A Novel Framework for Benchmarking Vision Models: Integrating Language Models and Explainable AI

Anonymous ACL submission

Abstract

While traditional metrics such as accuracy and Intersection over Union effectively quantify the performance of deep learning models, they provide limited insights into the interpretability and focus. Current evaluation approaches fail to capture how well models concentrate on relevant features or whether they rely on spurious correlations, which is crucial for building trust and transparency in AI systems. In this paper, we propose a novel framework for evaluating the performance and interpretability of deep learning vision models using CAM visualizations and large language models (LLMs)-based scoring system. We apply this framework to 14 state-of-the-art models, including ResNet, MobileNet, and Swin Transformer, using the ImageNet dataset as a benchmark. CAM methods are used to extract regions that highlight the attention map of the models. To assess the quality of these visual explanations, we introduce an LLM-powered scoring mechanism that evaluates the relevance of the highlighted regions to the predicted class. Through extensive experiments, we compare the interpretability of these models, showing how LLM-based evaluations can complement and enhance traditional model benchmarking. Our results show that this LLM-based evaluation method provides a more comprehensive understanding of model behavior and interpretability, advancing the field of explainable AI. We believe that this approach can be extended to other domains and explainability techniques, offering new possibilities for enhancing model transparency.

1 Introduction

The rapid advancements in deep learning, particularly in computer vision, have led to significant improvements in model performance across various tasks including object detection, image classification, and segmentation. These models are evaluated primarily using metrics such as accuracy and Intersection over Union (IoU), which measure the model's performance by comparing predictions to ground truth labels (Girshick, 2015; Fang et al., 2021; Pham et al., 2024). While these metrics provide a model's output accuracy, they fail to capture a crucial aspect: the interpretability of its decisions. Moreover, these metrics often rely on thresholds, such as a confidence score in classification or a pixel-wise threshold in IoU for segmentation, to determine the correctness of a prediction. These thresholds can obscure the true behavior of the model, as small variations in the predicted output may result in drastically different scores, despite the model focusing on relevant regions or objects in the input image. Thus, the metrics fail to reflect the interpretability of the model's predictions such as *how* or *why* a model makes predictions.

To address these limitations, explainable AI (xAI) has recently gained significant attention, which aims to enhance the transparency of model predictions. Broadly, xAI methods can be categorized into two types: model-agnostic (or result-based) and model-specific (or model-based) approaches (Lundberg and Lee, 2017). Model-agnostic techniques work for any model and often rely on feature importance analysis, which assesses the contribution of each input feature to the final prediction. In contrast, model-specific techniques leverage the internal structure of the model to provide explanations (Bach et al., 2015). Despite the pros and cons of both approaches, model-agnostic methods are commonly used with large models due to their convenience compared to approaches that rely on understanding a model's internal architecture (Loh et al., 2022; Minh et al., 2022). Popular model-agnostic approaches include saliency maps, Local Interpretable Model-agnostic Explanations (LIME), and Grad-CAM, which highlight important regions in the input data that influence model predictions, providing valuable insights into the decision-making process (Itti et al., 1998; Ribeiro et al., 2016; Selvaraju et al., 2017).

Large language models such as BERT, GPT-4, LLaMA-2, and Claude, have made significant contributions to natural language processing (NLP), wisely used in a wide range of tasks (Devlin et al., 2019; Achiam et al., 2023; Touvron et al., 2023b). LLMs have demonstrated remarkable abilities in understanding context, reasoning, and generating human-like text, making them valuable tools for summarizing and interpreting complex outputs from AI systems. Researchers have been investigating how LLMs can enhance the interpretability of deep learning models, for instance, by generating natural language explanations for model predictions or evaluating the quality of existing explanations. Moreover, Vision Language Models (VLMs), which extend LLMs to visual tasks like visual question answering (VQA) and image captioning, have emerged as powerful tools in AI research (Ranasinghe and Ryoo, 2023).

While LLMs are widely used to generate descriptions for images, their potential for interpreting and evaluating vision model results remains underexplored. Moreover, numerous vision models have emerged over the past decades, yet evaluations have primarily centered on accuracy, often overlooking the crucial aspect of interpretability. Given the importance of understanding model decisions for trust and transparency in AI systems, we propose an LLM-based framework to interpret predictions from vision models by analyzing their Grad-CAM visualizations. Our contributions are as follows:

- **Framework Development**: We propose a novel pipeline that integrates LLMs with Grad-CAM visualizations to enhance the interpretability of predictions made by vision models.
- **Evaluation Methodology**: We propose a systematic methodology for quantitatively assessing the interpretability of SOTA vision models by analyzing the relevance of highlighted features in Grad-CAM outputs.
- **Interpretation Benchmark**: We create a benchmark to evaluate vision models on the classification task using the ImageNet dataset.

2 Related Work

The need for transparency and interpretability in AI systems has led to the development of XAI, which encompasses a range of techniques to improve the interpretability of machine learning models (Gunning and Aha, 2019; Zhao et al., 2015). In computer vision, several approaches have been proposed to highlight regions a model focuses on, including CAM (Zhou et al., 2016), Grad-CAM (Selvaraju et al., 2017), Grad-CAM++ (Chattopadhyay et al., 2017), LayerCAM (Jiang et al., 2021), ScoreCAM (Wang et al., 2020a), Eigen-CAM, and XGradCAM (Oquab et al., 2015; Wang et al., 2020b). Saliency maps help researchers understand the main reason behind the vision models' decisions (Itti et al., 1998; Kümmerer et al., 2014; Zhao et al., 2015), thus helping them develop new architectures and methods to improve overall performances. Grad-CAM, for example, can produce class-discriminative localization maps by utilizing the gradients of the target class flowing into the last convolutional layer. This CAM method highlights the most relevant regions in the input image to the prediction, offering valuable insights into the model's decision-making process. The application of XAI ranges from medicine to remote sensing and transportation (Borys et al., 2023; Kakogeorgiou and Karantzalos, 2021; Kim and Joe, 2022). For example, XAI techniques have been used in medical image analysis to explain deep learning model predictions in critical tasks such as disease diagnosis, making these systems more interpretable for healthcare professionals (Borys et al., 2023).

