2 as

$$z_t = \frac{x_{t-1} - \mu_\theta(x_t, c)}{\sigma_t}, \quad t = T, ..., 1 \tag{5}$$

We follow the choice suggested in [21] and compute the noise maps through the standard forward diffusion process Equation 1, but using statistically independently-sampled noise for each timestep. This yields noise maps $\mathbf{z} = \{x_T, z_T, \ldots, z_1\}$ that are consistent with x_0.

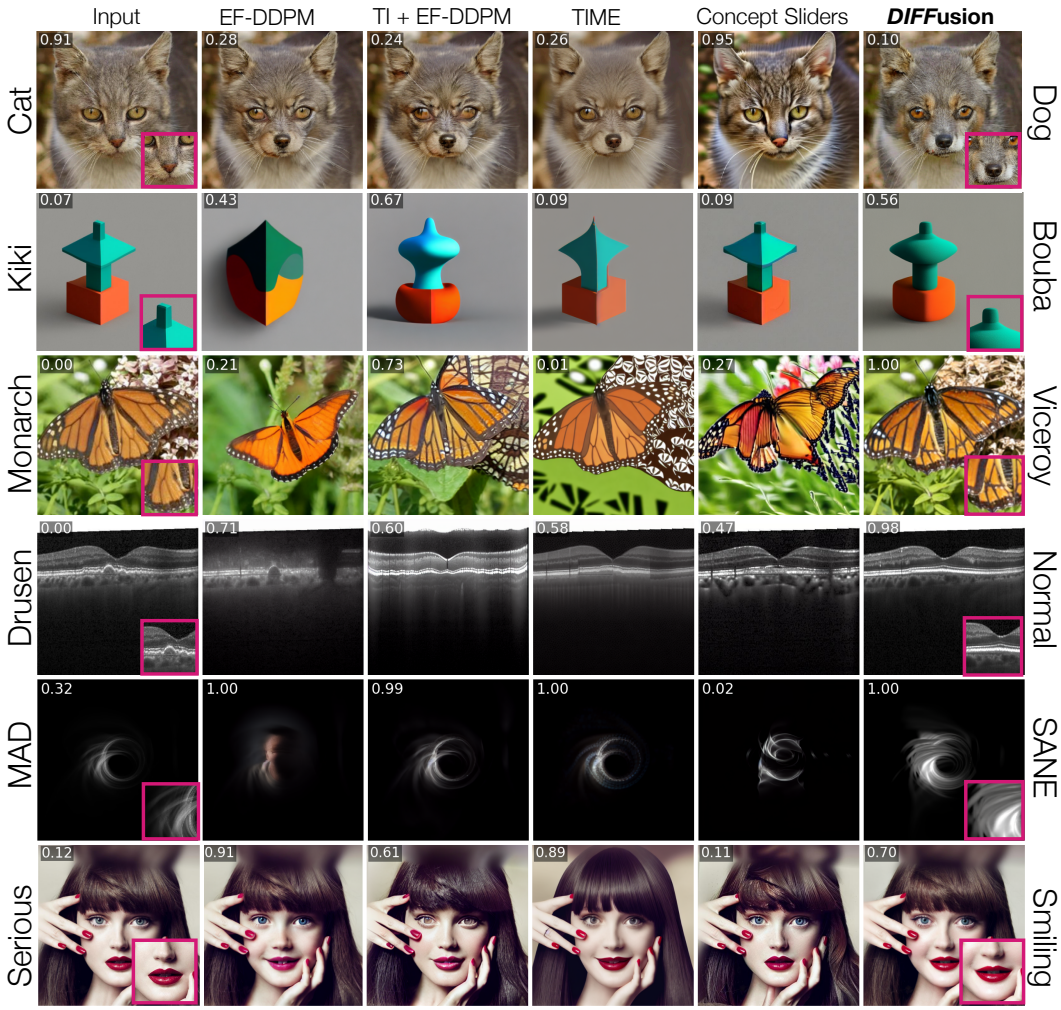


Figure 3: **Qualitative Results.** We present our qualitative results, where each row corresponds to one direction of our binary datasets. The first column contains the inputs, and each subsequent column contains the results from each baseline, with the last column containing the result from ***DIFF*usion**. In particular, the magnified boxes in the magenta frame show that our method is able to pick up on small discriminative cues. For example, when converting from MAD to SANE, the whisps become amplified and more uniform in brightness, and when converting from Monarch to Viceroy, a cross-sectional line is added on the wing. **Note:** The value in the top left corner of each image represents the probability predicted by the oracle classifier, as explained in Section 4.2. Values below 0.5 are classified as class 0, and above as class 1. Table 1 contains the class names and order.

