Analysis of Diffusion Model's Inference Mechanism Using XAI Techniques

Doeun Kim[○1], Jieun Byeon[1], Inae Park[1]

[1]Dept. Of Computer Science and Engineering, Ewha Womans Univ.

ABSTRACT

Recently, diffusion models have demonstrated high performance across various domains including image generation. These models generate high-quality outputs through a denoising process, yet their complex internal mechanisms pose challenges for intuitive understanding. This study employs XAI(eXplainable AI) techniques to elucidate the inference mechanism of diffusion models. Specifically, methods like Integrated Gradients, Gradient SHAP, and Occlusion are applied to analyze how the model focuses on different pixel regions over time-step during image generation.

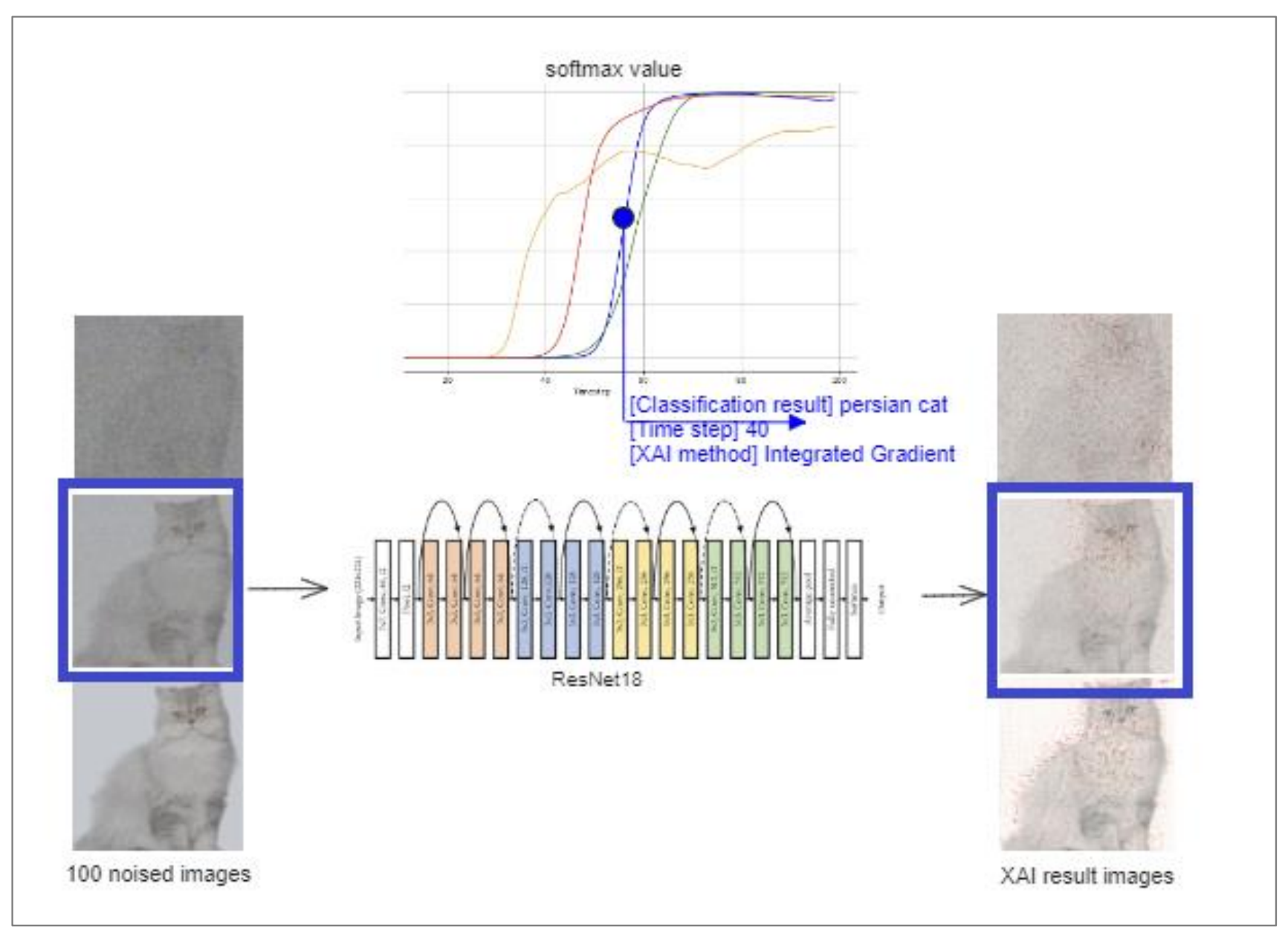
INTRODUCTION

In recent years, diffusion models have demonstrated their prowess in various applications such as image generation and anomaly detection. These models are grounded in principles of probabilistic modeling and stochastic processes, utilizing iterative noise addition and removal processes to handle data. Particularly, they excel in enhancing the quality and diversity of outputs through denoising procedures. However, due to their complex structure, diffusion models present limitations as "black-box" models, making it challenging to intuitively understand their internal operations. Therefore, there is a critical need for techniques that can provide deeper insights into and explanations of how these models function. To address this, we analyzed the focus areas of images generated at each time-step using techniques such as Integrated Gradient, Grad SHAP, and Occlusion.

BACKGROUND