On the other hand, evaluating vision models requires significant effort to construct a dataset, including collecting images and annotations. While benchmarks like ImageNet (Deng et al., 2009), COCO (Lin et al., 2014), MNIST (Lecun et al., 1998), CIFAR, and Cityscapes (Cordts et al., 2016) successfully evaluate the performance of vision models on many tasks such as image classification, detection, and segmentation, there remains a lack of understanding of how vision models achieve such results, thus slowing down the development of this field. Consequently, creating a benchmark that facilitates a deeper understanding of vision models through quantitative evaluation and natural language is necessary to drive further advancements. Such benchmarks provide how different architectures and models' modifications affect their performance and image understanding, which leads to a better understanding of transformer-based models such as ViT (Dosovitskiy, 2020), YOLOS (Fang et al., 2021), DETR (Carion et al., 2020), DaViT (Ding et al., 2022), CLIP (Radford et al., 2021), and convolution-based models like VGG (Simonyan

and Zisserman, 2014), AlexNet (Krizhevsky et al., 2012), DenseNet (Huang et al., 2017), EfficientNet (Tan and Le, 2019), and GoogLeNet (Szegedy et al., 2015). Further phenomena like vanishing gradient, bias, fairness, and exploding gradient can also be investigated with this type of benchmark.

Furthermore, large language models or LLMs, such as Qwen (Bai et al., 2023), Llama (Touvron et al., 2023a), Phi (Li et al., 2023), Mixtral (Jiang et al., 2023), and Gemma (Mesnard et al., 2024), perform effectively on many natural language tasks, including logical thinking. Their abilities extend their applications in many domains, including judging different models' performance. Some existing benchmarks such as MT Bench 101 (Bai et al., 2024), MT Bench (Zheng et al., 2023), Vicuna-Benchmark (Chiang et al., 2023), and AGIEval (Zhong et al., 2024) utilize LLMs to evaluate other language models. This indicates the reliability of LLMs as we use them as judges for many tasks that require human-level evaluation and explanation despite the risk of hallucination and bias. Further extension of LLM applications to even more domains (Vision, XAI, etc.) is needed to utilize their strength in enhancing the research process of different fields.

Despite the application of LLMs in judging other models in their natural language abilities, using vision language models or VLMs in similar tasks is relatively new. However, some research such as Prometheus-Vision (Lee et al., 2024), WildVision (Lu et al., 2024), and MLLM-as-a-Judge (Chen et al., 2024) indicate that VLMs can also be judges to evaluate models on their vision ability. This creates a new opportunity to apply VLMs on more domains. The advantage of this method is its flexibility in evaluating criteria depending on the target of the benchmark. Moreover, this flexibility also allows for the incorporation of various prompting strategies, such as Chain of Thought (Wei et al., 2024), Tree of Thought (Yao et al., 2023), and few-shot learning (Brown et al., 2020), to enhance the accuracy of the benchmark (Wei et al., 2022; Lyu et al., 2023; Reynolds and McDonell, 2021; Ma et al., 2023). However, while these methods and VLMs primarily focus on improving model descriptiveness and reducing LLM hallucinations, the internal interpretability of vision modules of VLMs is still a black box, making those applications of VLMs need more time to ensure quality and reliability.

Despite the significant advancements in VLMs in recent years, LLM applications to analyze interpretive visualizations, such as Grad-CAM, in visual models remain underexplored. To fill this gap, Nguyen et al. (2024) explored the potential of using LLMs to generate textual explanations for visual recognition based on the intensity of colors extracted from CAM methods. However, this approach is limited to text generation without evaluation or comparison of the generated attention maps from vision models. To further bridge this gap, we developed a pipeline that utilizes LLMs to evaluate predictions from vision models, scoring them and providing detailed explanations.

3 Methodology

In this section, we describe our proposed pipeline of utilizing CAM methods and existing vision language models to provide detail explanations and quantitative evaluation of vision models on their abilities to understand and comprehend images.

3.1 Pipeline

We introduce a novel pipeline to evaluate the ability of vision models to understand images. This pipeline combines many CAM methods to visualize the model's attention with vision-language models for evaluation and scoring. The entire pipeline to assess vision models is illustrated in Figure 1.

The pipeline starts by feeding an image to vision models and getting a predicted result of models on the image. After that, different methods to extract models' attention, including CAM, LayerCAM, and more, are utilized to get an attention map of vision models on the image. Then, we apply an activation function, a modified version of the *sigmoid* function, to each value of the attention map to get a mask for each image. The activation function is illustrated in Equation 1, where v_{xy} is the value of the attention map at position (x, y), indicating the importance of the pixel at that position, and M_{xy} is the activated value at position (x, y). This function maps the input v_{xy}, ranging from 0 to 1, to the output M_{xy}, which is also between 0 and 1.

$$M_{xy} = \frac{1}{1 + \exp\left(25 \cdot (0.4 - v_{xy})\right)} \qquad (1)$$

After achieving the mask, we apply it to the original image to hide regions the vision model pays less attention to according to the CAM method. This process is formulated in Equation 2, where we multiply each pixel in the original image I with the

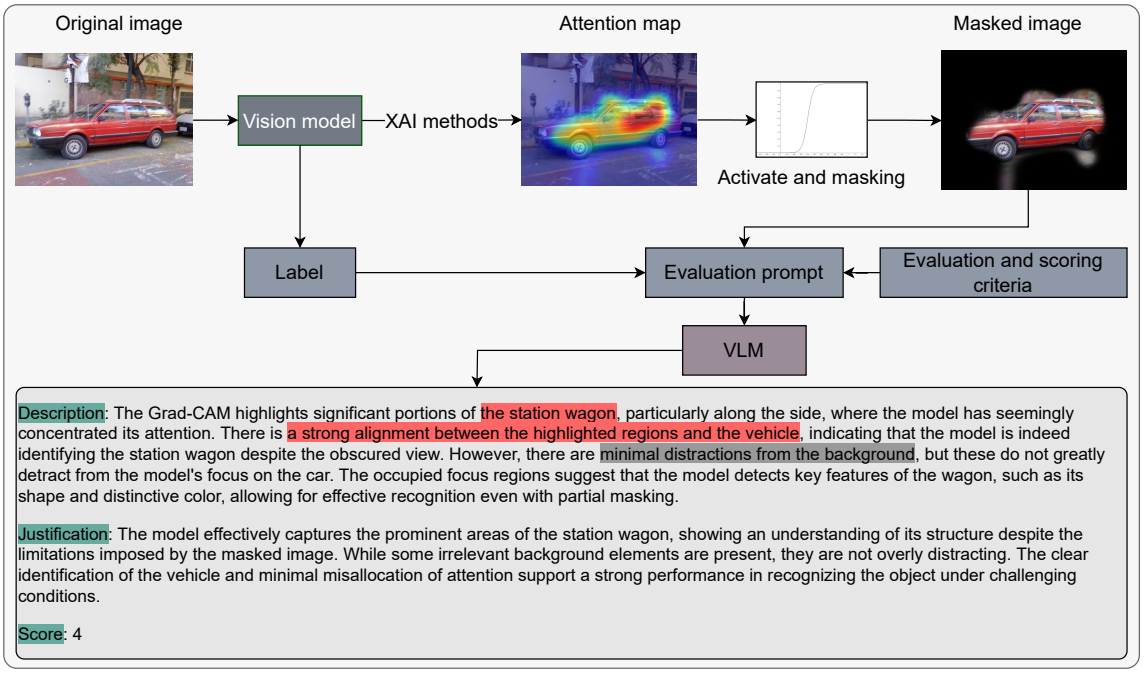


Figure 1: The pipeline evaluates vision models' ability to understand an image. The VLM model can describe, justify, and score the input image and the corresponding attention map. In the description, the model's interpretation of positive objects is highlighted in red, while gray illustrates the negative description.