Conditioning. We generate edits that flip the category through arithmetic operations on c, resulting in $\hat{c}$. We apply an additive translation to the conditioning vector $c = E(x)$:

$$\hat{c} = c + \omega \Delta c \tag{6}$$

where c is the CLIP image embedding of the original image, Δc is a direction that moves the class from the original class to the target class, and ω is a scaler that varies the direction's strength. We calculate this translation through the difference of means for each class:

$$\Delta c = \mathbb{E}_{x_p}\left[E(x_p)\right] - \mathbb{E}_{x_n}\left[E(x_n)\right] \tag{7}$$

such that x_p is an image of class p and x_n is an image of class n (e.g, positive and negative classes). We normalize all the image embeddings with L2 norm prior to the arithmetic.

Sampling. We use $\hat{c}$ as the conditioning vector for DDPM sampling, paired with the inverted noise maps, z, to generate the counterfactual image. As suggested in [21], we run the generation process

starting from timestep $T - T_{skip}$, where T_{skip} is a parameter controlling the resemblance to the input image. Therefore, similar to Equation 2, denoting the denoised edited image at timestep t as $\hat{x}_t$ we have,

$$\hat{x}_{t-1} = \mu_\theta(\hat{x}_t, \hat{c}) + \sigma_t z_t, \quad t = T - T_{skip}, ..., 1 \tag{8}$$

This approach allows us to systematically steer the image generation toward the target class by adjusting the manipulation scale ω, while maintaining key structural features of the original image through T_{skip}. Intuitively, a larger T_{skip} results in fewer denoising steps under the manipulated condition $\hat{c}$, leading to greater adherence to the input image.

Domain Tuning We use a pre-trained diffusion model [46] that conditions on CLIP image embeddings. When adapting to a new domain, we fine-tune the model using LoRA [20], training only its cross-attention and corresponding projection layers. As discussed in B.2, we find that domain tuning is beneficial for the Butterfly [51] and Retina [27] datasets, but has minimal impact on the other datasets.

Implementation Details. For inversion, we adapt the edit-friendly DDPM inversion scheme [21] to our diffusion decoder [46]. Specifically, we use CFG [18] in both inversion and generation. We first aim to find guidance scale parameters that achieve perfect reconstruction, and then use these guidance scales for our method. This process is further discussed in B.3. To generate counterfactuals, we manipulate the conditioning space using Equation 6, adjusting the manipulation guidance scale per dataset ($\omega = 1.0$ for AFHQ, $\omega = 2.0$ for the rest of the datasets). We then sample for $T - T_{skip}$ steps, where $T = 100$ and the choice of the T_{skip} parameter is further discussed in Section 4.2.

4 Experiments

4.1 Datasets and Baselines

Datasets. We quantitatively benchmark on datasets from diverse domains. We also note the corresponding directions under examination for each dataset in Table 1. We evaluate on AFHQ [8], CelebaHQ [30] and KikiBouba [1] as our non-scientific datasets. We also evaluate on three scientific datasets. The first is Retina [27], a dataset of retina cross-sections, both diseased and healthy. The second is Black Holes, which is a dataset of images taken from fluid simulations of accretion flows around a black hole [53]. The simulations assume general relativistic magnetohydrodynamics (GRMHD) under one of two regimes: magnetically arrested (MAD) or standard and normal evolution (SANE) [25]. Finally, we also evaluate on Monarch and Viceroy, a fine-grained species classification task. Monarch butterflies evolved to be mimics of Viceroys, and the two species are notoriously difficult to tell apart.