Recent advancements in diffusion models have revolutionized various domains, particularly in image generation. These models employ a denoising process to produce high-quality images but are hindered by their intricate internal mechanisms, challenging intuitive comprehension. Understanding these mechanisms is crucial for enhancing model interpretability and performance.

METHOD



Experimental Setup

In this study, we designed experiments based on Denoising Diffusion Probabilistic Models (DDPM). We utilized a pre-trained ResNet-18 classifier on the ImageNet1000 dataset. For our analysis, we selected the Persian cat, Siamese cat, Egyptian cat, and tiger labels corresponding to the ImageNet1000 classes. For each label image, we generated 100 noisy images by adding pre-scheduled Gaussian noise over 100 timesteps.

XAI Techniques

We applied Integrated Gradients, Gradient SHAP, and Occlusion techniques to evaluate the contribution of each model prediction. These techniques allowed us to visualize the important regions of the noised images over time.

Analysis of Likelihood Increase

To identify the time step at which the likelihood of the noised images begins to increase significantly, we implemented a softmax function. We targeted the regions where the gradient showed a steep rise and conducted an in-depth analysis of the XAI results for these specific intervals.

RESULTS

This study utilizes three XAI(eXplainable AI) techniques and 100 sampled noise images to elucidate the operational mechanisms of Diffusion Models, which belong to the Black Box Model category. To understand the features that lead to different classifications and generations of similar images, a comparison was conducted using three types of cats and a tiger class.

Each image is arranged clockwise from the top left as follows: Persian Cat, Siamese Cat, Tiger, and Egyptian Cat. For each class, noise sampling was conducted over 100 timesteps, followed by the application of XAI techniques.