corresponding value in the computed mask M in Equation 1 to achieve the final masked image A.

$$A_{xy} = I_{xy} \cdot M_{xy} \tag{2}$$

The result of this process is an image that is blacked out for the most part except for areas that the model considered to be important for its output. Next, the masked image and the predicted label of the vision model are fed to a VLM for assessment and scoring. In this process, VLM plays a critical role in understanding what the vision model does with the image to perform evaluation and scoring. In the pipeline, VLMs are asked to find the relevance between the predicted result of the vision model and the visible object(s) in the masked image and then explain further how the model focuses on different objects. Finally, VLMs are required to score every pair of masked images and labels. This score indicates the ability of the vision model to focus on relevant information for its prediction.

The main reason we use the masked image instead of the heatmap overlay to explain the vision model's attention is to prevent degrading the quality of the image, which can affect the evaluation results of VLMs. Using heatmap overlay can hide away important features of the object(s), thus reducing VLM's ability to understand the attention regions and lowering its accuracy. Furthermore, many VLMs are not trained on images with heatmap overlay and CAM images, which can confuse them in recognizing regions with and without attention. By blacking out the areas without the model's focus and maintaining the remaining areas, we will not sacrifice the image quality on attended objects while ensuring that VLMs can only see and focus on the main regions they need for the evaluation.

3.2 Evaluation and Scoring Criteria

In our pipeline, VLM is crucial in understanding vision models through samples and evaluating their abilities. However, their evaluation results can be inconsistent with different samples. Every time the input context, including the masked image and the label, changes, their evaluation criteria can vary, either more strict or tolerant. This raises the problem of inaccurate evaluation and comparison of models.

To tackle this problem, we create a set of criteria and specify how the model should score for each case. The target is to ensure the evaluations and scores given by VLMs are consistent between all vision models and all samples (consistency for every masked image and label pair). First, we require the VLM to generate a short paragraph about how it comprehends the masked image, which parts the

vision model focuses on, and whether it is relevant to the predicted result of the vision model. This step enhances VLM's ability to understand context, giving it a clear view of the performance of the vision model. After that, VLMs are required to generate a justification and score based on our defined criteria. The criteria are shown in Table 1.

Score	Description
0	Scattered attention; no clear target, indicating a lack of understanding.
1	Attention towards unrelated elements, showing a misunderstanding of the object.
2	Captures a small fragment of the object, missing critical features.
3	Limited identification of object, with irrelevant parts included.
4	Predominantly focuses on object, with minor distractions.
5	Accurately captures the entire object with no distractions.

Table 1: Our criteria for VLMs to score vision models during the evaluation process.

The short paragraph generated from VLM helps increase the accuracy of VLMs during the assessment process by ensuring that the model will not conclude until it has gathered enough evidence, similar to CoT (Wei et al., 2024). This description is prompted to include supportive information about the masked image and reduce the variance and indecision of VLMs when scoring. The short description contains the following information:

- Focus Accuracy: Which areas of the image are focused by the vision model?
- Object Recognition: Is the focused object correctly recognized by the model?
- Object Coverage: Does the attention map cover all or most parts of targets?
- Irrelevant Focus: How many irrelevant objects are captured by the attention map?
- Explanation: What is the reason behind the model's attention, is it fooled by the image?

Overall, the information in the VLMs descriptions covers a wide range of knowledge about the pair of masked images and labels, including focus areas, targets, and some explanation of the model's attention. Those data enrich the context, create more evidence for VLMs, and further guide them to score less biased, more consistent, and accurate.

4 Experiment

In this section, we create a small benchmark, XVisionMini, with our pipeline, and evaluate various widely known vision models. The experiment aims to show the practicals of the proposed pipeline and provide a different aspect of the model's abilities, besides their performances on a specific task, for further research and development on vision models.

4.1 Experimental Setup

This section describes how our benchmark, XVisionMini, is set up and prepared to evaluate different vision models on their image understanding capabilities. The setup includes an image dataset, two VLMs, three methods for extracting attention maps, a method for selecting the model's attention layer for XAI techniques, and our chosen summarizing metrics. Finally, we provide a list of the vision models for evaluation in this experiment.

4.1.1 Dataset and Evaluation VLMs

For our benchmark, we first need images to obtain attention maps from the vision models, which will be used to analyze their capabilities. To achieve this, we randomly selected 500 photos from the test set of the ILSVRC dataset (Addison Howard, 2018). Next, we chose several VLMs to evaluate vision models on masked images. GPT-4o-mini and Gemini-1.5-flash were selected because of their natural language processing and image understanding abilities. Additionally, their cost-effectiveness is crucial for creating a benchmark, conducting various experiments, and generating detailed explanations, justifications, and scores to assess the image understanding capabilities of different models.

4.1.2 XAI Methods

Choosing XAI methods to understand vision models is also important for this benchmark as they affect given judgment and scores by VLMs. Furthermore, each combination of a vision model and an XAI method results in different evaluation outcomes. While some combinations can help explain the model correctly, others may not work very well to provide the correct attention map of the model. For that reason, choosing one method to inspect the model's behaviors and evaluate the model's