Table 1: Datasets and their classification tasks.

Dataset	Class 0 / Class 1
AFHQ [8]	Dog / Cat
KikiBouba [1]	Kiki / Bouba
Retina [27]	Drusen / Normal
Black-Holes	MAD / SANE
Butterfly [51]	Monarch / Viceroy
CelebA-HQ [30]	Smile / No-Smile

Baselines. We use TIME [24] as our counterfactual baseline, and replace black-box classifier labels with ground truth labels. For editing baselines, we compare against Stable Diffusion [42] with EF-DDPM inversion [21] using class-name prompts. To better accommodate visual concepts, we implemented another baseline that uses Textual Inversion [13] for each class of images and then applies source and target prompts based on the desired edit direction. We term this baseline TI + EF-DDPM. Lastly, we use the visual sliders objective of Concept Sliders [14] that provides a visual counterpart to text-driven attribute edits. To ensure a robust evaluation, we experimented with varying the rank and number of images used for defining the concept direction, selecting the best configuration for each dataset. Since the original method assumes paired data, we adapted it for unpaired settings.

4.2 Editing Results

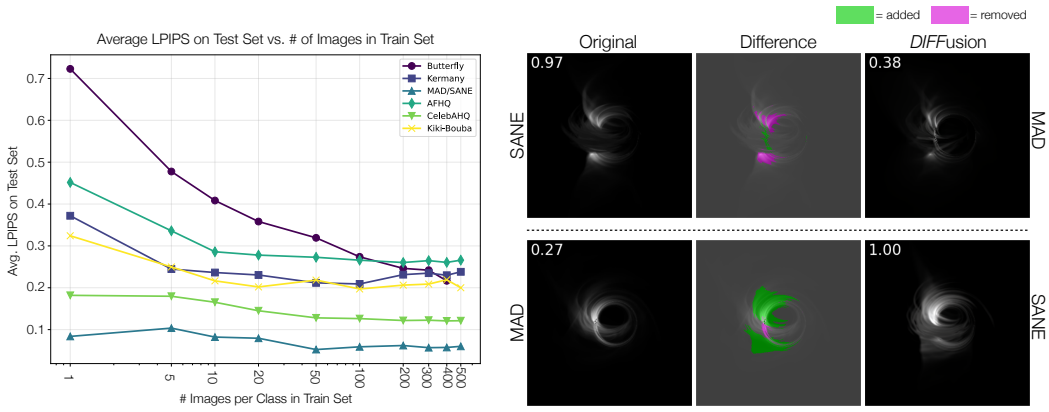
We quantitatively evaluate how well our method can make minimal edits to the image to flip the classifier's prediction. For evaluation, we take a balanced sample of 50 images per class from the validation set of each dataset, totaling 100 images from each dataset. Since our method can generate

Table 2: Performance comparison across datasets. SR = Success Ratio, LPIPS = Perceptual Distance. In **bold** are the best results, and in underline are the second-best results.

Method	Science Datasets								Regular Datasets			
	Retina		Butterfly		KikiBouba		Black-Holes		AFHQ		CelebA-HQ	
	SR↑	LPIPS↓	SR↑	LPIPS↓	SR↑	LPIPS↓	SR↑	LPIPS↓	SR↑	LPIPS↓	SR↑	LPIPS↓
EF-DDPM	0.39	0.272	0.86	0.328	0.68	0.343	0.73	0.117	**1.0**	**0.187**	**1.0**	**0.104**
TI+EF-DDPM	0.89	0.330	**1.0**	0.289	0.97	0.332	0.5	**0.045**	**1.0**	0.211	**1.0**	0.181
TIME	0.50	0.358	0.13	0.320	0.17	**0.170**	0.52	0.086	0.95	0.217	0.79	0.166
Concept Sliders	0.48	0.248	0.27	0.362	0.13	0.206	0.53	0.155	0.49	0.375	0.21	0.238
***DIFF*usion**	**0.98**	**0.217**	**1.0**	**0.218**	**0.98**	0.176	**1.0**	0.076	**1.0**	0.245	**1.0**	0.116

different strengths of edits, to pick the minimal edit, we generate 10 edits with varying strengths using the T_{skip} parameter, as does the TIME baseline [24], testing from highest to lowest T_{skip}, and select the first edit that flips the classifier prediction while maximizing LPIPS similarity to the original image.