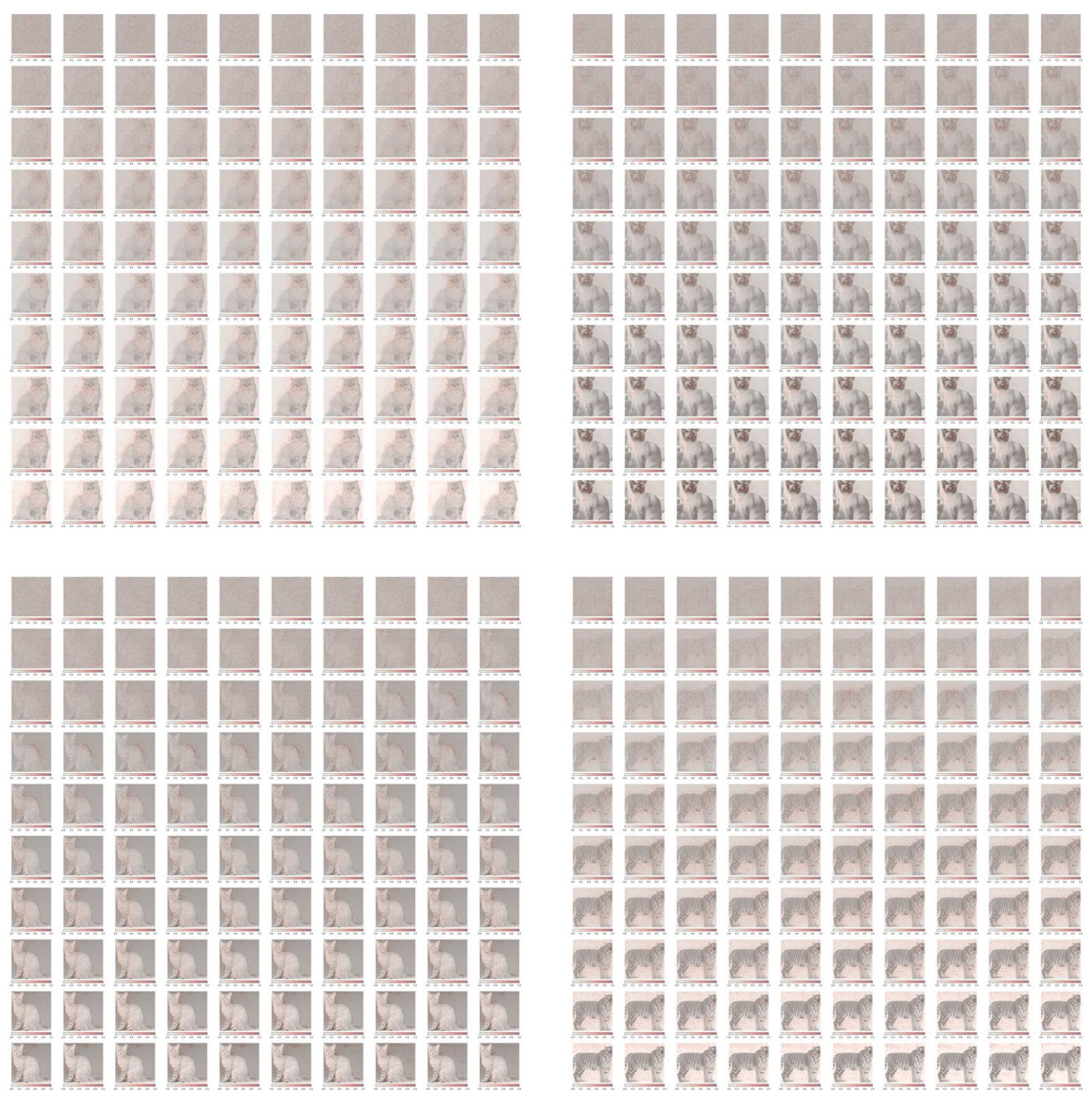


Fig2. Integrated Gradient Experiment results

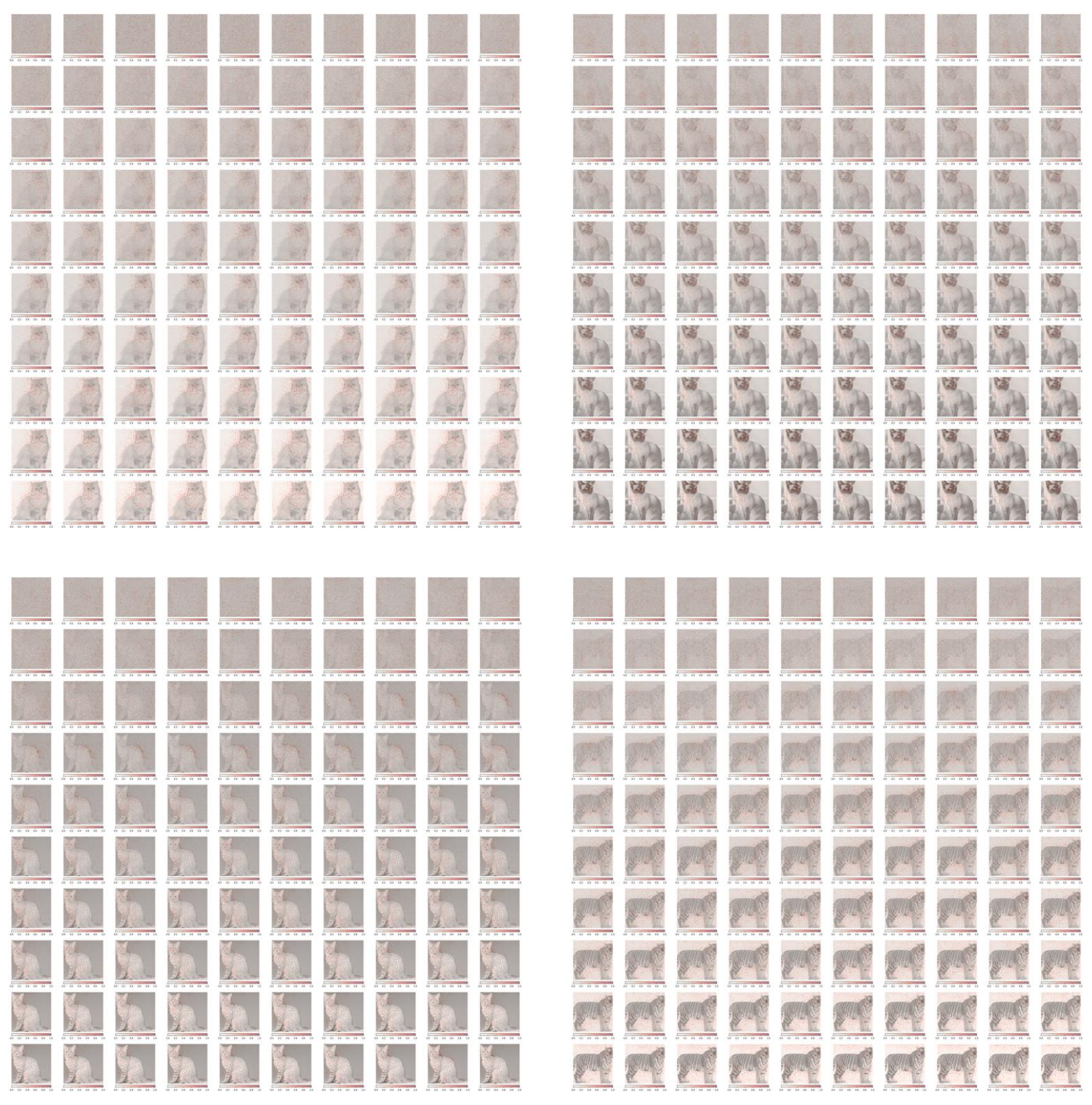


Fig3. Grad SHAP Experiment results

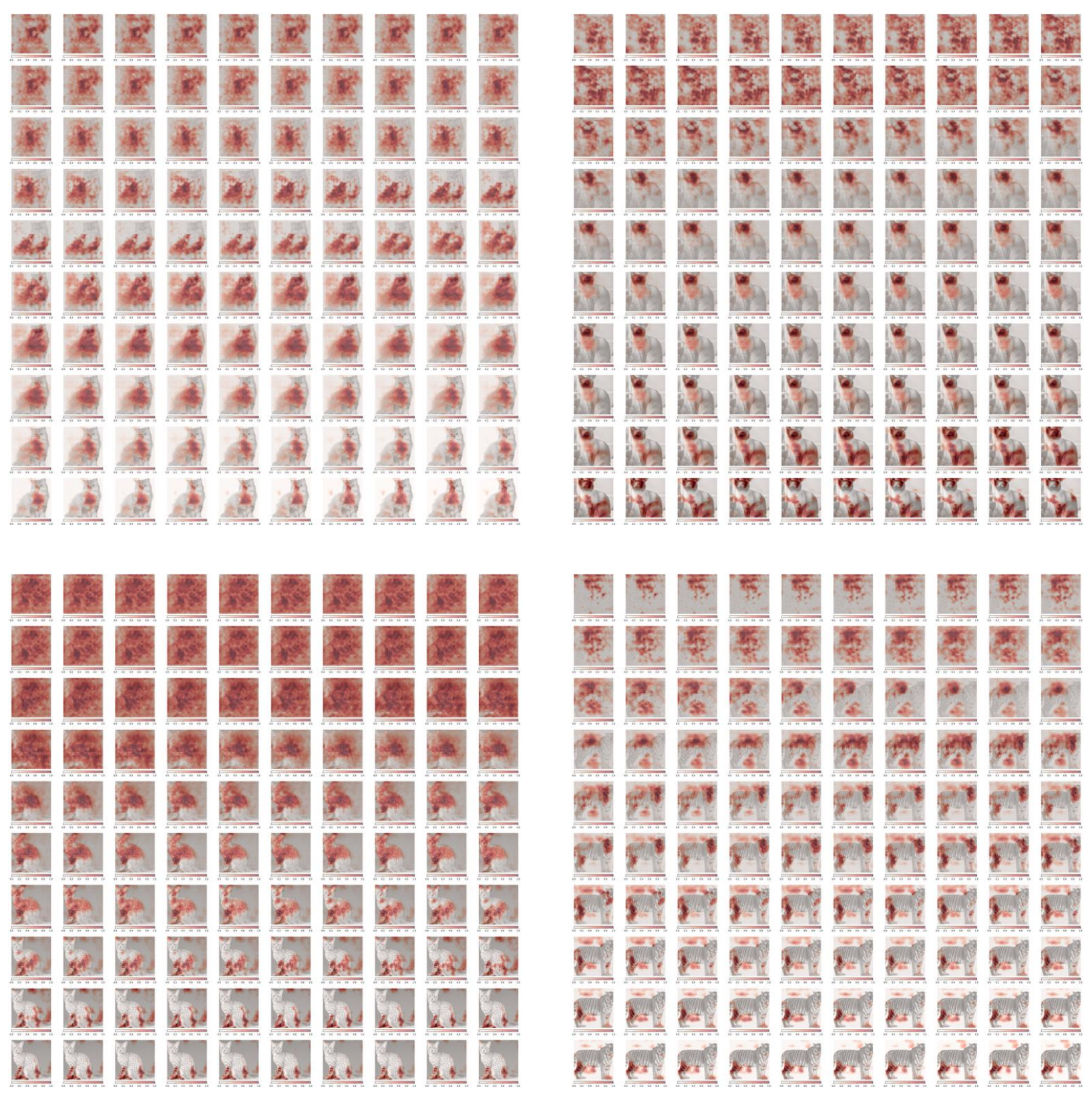