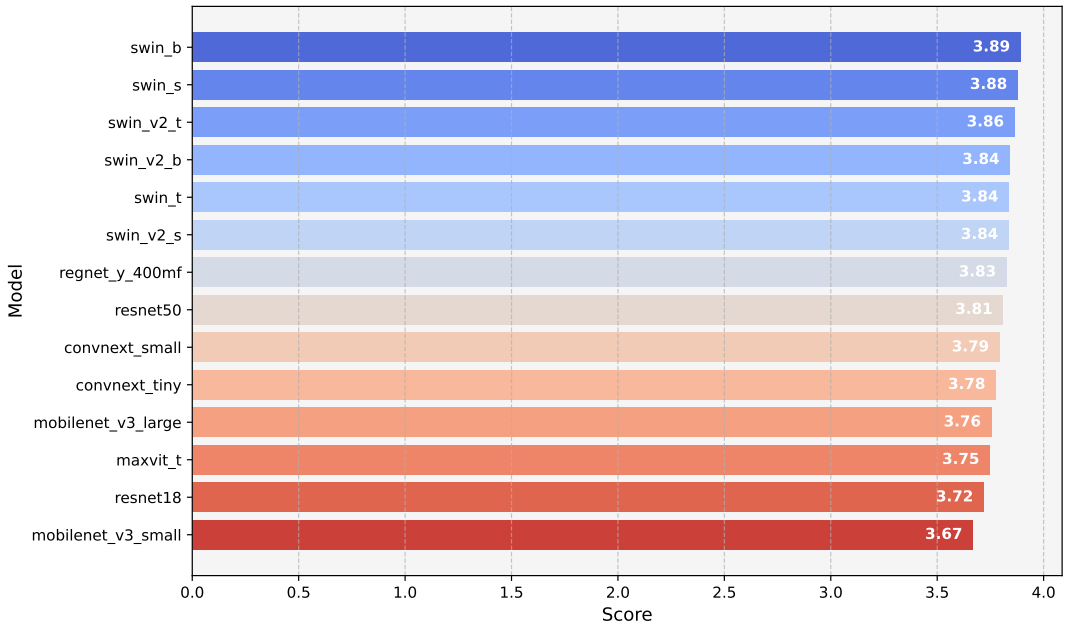


Figure 2: Average maximum results of 14 vision models on our benchmark represent the final evaluation of their image understanding capabilities.

image understandability can result in bias and incorrect judgments, thus reducing the quality of the benchmark. Consequently, we choose three XAI methods, including GradCAM, GradCAM++, and SmoothGradCAM++, to provide more diversity and prevent using only incompatible combinations of model and explanation methods for assessment.

4.1.3 Explained Layer of Vision Models

To extract the attention map of each vision model, it is necessary to choose a target layer to get the information. While early layers of the model might not gather enough information to explain the model behaviors, using layers in the early stages of vision models could lead to unexpected behaviors and lower the accuracy of the evaluation process and general score of that model. Furthermore, layers close to the model output provide much better visualization to model attention and therefore, are used to extract the saliency map in this experiment. In addition, the last layers of vision models gather much knowledge of the model on an image from many previous steps, including spatial information, which makes them more meaningful and provides better visual explanations for VLMs to evaluate.

4.1.4 Evaluation Metrics

As each model has many different scores from different combinations of VLMs and XAI methods, it is necessary to create metrics to combine those results and provide a general understanding of vision models. In this experiment, we use three metrics to combine the results, including the average score, the max score, and the average of max scores. However, only the average of max scores method is chosen due to its effectiveness in providing accurate evaluation, the remaining two methods are used to provide extra information on vision models.

The initial idea for combining the results in this benchmark is to use the average score of all results (Average). The benefit of this method is that it considers all results, and prevents bias toward any single method, which could lead to an unfair judgment of vision models. Despite this solution being widely applied in different benchmarks, in XVisionMini, this method might not be a good option due to the varying compatibility between vision models and different XAI methods. For example, some models can be well explained by GradCAM but not by SmoothGradCAM or CAM methods. Consequently, averaging all results can lower the grade of the vision model that is well explainable with different explanation techniques. In contrast, models with high compatibility with many XAI methods can have high scores despite their average results for all methods. Such a scenario does not help explain and evaluate vision models on their image understandability and can provide bias and false judgment. As our idea in this benchmark is

assessing the model's performance in image understandability, this metric is not useful and therefore, we must look for different combination methods.

To address this, we use two different methods to combine the evaluation results: the Max Score and the Average of Max Scores. The Max Score takes the highest score from the evaluation results as the model's final grade. This ensures that the model is graded using the most suitable XAI method, which helps explain the attention ability effectively. However, a limitation of this method in our benchmark is that it only considers the result of one VLM, potentially leading to inaccurate results due to the hallucination and bias of existing language models. To tackle this, we take max scores evaluated by each VLM and then take the average of those, this result can ensure the incompatible combination of vision models and model introspection techniques will not be used in the final result while reducing the bias and hallucination of one VLM by incorporating many VLMs in the evaluation process.

4.1.5 Experiment Vision Models

This experiment evaluates 14 vision models, including families such as Swin Transformer (Liu et al., 2021), Swin Transformer V2 (Liu et al., 2022a), ResNet (He et al., 2016), RegNet (Xu et al., 2023), MobileNetV3 (Howard et al., 2019), MaxViT (Tu et al., 2022), and ConvNeXt (Liu et al., 2022b). This selection contains six transformer-based models, seven convolution-based models, and one hybrid model. Furthermore, this lineup has a diversity of vision models from a traditional one like ResNet to modern models such as Swin Transformer and ConvNeXt. This diversity helps us understand how different architecture and modification of each vision model affect their focus abilities on the target and utilize image features for their prediction. The comparison between models in this setup can provide comprehensive insights and detailed explanations of the model's strengths and weaknesses.

4.2 Experimental Result

The full experiment results are reported in Table 2 while our ranking of models is shown in Figure 2. Particularly, Figure 2 shows the detailed score each model can achieve from different VLMs and XAI methods combination and the overall scores of vision models on three different metrics, including Average Score, Max Score, and Average of Max Scores. In general, vision models with attention architecture are highly evaluated by both VLMs on various scenarios as we compare them with convolution-based models. Among the transformer-based models, the attention map of the Swin transformer family, especially the small and base versions, demonstrates a better and more accurate attention map to explain their final answers compared to others. Moreover, as the size of the Swin model increases, Figure 2 shows that the explainability of the model on images also increases, which can be the main reason behind the improved performance.

Meanwhile, the Swin Transformer V2 family has lower overall scores than the original Swin Transformer in this evaluation. Interestingly, all versions of Swin Transformer V2 achieve similar and comparable scores on this benchmark despite the differences in the model's size. This might suggest that the differences in their performance might not come from the difference in their ability to focus on the right object. Furthermore, this also shows that the attention map of this family might not be affected by the differences in the model's size. Further research on the phenomenon can lead to new techniques to improve image models in image comprehension and overall performance.

Next, the tiny version of the MaxViT model, a hybrid of transformer and CNN architecture, achieves a pretty low score across all metrics despite its strong performance on ImageNet1k. Its overall score in this benchmark is 3.75, just higher than the results of ResNet18 and MobileNetV3 small, indicating its attention map is less accurate and the model uses more information from surrounding elements for its prediction. However, this result can be affected by the small size or the incompatibility of the MaxViT's architecture with the three XAI methods used in this experiment.