Metrics. We evaluate our method using two key metrics. Success Ratio (SR): Also known as Flip-Rate, quantifies the ability of a method to flip an oracle classifier's decision. The oracle classifier we use is an ensemble of ResNet-18 [16], MobileNet-V2 [45], and EfficientNet-B0 [49], trained on each dataset. LPIPS [57]: Measures the perceptual similarity between the input and generated image, by capturing feature-level difference in a learned embedding space.



(a) (b)

Figure 4: (a) **Varying number of images.** Average LPIPS vs. number of images used per class. LPIPS stabilizes around 50 images for most datasets, reflecting improved identity fidelity and subtle class-distinctive feature shifts with increased embedding samples. (b) **Difference Overlay.** We visualize the difference between the input image and the counterfactual from *DIFF*usion. From SANE to MAD we notice a highlighting of the photon ring (**green**). From MAD to SANE we notice that the ring becomes less pronounced (**magenta**), and wisps appear (**green**).

Quantitative Results. As seen in Table 2, our method achieves the highest SR across all datasets compared to baseline approaches. In terms of LPIPS, it shows significant improvements over previous methods on datasets where language struggles to capture visual details (e.g., Black-Holes, KikiBouba), unlike datasets with common objects like AFHQ. It also performs either best or competitively on the remaining natural-image datasets. Additionally, while TI + EF-DDPM improves the same text-based baseline, it still struggles with images that are hard to describe textually, such as Black-Holes.

Qualitative Results. In Figure 3, we present class transitions for all baselines and *DIFF*usion. On familiar datasets like CelebA-HQ and AFHQ, our method performs well, similar to baselines. However, its strengths stand out in datasets where language may not fully capture visual details. For KikiBouba, only our method and TI + EF-DDPM round Kiki's edges, though the baseline changes

the original colors, while ours keeps them intact. In the Butterfly dataset, the baselines miss the cross-sectional line, and in the Retina dataset, only our approach removes Drusen while preserving image identity. For the Black-Holes dataset, our method flips the classifier's prediction with notable visual differences, as also highlighted in Figure 4b. These results suggest our method handles subtle visual nuances particularly well.

4.3 Teaching Results

We evaluate our method's effectiveness in teaching people subtle visual differences between classes.

User Study Design. We divided participants into three groups of 10 people each. Group 1 studied only unpaired images. Group 2 studied videos transitioning from original images to counterfactual images generated by the best baseline. Group 3 studied videos transitioning from original images to counterfactual images generated by our method. Since Groups 2 and 3 viewed transitions from real to edited images, they were also exposed to the unpaired image distribution seen by Group 1. All participants studied their respective materials for 3 minutes to learn to distinguish between the two classes before taking a test. The test required labeling 50 images, evenly distributed with 25 images from each class.

User Study Results. We assess *DIFF*usion for teaching via a user study on the Black Holes and Butterfly datasets [51], shown in Table 3 and Figure 5. For Black Holes, unpaired material gave a 78% average score, but our counterfactuals boosted this to 90%, with 40% of users hitting near-perfect scores (96%+), surpassing baselines and counterfactuals. For Butterfly, unpaired data led to varied scores, but our counterfactuals raised 9 out of 10 users above 80%, standardizing understanding effectively. P-tests confirm significance: Black Holes (p = 0.016 vs. 0.811 for baseline) and Butterfly ($p = 0.004$ vs. 0.897 for baseline), both $p < 0.05$. Our counterfactuals consistently outperform alternatives, demonstrating the usefulness of our method for teaching humans subtle visual differences.