Fig4. Occlusion Experiment results

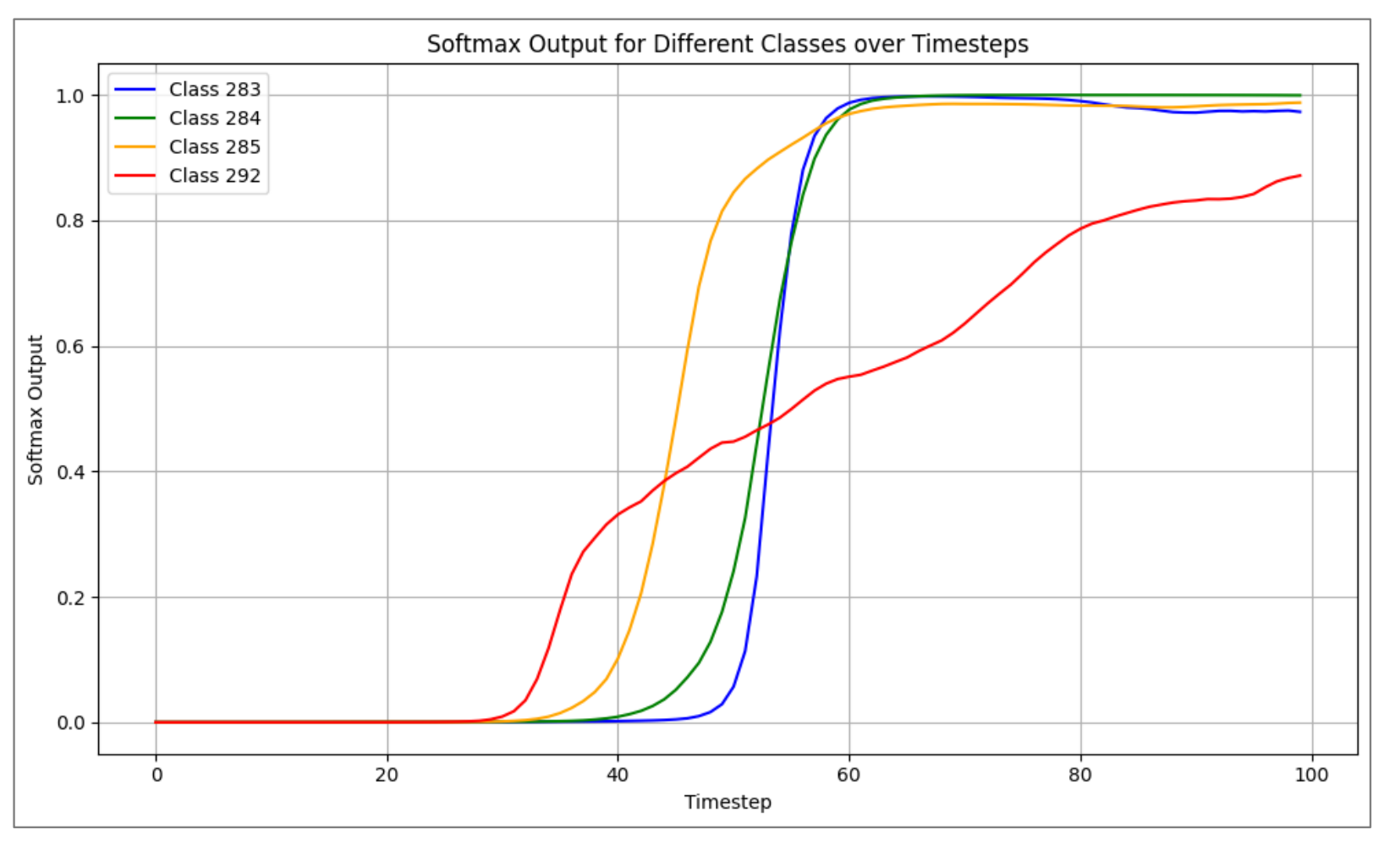


Fig5. Softmax Output

In the early stages of the time steps, all classes performed classification based on the