In contrast to transformer-based models, vision models with convolution-based architecture produce less accurate attention maps. This result is understandable due to the difference between their structure and ideology. Furthermore, among the convolution-based models, the RegNet model presents the most precise attention map with a 3.83 overall score. This result is compatible with the small and base versions of Swin Transformer V2, indicating a strong understanding of images and objects of the RegNet model. Furthermore, ResNet50 surprisingly has a high score in this benchmark at 3.81 despite being older than many models. Next, the ConvNeXt family, including the tiny and small versions, have good explainability for their predicted results with 3.78 and 3.79 overall scores

Model	GPT-4o-mini			Gemini-1.5-flash					
	GradCAM	*GradCAM++*	*SGradCAM++*	*GradCAM*	*GradCAM++*	*SGradCAM++*	*Average*	*Max*	***Average Max***
Transformers-based									
swin_v2_t	3.80	3.84	3.78	3.89	3.88	3.62	3.80	3.89	3.86
swin_v2_s	3.81	3.81	3.78	3.86	3.86	3.60	3.79	3.86	3.84
swin_v2_b	3.79	3.79	3.76	3.87	3.90	3.65	3.79	3.90	3.84
swin_t	3.82	3.83	3.76	3.85	3.84	3.66	3.79	3.85	3.84
swin_s	3.83	3.83	3.80	**3.93**	**3.91**	**3.72**	3.83	**3.93**	3.88
swin_b	**3.86**	**3.86**	**3.83**	3.92	**3.91**	**3.72**	**3.85**	3.92	**3.89**
Hybrid									
maxvit_t	3.76	3.56	3.55	3.74	2.75	2.46	3.30	3.76	3.75
Convolution-based									
resnet50	3.76	3.73	3.74	3.83	3.86	3.64	3.76	3.86	3.81
resnet18	3.64	3.66	3.57	3.78	3.77	3.53	3.66	3.78	3.72
regnet_y_400mf	3.76	3.79	3.72	3.86	3.86	3.70	3.78	3.86	3.83
mobilenet_v3_small	3.58	3.64	3.51	3.65	3.70	3.45	3.59	3.70	3.67
mobilenet_v3_large	3.71	3.72	3.67	3.74	3.80	3.66	3.71	3.80	3.76
convnext_tiny	3.75	3.79	3.69	3.74	3.77	3.50	3.71	3.79	3.78
convnext_small	3.77	3.64	3.59	3.82	2.09	3.18	3.35	3.82	3.79

Table 2: Benchmark of models using GPT-4o-mini and Gemini-1.5-flash and XAI methods with average scores. In the table, the Smooth GradCAM method is denoted as SGradCAM++. The highest scores are highlighted in **bold**.

respectively. Lastly, MobileNetV3 models and ResNet18 come in the last positions, together with the MaxViT tiny model, showing that their attention may not always be on their target elements.

5 Conclusion

In this paper, we proposed a novel framework that integrates CAM visualizations with LLM-based scoring to evaluate the interpretability of vision models. Furthermore, we also propose a new benchmark based on our pipeline and evaluate 14 vision models of different architectures. The result shows that our pipeline can assess the model's abilities in image understanding and provide a more comprehensive evaluation of vision models. Finally, the benchmark offers a more detailed analysis of the model's performance, including a description and quantitative assessment, which is useful for further research to understand vision models.

6 Limitations

Even though the proposed pipeline can help evaluate, explain, and advance the development of vision models, our benchmark, the XVisionMini, has limitations that should be addressed in the future.

The first limitation of the XVisionMini is the number of images in the dataset, which is just 500 images collected from the ImageNet1k dataset. Despite enough to examine our pipeline, a larger version of this dataset is necessary to provide more accurate results in explaining and evaluating the image comprehension ability of vision models. Future similar benchmarks should address this problem.

Finally, only three CAM methods were used to extract the attention map of vision models on an image. While they work quite well with vision models in the experiment, more model explanation methods should be included to provide better visualization for different vision models. This can help reduce using inappropriate model explanation techniques, which leads to incorrect evaluation of vision models. Furthermore, some vision models cannot extract attention maps using existing XAI techniques because of architectures. Further research should address this problem to evaluate image understanding abilities for all vision models.

7 Potential risk

While large language models (LLMs) have been used in evaluation in many NLP benchmarks due to the advancement of ChatGPT and Gemini, VLMs, a version of LLM with vision module integration, have not been widely used for evaluation. This may need more time to inspect their advantages and disadvantages.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Wendy Kan Addison Howard, Eunbyung Park. 2018. Imagenet object localization challenge.

Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. 2015. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7):e0130140.

Ge Bai, Jie Liu, Xingyuan Bu, Yancheng He, Jiaheng Liu, Zhanhui Zhou, Zhuoran Lin, Wenbo Su, Tiezheng Ge, Bo Zheng, and Wanli Ouyang. 2024. MT-bench-101: A fine-grained benchmark for evaluating large language models in multi-turn dialogues. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7421–7454, Bangkok, Thailand. Association for Computational Linguistics.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenhang Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, K. Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Yu Bowen, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xing Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023. Qwen technical report. *ArXiv*, abs/2309.16609.

Katarzyna Borys, Yasmin Alyssa Schmitt, Meike Nauta, Christin Seifert, Nicole Krämer, Christoph M Friedrich, and Felix Nensa. 2023. Explainable ai in medical imaging: An overview for clinical practitioners–saliency-based xai approaches. *European journal of radiology*, 162:110787.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. 2020. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer.

Aditya Chattopadhyay, Anirban Sarkar, Prantik Howlader, and Vineeth N. Balasubramanian. 2017. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 839–847.

Dongping Chen, Ruoxi Chen, Shilin Zhang, Yaochen Wang, Yinuo Liu, Huichi Zhou, Qihui Zhang, Yao Wan, Pan Zhou, and Lichao Sun. 2024. MLLM-as-a-judge: Assessing multimodal LLM-as-a-judge with vision-language benchmark. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 6562–6595. PMLR.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. 2016. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Mingyu Ding, Bin Xiao, Noel Codella, Ping Luo, Jingdong Wang, and Lu Yuan. 2022. Davit: Dual attention vision transformers. In *European conference on computer vision*, pages 74–92. Springer.

Alexey Dosovitskiy. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Yuxin Fang, Bencheng Liao, Xinggang Wang, Jiemin Fang, Jiyang Qi, Rui Wu, Jianwei Niu, and Wenyu Liu. 2021. You only look at one sequence: Rethinking transformer in vision through object detection. *Advances in Neural Information Processing Systems*, 34:26183–26197.

Ross Girshick. 2015. Fast r-cnn. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 1440–1448.

David Gunning and David Aha. 2019. Darpa's explainable artificial intelligence (xai) program. *AI magazine*, 40(2):44–58.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.

Andrew Howard, Mark Sandler, Bo Chen, Weijun Wang, Liang-Chieh Chen, Mingxing Tan, Grace Chu, Vijay Vasudevan, Yukun Zhu, Ruoming Pang, Hartwig Adam, and Quoc Le. 2019. Searching for mobilenetv3. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1314–1324.

Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. 2017. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708.

Laurent Itti, Christof Koch, and Ernst Niebur. 1998. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on pattern analysis and machine intelligence*, 20(11):1254–1259.

Albert Qiaochu Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, L'elio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *ArXiv*, abs/2310.06825.

Peng-Tao Jiang, Chang-Bin Zhang, Qibin Hou, Ming-Ming Cheng, and Yunchao Wei. 2021. Layercam: Exploring hierarchical class activation maps for localization. *IEEE Transactions on Image Processing*, 30:5875–5888.

Ioannis Kakogeorgiou and Konstantinos Karantzalos. 2021. Evaluating explainable artificial intelligence methods for multi-label deep learning classification tasks in remote sensing. *International Journal of Applied Earth Observation and Geoinformation*, 103:102520.

Hong-Sik Kim and Inwhee Joe. 2022. An xai method for convolutional neural networks in self-driving cars. *PLoS one*, 17(8):e0267282.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25.

Matthias Kümmerer, Lucas Theis, and Matthias Bethge. 2014. Deep gaze i: Boosting saliency prediction with feature maps trained on imagenet. *arXiv preprint arXiv:1411.1045*.

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

Seongyun Lee, Seungone Kim, Sue Hyun Park, Geewook Kim, and Minjoon Seo. 2024. Prometheus-vision: Vision-language model as a judge for fine-grained evaluation. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 11286–11315, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Yuan-Fang Li, Sébastien Bubeck, Ronen Eldan, Allison Del Giorno, Suriya Gunasekar, and Yin Tat Lee. 2023. Textbooks are all you need ii: phi-1.5 technical report. *ArXiv*, abs/2309.05463.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer.

Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, Furu Wei, and Baining Guo. 2022a. Swin transformer v2: Scaling up capacity and resolution. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11999–12009.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9992–10002.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. 2022b. A convnet for the 2020s. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11966–11976.

Hui Wen Loh, Chui Ping Ooi, Silvia Seoni, Prabal Datta Barua, Filippo Molinari, and U Rajendra Acharya. 2022. Application of explainable artificial intelligence for healthcare: A systematic review of the last decade (2011–2022). *Computer Methods and Programs in Biomedicine*, 226:107161.

Yujie Lu, Dongfu Jiang, Wenhu Chen, William Yang Wang, Yejin Choi, and Bill Yuchen Lin. 2024. Wildvision: Evaluating vision-language models in the wild with human preferences. *arXiv preprint arXiv:2406.11069*.

Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30.

Qing Lyu, Shreya Havaldar, Adam Stein, Li Zhang, Delip Rao, Eric Wong, Marianna Apidianaki, and Chris Callison-Burch. 2023. Faithful chain-of-thought reasoning. *arXiv preprint arXiv:2301.13379*.

Huan Ma, Changqing Zhang, Yatao Bian, Lemao Liu, Zhirui Zhang, Peilin Zhao, Shu Zhang, Huazhu Fu, Qinghua Hu, and Bingzhe Wu. 2023. Fairness-guided few-shot prompting for large language models. *Advances in Neural Information Processing Systems*, 36:43136–43155.

Gemma Team Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, L. Sifre, Morgane Riviere, Mihir Kale, J Christopher Love, Pouya Dehghani Tafti, L'eonard Hussenot, Aakanksha Chowdhery, Adam Roberts, Aditya Barua, Alex Botev, Alex Castro-Ros, Ambrose Slone, Am'elie H'eliou, Andrea Tacchetti, Anna Bulanova, Antonia Paterson, Beth Tsai, Bobak Shahriari, Charline Le Lan, Christopher A. Choquette-Choo, Cl'ement Crepy, Daniel Cer, Daphne Ippolito, David Reid, Elena Buchatskaya, Eric Ni, Eric Noland, Geng Yan, George Tucker, George-Christian Muraru, Grigory Rozhdestvenskiy, Henryk Michalewski, Ian Tenney, Ivan Grishchenko, Jacob Austin, James Keeling, Jane Labanowski, Jean-Baptiste Lespiau, Jeff Stanway, Jenny Brennan, Jeremy Chen, Johan Ferret, Justin Chiu, Justin Mao-Jones, Katherine Lee, Kathy Yu, Katie Millican, Lars Lowe Sjoesund, Lisa Lee, Lucas Dixon, Machel Reid, Maciej Mikula, Mateo Wirth, Michael Sharman, Nikolai Chinaev, Nithum Thain, Olivier Bachem, Oscar Chang, Oscar Wahltinez, Paige Bailey, Paul Michel, Petko Yotov, Pier Giuseppe Sessa, Rahma Chaabouni, Ramona Comanescu, Reena Jana, Rohan Anil, Ross McIlroy, Ruibo Liu, Ryan Mullins, Samuel L Smith, Sebastian Borgeaud, Sertan Girgin, Sholto Douglas, Shree Pandya, Siamak Shakeri, Soham De, Ted Klimenko, Tom Hennigan, Vladimir Feinberg, Wojciech Stokowiec, Yu hui Chen, Zafarali Ahmed, Zhitao Gong, Tris Brian Warkentin, Ludovic Peran, Minh Giang, Cl'ement Farabet, Oriol Vinyals, Jeffrey Dean, Koray Kavukcuoglu, Demis Hassabis, Zoubin Ghahramani, Douglas Eck, Joelle Barral, Fernando Pereira, Eli Collins, Armand Joulin, Noah Fiedel, Evan Senter, Alek Andreev, and Kathleen Kenealy. 2024. Gemma: Open models based on gemini research and technology. *ArXiv*, abs/2403.08295.

Dang Minh, H Xiang Wang, Y Fen Li, and Tan N Nguyen. 2022. Explainable artificial intelligence: a comprehensive review. *Artificial Intelligence Review*, pages 1–66.

Truong Thanh Hung Nguyen, Tobias Clement, Phuc Truong Loc Nguyen, Nils Kemmerzell, Van Binh Truong, Vo Thanh Khang Nguyen, Mohamed Abdelaal, and Hung Cao. 2024. Langxai: Integrating large vision models for generating textual explanations to enhance explainability in visual perception tasks. *arXiv preprint arXiv:2402.12525*.

Maxime Oquab, Leon Bottou, Ivan Laptev, and Josef Sivic. 2015. Is object localization for free? - weakly-supervised learning with convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Tan-Hanh Pham, Xianqi Li, and Kim-Doang Nguyen. 2024. Seunet-trans: A simple yet effective unet-transformer model for medical image segmentation. *IEEE Access*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Kanchana Ranasinghe and Michael S Ryoo. 2023. Language-based action concept spaces improve video self-supervised learning. *Advances in Neural Information Processing Systems*, 36:74980–74994.