Table 3: User Study Results - Mean Accuracy (%)

Method	Black Holes Mean±SD	Butterfly Mean±SD	Avg. Impr.
Unpaired	78.6±13.7	61.6±22.8	—
Baseline	77.2±11.5	62.8±16.8	-0.1%
Ours	90.8±4.8	87.8±10.4	+19.2%

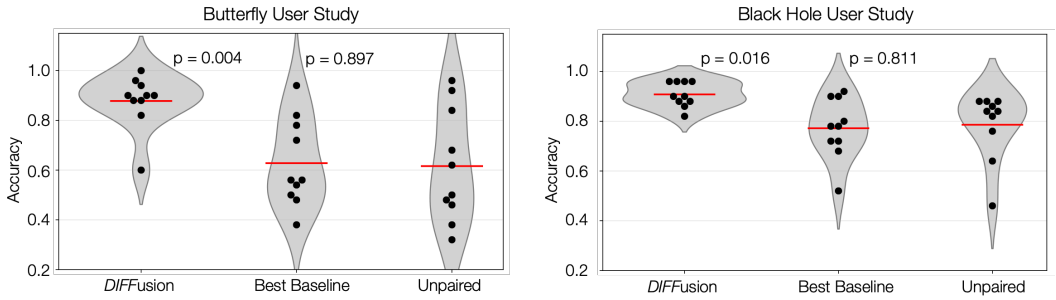


Figure 5: **User Study Results**. We plot the results from user studies across users who studied our counterfactuals, users who studied the best baseline counterfactuals, and users who studied unpaired images. For both Butterfly and Black Hole datasets, we observe that the users who studied our counterfactuals significantly outperformed the other groups. The violin plots illustrate the distribution of user percentages, where the width of each grey shape represents the density of data points.

4.4 Method Analysis

Varying Dataset Size. In Figure 4a, we examine the impact of varying the number of images per class on the average LPIPS metric across the test sets. We notice that for most datasets, the LPIPS stops improving at around 50 images. In Section B.4, we show qualitative results as the number of images changes. We notice that as the number of images incorporated into the average embeddings

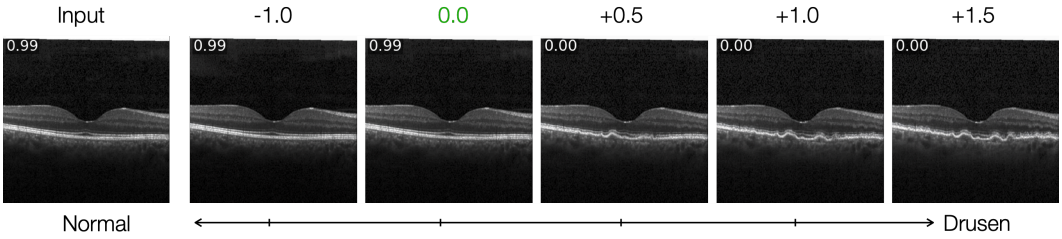


Figure 6: **Interpolation.** By varying the manipulation scale $\omega \in \{-0.5, 0.0, 0.5, 1.0, 1.5\}$, we can adjust the manipulation strength, allowing for smooth interpolation between the two classes. Notably, when $\omega = 0$, we can reconstruct the original image while preserving the classifier's probabilities.

increases, the fidelity to the original image's identity improves, while subtly altering the features that are distinctive between classes.

Interpolation. In Figure 6, we present qualitative results demonstrating the effects of varying the manipulation scale, w, on an instance of a Normal retina. The manipulation scale, which can take positive or negative values, modulates the transformation direction. Positive values of w shift the features toward Drusen from the Normal retina, while negative values make the image smoother.