overall outline, but proper classification began around timesteps 40-60.

CONCLUSION

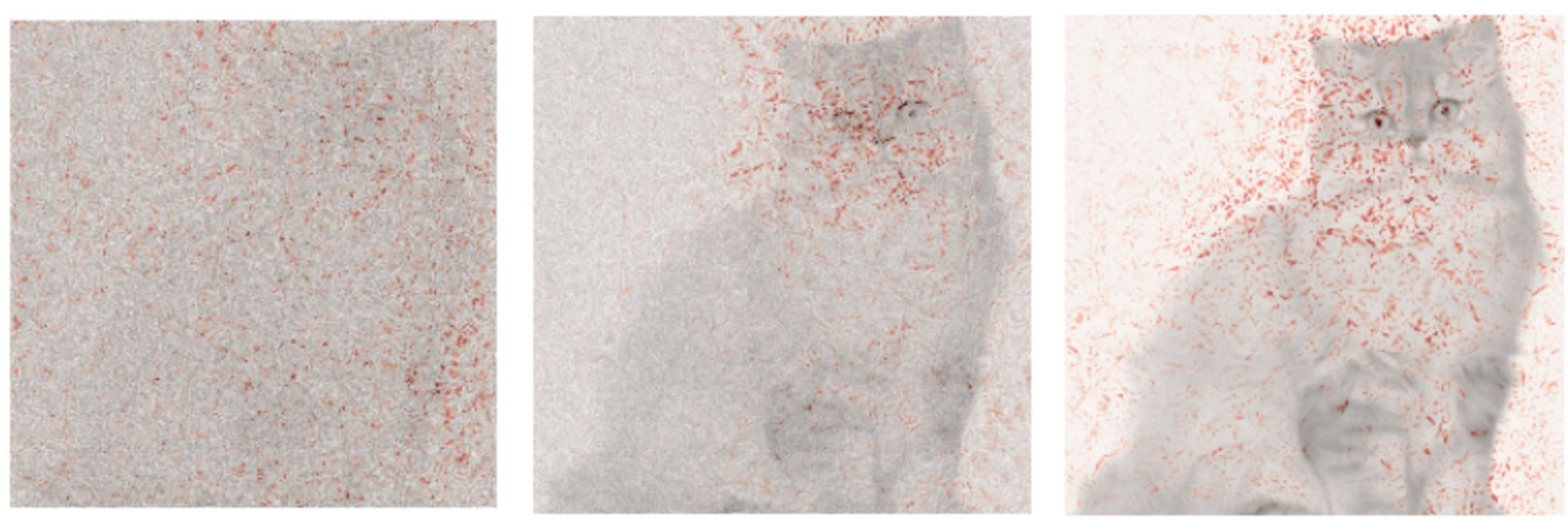


Fig6. Integrated Gradient, Persian cat (From the left, time steps 20, 50, and 90)

Applying XAI techniques to the images from time steps 40-60 revealed that, in the case of cats, the main features could generally be considered as distinctive and detailed elements like the face.