Laria Reynolds and Kyle McDonell. 2021. Prompt programming for large language models: Beyond the few-shot paradigm. In *Extended abstracts of the 2021 CHI conference on human factors in computing systems*, pages 1–7.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144.

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626.

Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. 2015. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9.

Mingxing Tan and Quoc Le. 2019. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. Llama: Open and efficient foundation language models. *ArXiv*, abs/2302.13971.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Zhengzhong Tu, Hossein Talebi, Han Zhang, Feng Yang, Peyman Milanfar, Alan Bovik, and Yinxiao Li. 2022. Maxvit: Multi-axis vision transformer. In *Computer Vision – ECCV 2022*, pages 459–479, Cham. Springer Nature Switzerland.

Haofan Wang, Zifan Wang, Mengnan Du, Fan Yang, Zijian Zhang, Sirui Ding, Piotr Mardziel, and Xia Hu. 2020a. Score-cam: Score-weighted visual explanations for convolutional neural networks. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 111–119.

Haofan Wang, Zifan Wang, Mengnan Du, Fan Yang, Zijian Zhang, Sirui Ding, Piotr Mardziel, and Xia Hu. 2020b. Score-cam: Score-weighted visual explanations for convolutional neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 24–25.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2024. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Jing Xu, Yu Pan, Xinglin Pan, Steven Hoi, Zhang Yi, and Zenglin Xu. 2023. Regnet: Self-regulated network for image classification. *IEEE Transactions on Neural Networks and Learning Systems*, 34(11):9562–9567.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. In *Advances in Neural Information Processing Systems*, volume 36, pages 11809–11822. Curran Associates, Inc.

Rui Zhao, Wanli Ouyang, Hongsheng Li, and Xiaogang Wang. 2015. Saliency detection by multi-context deep learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1265–1274.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Advances in Neural Information Processing Systems*, volume 36, pages 46595–46623. Curran Associates, Inc.

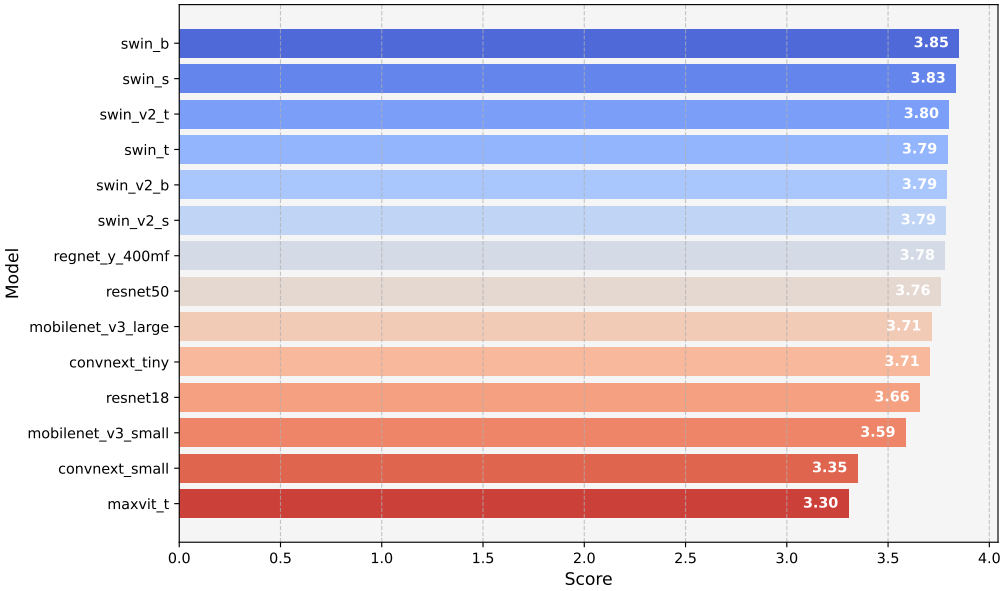
Wanjun Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied, Weizhu Chen, and Nan Duan. 2024. AGIEval: A human-centric benchmark for evaluating foundation models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2299–2314, Mexico City, Mexico. Association for Computational Linguistics.

Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. 2016. Learning deep features for discriminative localization. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2921–2929.

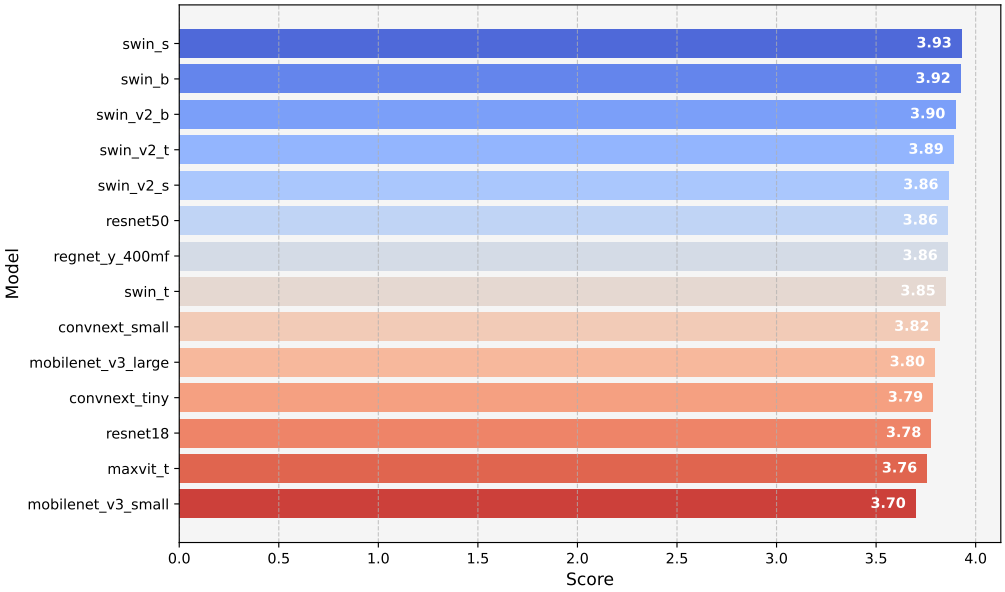
A Appendix

A.1 More Evaluation Result

Figure 3 shows more ranking results with different result combination methods. While the Average method reduces scores of MaxVit and ConvNeXt small due to some incompatible combinations between models and CAM methods, the Max Score can handle this problem effectively.



(a) Ranking results of 14 vision models on our benchmark as we use Average Score.



(b) Ranking results of 14 vision models on our benchmark as we use Max Score.

Figure 3: More ranking results with different score combination methods.

A.2 Examples of Model's Evaluation

We present additional qualitative results of our benchmark to analyze the effectiveness of our method and evaluation metrics. The example shown in Figure 4 demonstrates how the model's attention can sometimes focus on irrelevant features but does not lead to reduced interpretability.