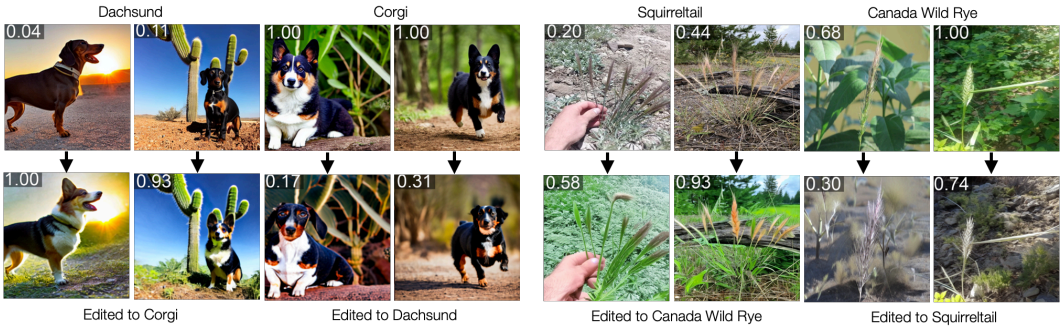


Figure 7: **Dataset Bias**. *DIFF*usion can reveal dataset bias. Squirreltail-to-Canada Wild Rye shifts emphasize environmental backgrounds over plant traits, reflecting iNaturalist's contextual bias, while Dachshund-to-Corgi edits prioritize foreground dog features, highlighting variable bias impact.

4.5 Visualizing Dataset Bias

Our method edits images using differences between class mean embeddings, making it sensitive to dataset bias. If distinguishing features reflect unintended biases rather than targeted traits, edits deviate from our intent. This is both a limitation - preventing precise control, and a strength, as it visualizes dataset biases, revealing underlying structure. We show how dataset bias is captured by our method in Figure 7. In iNaturalist [51], counterfactuals from Squirreltail (dry climates) to Canada Wild Rye (humid) shift backgrounds more than plant structure, suggesting environmental bias dominates. Conversely, using the Spawrious [32] dataset, Dachshund-to-Corgi counterfactuals prioritize dog features (e.g., shape, size) over jungle-to-desert backgrounds. We attribute this to stronger foreground differences in dogs and clearer object-background separation, unlike plants blending into settings in iNaturalist data. The effect of dataset bias on edits varies with class prominence and context.

5 Discussion and Limitations

*DIFF*usion generates counterfactuals to support visual expertise training across domains with limited data. It reveals dataset biases, often shifting unintended features due to embedding reliance, which limits precise control. Additionally, the arithmetic is very simple: a difference of averages, highlighting a trade-off between flexibility and specificity. Future work could explore disentanglement or guidance mechanisms to enhance edit precision in specialized applications.

References

[1] Morris Alper and Hadar Averbuch-Elor. Kiki or bouba? sound symbolism in vision-and-language models, 2024.

[2] Moab Arar, Rinon Gal, Yuval Atzmon, Gal Chechik, Daniel Cohen-Or, Ariel Shamir, and Amit H. Bermano. Domain-agnostic tuning-encoder for fast personalization of text-to-image models, 2023.

[3] Maximilian Augustin, Yannic Neuhaus, and Matthias Hein. Dig-in: Diffusion guidance for investigating networks – uncovering classifier differences neuron visualisations and visual counterfactual explanations, 2024.

[4] Manuel Brack, Felix Friedrich, Katharina Kornmeier, Linoy Tsaban, Patrick Schramowski, Kristian Kersting, and Apolinário Passos. Ledits++: Limitless image editing using text-to-image models, 2024.

[5] Tim Brooks, Aleksander Holynski, and Alexei A. Efros. Instructpix2pix: Learning to follow image editing instructions, 2023.

[6] Mia Chiquier and Carl Vondrick. Muscles in action. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22091–22101, 2023.

[7] Mia Chiquier, Utkarsh Mall, and Carl Vondrick. Evolving interpretable visual classifiers with large language models. In *European Conference on Computer Vision*, pages 183–201. Springer, 2024.

[8] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. Stargan v2: Diverse image synthesis for multiple domains. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8185–8194, 2020.

[9] Guillaume Couairon, Jakob Verbeek, Holger Schwenk, and Matthieu Cord. Diffedit: Diffusion-based semantic image editing with mask guidance, 2022.

[10] Prafulla Dhariwal and Alex Nichol. Diffusion models beat gans on image synthesis, 2021.

[11] Lisa Dunlap, Krishna Mandal, Trevor Darrell, Jacob Steinhardt, and Joseph E Gonzalez. Vibecheck: Discover and quantify qualitative differences in large language models. *arXiv preprint arXiv:2410.12851*, 2024.