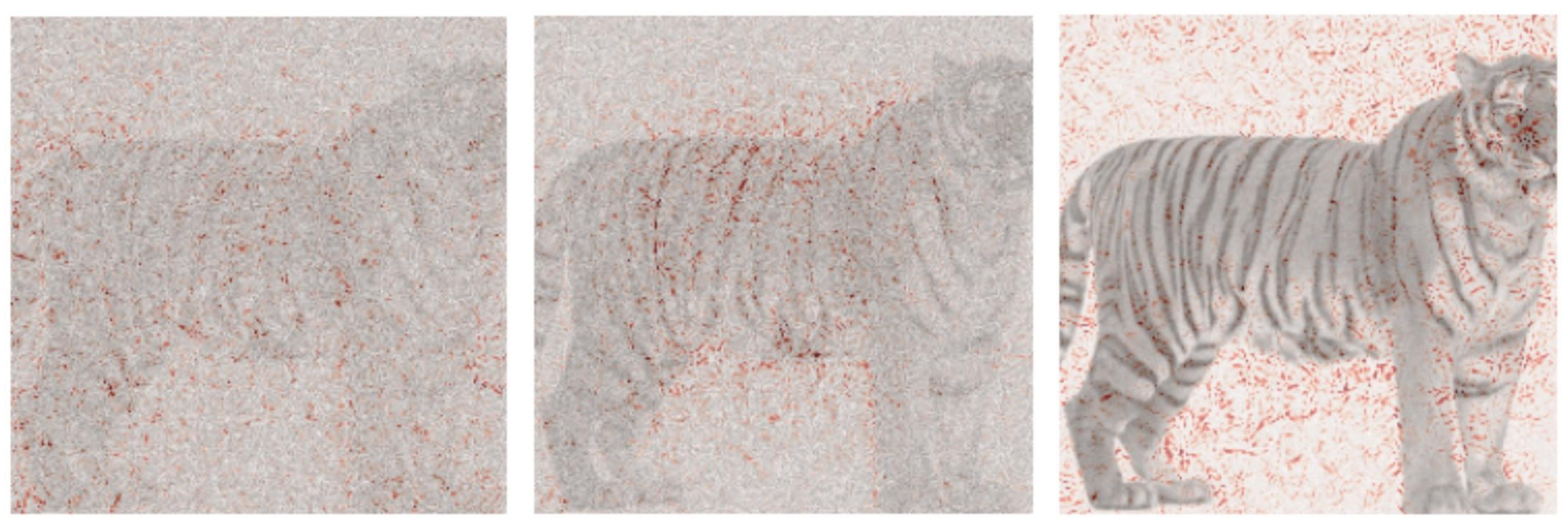


Fig7. Integrated Gradient, Tiger (From the left, time steps 20, 30, and 90)

For tigers, the likelihood increased based on strong features like stripes, and ultimately, when facial features similar to those of cats were used, the likelihood peaked.

This suggests that the image generation process of Diffusion Models is similar to the sequential process humans follow when drawing pictures. In other words, using XAI techniques, it was possible to observe a transition from focusing on the overall outline to concentrating on detailed elements.

References

[1] Denoising Diffusion Probabilistic Model, Jonathan Ho, Ajay Jain, Pieter Abbeel (2020) arXiv:2006.11239
[2] Axiomatic Attribution for Deep Networks, Mukund Sundararajan, Ankur arXiv:1703.01365 Taly, Qiqi Yan (2017)
[3] SHAPLEY EXPLANATION NETWORKS, Rui Wang Xiaoqian Wang David I. Inouye (2021) arXiv:2104.02297
[4] Occlusion Sensitivity Analysis with Augmentation Subspace Perturbation in Deep Feature Space, Pedro H. V. Valois, Koichiro Niinuma, Kazuhiro Fukui (2023) arXiv:2311.15022