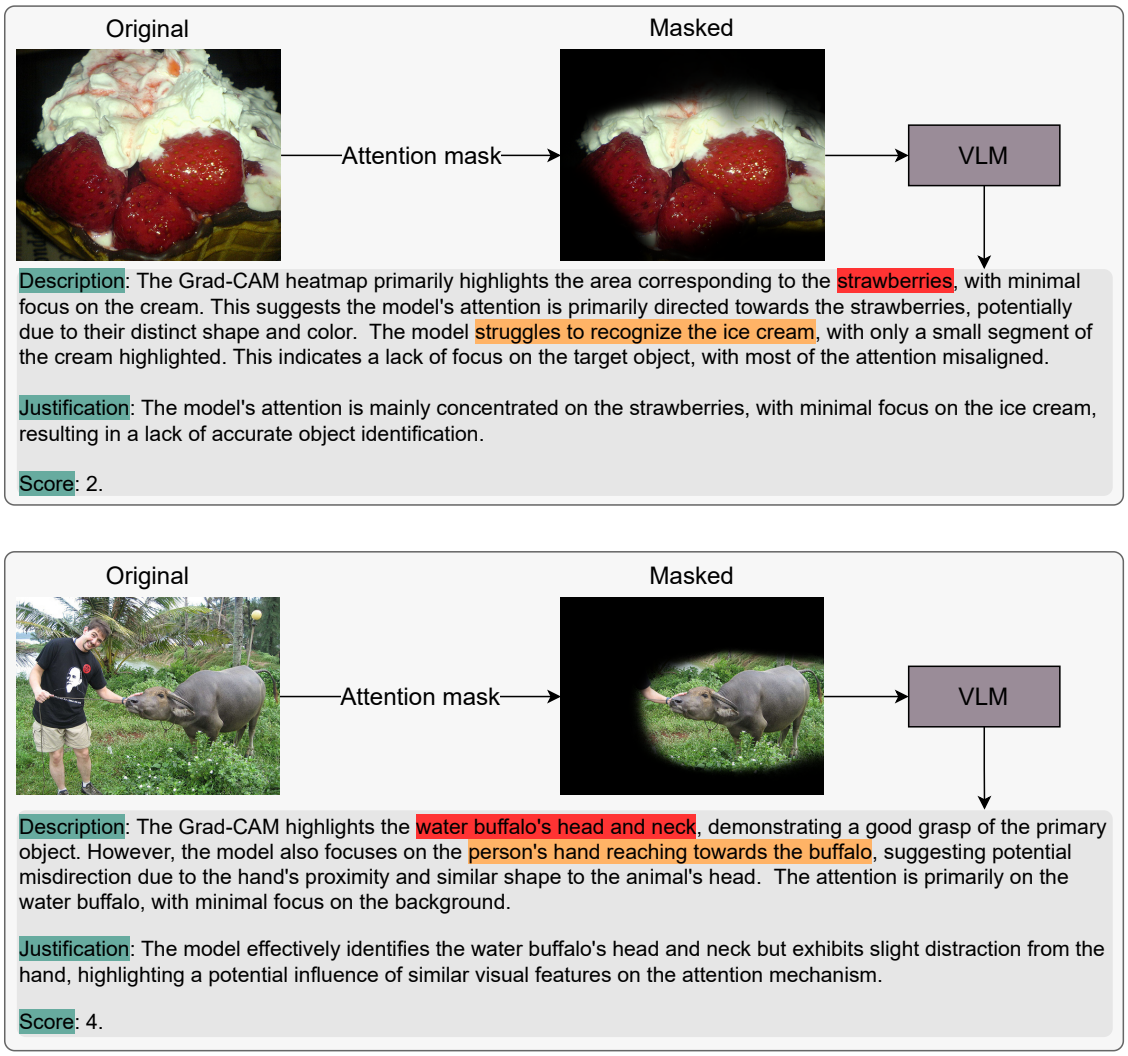


Figure 4: Two prediction examples of the XVisionMini benchmark.

A.3 Prompting

The prompt used for the evaluation framework consists of an image description, evaluation criteria, scoring, and output format. The task involves analyzing a masked image where the model's focused areas are highlighted, while irrelevant regions are blacked out. Key criteria for evaluation include focus accuracy, object recognition, object coverage, and potential distractions from background or irrelevant elements. The evaluator is instructed to analyze the model's attention on the object and provide an explanatory analysis, considering factors like visual challenges or misleading elements. A score from 0 to 5 is assigned, with specific descriptions for each score, reflecting the model's attention and recognition performance. The output includes a concise evaluation and score with justification.

Prompt to get sample description and score

Task: Evaluate the Model's Attention Mechanism Using the Provided Masked Image.

- Image Description:
  - The image is masked with a Grad-CAM heatmap, where only the areas the model focuses on are visible, while all other regions are blacked out.
  - The model is attempting to focus on the object.
- Evaluation Criteria:
  - Focus Accuracy: Analyze which part of the image the Grad-CAM is highlighting. Is the model's attention placed accurately on the object, or is it scattered across other areas?
  - Object Recognition: Determine if the model is correctly recognizing the object. Is the attention primarily on the correct object, or does the model focus on irrelevant areas?
  - Object Coverage: Evaluate how much of the object is being captured by the model's attention. Is the entire object covered, only a small part, or none at all?
  - Background and Irrelevant Focus: Check for any significant focus on the background or irrelevant objects. Does this distract the model from the primary object?
  - Explanatory Analysis: Provide possible reasons for the model's attention pattern. Consider whether the model is being misled by similarly shaped or colored objects, complex backgrounds, or other visual challenges.
- Scoring:

  Assign a score between 0 and 5 based on the relevance and accuracy of the model's attention:
  - 0: The model's attention is completely irrelevant to the object, leading to a wrong result.
  - 1: The model fails to recognize the object entirely, focusing on irrelevant areas.
  - 2: The model captures only a small part of the object.
  - 3: The object is recognized, but the attention also covers irrelevant parts or other objects.
  - 4: Most of the object is detected correctly, with minimal distraction from irrelevant areas or the background.
  - 5: The model perfectly captures the entire object without being distracted by irrelevant areas or the background.
- Output Format:
  - Evaluation: Provide a concise evaluation (5-6 sentences), discussing: Where the Grad-CAM is focusing. Whether the attention aligns with the object. Whether there is any significant focus on irrelevant areas or the background. Explain why the model might be focusing on specific regions.
  - Score: Assign a score from 0 to 5, justifying your rating based on the model's performance in recognizing the object and avoiding distractions.
  - The format must be presented as follows:
    * Evaluation: [evaluation]
    * Score: [score], [justification]