[12] Karim Farid, Simon Schrodi, Max Argus, and Thomas Brox. Latent diffusion counterfactual explanations, 2023.

[13] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H. Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion, 2022.

[14] Rohit Gandikota, Joanna Materzynska, Tingrui Zhou, Antonio Torralba, and David Bau. Concept sliders: Lora adaptors for precise control in diffusion models, 2023.

[15] Zinan Guo, Yanze Wu, Zhuowei Chen, Lang Chen, Peng Zhang, and Qian He. Pulid: Pure and lightning id customization via contrastive alignment, 2024.

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015.

[17] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control, 2022.

[18] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance, 2022.

[19] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models, 2020.

[20] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021.

[21] Inbar Huberman-Spiegelglas, Vladimir Kulikov, and Tomer Michaeli. An edit friendly ddpm noise space: Inversion and manipulations, 2024.

[22] Guillaume Jeanneret, Loïc Simon, and Frédéric Jurie. Diffusion models for counterfactual explanations, 2022.

[23] Guillaume Jeanneret, Loïc Simon, and Frédéric Jurie. Adversarial counterfactual visual explanations, 2023.

[24] Guillaume Jeanneret, Loïc Simon, and Frédéric Jurie. Text-to-image models for counterfactual explanations: A black-box approach. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 4757–4767, 2024.

[25] Hong-Xuan Jiang, Yosuke Mizuno, Christian M Fromm, and Antonios Nathanail. Two-temperature grmhd simulations of black hole accretion flows with multiple magnetic loops. *Monthly Notices of the Royal Astronomical Society*, 522(2):2307–2324, 2023.

[26] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *nature*, 596(7873):583–589, 2021.

[27] Daniel S Kermany, Michael Goldbaum, Wenjia Cai, Carolina CS Valentim, Huiying Liang, Sally L Baxter, Alex McKeown, Ge Yang, Xiaokang Wu, Fangbing Yan, et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. *cell*, 172(5):1122–1131, 2018.

[28] Black Forest Labs. Flux. https://github.com/black-forest-labs/flux, 2024.

[29] Oran Lang, Yossi Gandelsman, Michal Yarom, Yoav Wald, Gal Elidan, Avinatan Hassidim, William T. Freeman, Phillip Isola, Amir Globerson, Michal Irani, and Inbar Mosseri. Explaining in style: Training a gan to explain a classifier in stylespace, 2021.

[30] Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. Maskgan: Towards diverse and interactive facial image manipulation. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5548–5557, 2020.

[31] Ce Liu, Antonio Torralba, William T Freeman, Frédo Durand, and Edward H Adelson. Motion magnification. *ACM transactions on graphics (TOG)*, 24(3):519–526, 2005.

[32] Aengus Lynch, Gbètondji J-S Dovonon, Jean Kaddour, and Ricardo Silva. Spawrious: A benchmark for fine control of spurious correlation biases, 2023.

[33] Utkarsh Mall, Cheng Perng Phoo, Mia Chiquier, Bharath Hariharan, Kavita Bala, and Carl Vondrick. Disciple: Learning interpretable programs for scientific visual discovery. *arXiv preprint arXiv:2502.10060*, 2025.

[34] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations, 2022.

[35] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models, 2022.

[36] Alexander Quinn Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob Mcgrew, Ilya Sutskever, and Mark Chen. GLIDE: Towards photorealistic image generation and editing with text-guided diffusion models. In *Proceedings of the 39th International Conference on Machine Learning*, pages 16784–16804. PMLR, 2022.

[37] Tae-Hyun Oh, Ronnachai Jaroensri, Changil Kim, Mohamed Elgharib, Fr'edo Durand, William T Freeman, and Wojciech Matusik. Learning-based video motion magnification. In *Proceedings of the European conference on computer vision (ECCV)*, pages 633–648, 2018.

[38] Gaurav Parmar, Krishna Kumar Singh, Richard Zhang, Yijun Li, Jingwan Lu, and Jun-Yan Zhu. Zero-shot image-to-image translation, 2023.

[39] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents, 2022.

[40] Anton Razzhigaev, Arseniy Shakhmatov, Anastasia Maltseva, Vladimir Arkhipkin, Igor Pavlov, Ilya Ryabov, Angelina Kuts, Alexander Panchenko, Andrey Kuznetsov, and Denis Dimitrov. Kandinsky: an improved text-to-image synthesis with image prior and latent diffusion, 2023.

[41] Pau Rodriguez, Massimo Caccia, Alexandre Lacoste, Lee Zamparo, Issam Laradji, Laurent Charlin, and David Vazquez. Beyond trivial counterfactual explanations with diverse valuable explanations, 2021.

[42] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10674–10685, 2022.

[43] Bernardino Romera-Paredes, Mohammadamin Barekatain, Alexander Novikov, Matej Balog, M Pawan Kumar, Emilien Dupont, Francisco JR Ruiz, Jordan S Ellenberg, Pengming Wang, Omar Fawzi, et al. Mathematical discoveries from program search with large language models. *Nature*, 625(7995):468–475, 2024.

[44] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Rapha Gontijo Lopes, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding, 2022.

[45] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks, 2019.

[46] Arseniy Shakhmatov, Anton Razzhigaev, Aleksandr Nikolich, Vladimir Arkhipkin, Igor Pavlov, Andrey Kuznetsov, and Denis Dimitrov. kandinsky 2.2. https://github.com/ai-forever/Kandinsky-2, 2023.

[47] Bartlomiej Sobieski and Przemysław Biecek. Global counterfactual directions, 2024.

[48] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models, 2022.

[49] Mingxing Tan and Quoc V. Le. Efficientnet: Rethinking model scaling for convolutional neural networks, 2020.

[50] Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1921–1930, 2023.

[51] Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The inaturalist species classification and detection dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8769–8778, 2018.

[52] Bram Wallace, Akash Gokul, and Nikhil Naik. Edict: Exact diffusion inversion via coupled transformations, 2022.

[53] George N Wong, Ben S Prather, Vedant Dhruv, Benjamin R Ryan, Monika Mościbrodzka, Chi-kwan Chan, Abhishek V Joshi, Ricardo Yarza, Angelo Ricarte, Hotaka Shiokawa, et al. Patoka: Simulating electromagnetic observables of black hole accretion. *The Astrophysical Journal Supplement Series*, 259(2): 64, 2022.

[54] Chen Henry Wu and Fernando De la Torre. Unifying diffusion models' latent space, with applications to cyclediffusion and guidance, 2022.

[55] Hao-Yu Wu, Michael Rubinstein, Eugene Shih, John Guttag, Frédo Durand, and William Freeman. Eulerian video magnification for revealing subtle changes in the world. *ACM transactions on graphics (TOG)*, 31 (4):1–8, 2012.

[56] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models, 2023.

[57] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018.

[58] Xiaojin Zhu. Machine teaching: An inverse problem to machine learning and an approach toward optimal education. In *Proceedings of the AAAI conference on artificial intelligence*, 2015.

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes] , [No] , or [NA] .
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes] " is generally preferable to "[No] ", it is perfectly acceptable to answer "[No] " provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No] " or "[NA] " is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS Paper Checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Please see the abstract and the paper.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Please see the limitations section at the very end.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: NA

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Yes, and implementation details can be found in the supplemental material.

   Guidelines:

   - The answer NA means that the paper does not include experiments.

- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We release code, and data is publicly available except for the Black Holes dataset.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).

- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: The main paper justifies hyperparams, optimizer, etc, but the data splits can be found in the code release.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [NA]

   Justification: We follow common practice.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: See supplemental.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.

- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: .

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: See supplemental.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [Yes]

    Justification: See supplemental.

    Guidelines:

    - The answer NA means that the paper poses no such risks.

- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: See supplemental.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    No new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [Yes]

    Justification: See supplemental.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [Yes]

    Justification: See supplemental.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: NA.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (https://neurips.cc/Conferences/2025/LLM) for what should or should not be described.