

# LangXAI: Integrating Large Vision Models for Generating Textual Explanations to Enhance Explainability in Visual Perception Tasks

**Hung Nguyen<sup>1,3</sup>, Tobias Clement<sup>2</sup>, Loc Nguyen<sup>2</sup>, Nils Kemmerzell<sup>2</sup>,  
Binh Truong<sup>3</sup>, Khang Nguyen<sup>3</sup>, Mohamed Abdelaal<sup>4</sup>, Hung Cao<sup>1</sup>**

<sup>1</sup>Analytics Everywhere Lab, University of New Brunswick, Canada

<sup>2</sup>Friedrich-Alexander-University Erlangen-Nürnberg, Germany

<sup>3</sup>Quy Nhon AI, FPT Software, Vietnam

<sup>4</sup>Software AG, Germany

hung.ntt@unb.ca, {tobias.clement, loc.pt.nguyen, nils.kemmerzell}@fau.de,  
{binhtv8, khangnvt1}@fpt.com, mohamed.abdelaal@softwareag.com, hcao3@unb.ca

## Abstract

LangXAI is a framework that integrates Explainable Artificial Intelligence (XAI) with advanced vision models to generate textual explanations for visual recognition tasks. Despite XAI advancements, an understanding gap persists for end-users with limited domain knowledge in artificial intelligence and computer vision. LangXAI addresses this by furnishing text-based explanations for classification, object detection, and semantic segmentation model outputs to end-users. Preliminary results demonstrate LangXAI's enhanced plausibility, with high BERTScore across tasks, fostering a more transparent and reliable AI framework on vision tasks for end-users.

## 1 Introduction

In the field of artificial intelligence (AI), making complex AI decisions comprehensible is essential, especially in the application context requiring a large demand for explainability, such as healthcare, and banking. Explainable AI (XAI) is vital for achieving this, yet image-based XAI methods currently demand substantial AI and computer vision (CV) knowledge, often necessitating a domain expert to describe explanations to end-users [Jin *et al.*, 2019; Nguyen *et al.*, 2023b]. Large Vision Models (LVMs), evolving from Large Language Models (LLMs) for visual tasks, present a promising approach to address this issue. LVMs adeptly interpret visual data in human-like ways, improving AI system transparency. Recognizing the potential of XAI and LVM synergy, we aim to develop an innovative XAI framework that incorporates advanced LVMs to elucidate a broad spectrum of CV tasks comprehensively.

More specifically, we introduce LangXAI, a framework that employs LVMs to deliver clear text-based explanations for AI visual decision processes. LangXAI aims to increase the transparency of black-box models, making it easier for users without extensive specialized knowledge to understand AI decisions. Furthermore, during the validation stage, we

employ a diverse set of metrics to thoroughly evaluate the model. Our initiative is dedicated to enhancing both AI transparency and trustworthiness through accessible explanations.

## 2 Related Work

We review two key aspects of our study. First, we investigate the current development of XAI in the fields of CV, with a specific focus on classification, semantic segmentation, and object detection. Additionally, we explore the emerging landscape of LVMs and their rapid development.

### 2.1 Explainable AI (XAI) in CV

XAI methods for CV tasks can be categorized into two distinct groups based on their purposes, as illustrated in Figure 2. While classification and semantic segmentation share common XAI techniques due to their inherent similarities, object detection necessitates different approaches owing to its dual focus on classification and localization. In contrast to classification tasks that take into account the entire image, object detectors primarily employ convolutional layers instead of fully connected ones. This design choice confines the receptive field of the desired output to just a segment of the input image [Truong *et al.*, 2023].

Within the context of working mechanisms, all XAI methods utilized in this study fall into two classes: Gradient-based and Perturbation-based. Gradient-based methods assess feature significance by computing gradients of the output concerning the extracted features using backpropagation. These methods then assign estimated attribution scores, identify the path that maximizes a specific output, and highlight critical input features, such as pixels in this study [Rodrigues *et al.*, 2024]. We take into consideration several gradient-based algorithms, such as GradCAM [Selvaraju *et al.*, 2017], GradCAM++ [Chattopadhyay *et al.*, 2018], SeCAM [Cao *et al.*, 2023], HiResCAM [Draelos and Carin, 2020], and G-CAME [Nguyen *et al.*, 2023a]. In contrast, perturbation-based methods modify input images to track changes in the output. Significant output alternations indicate input relevance, especially when the target class is modified. These methods support iterative testing and offer visualizations of crucial input segments [Paralić *et al.*, 2023]. In this research, several

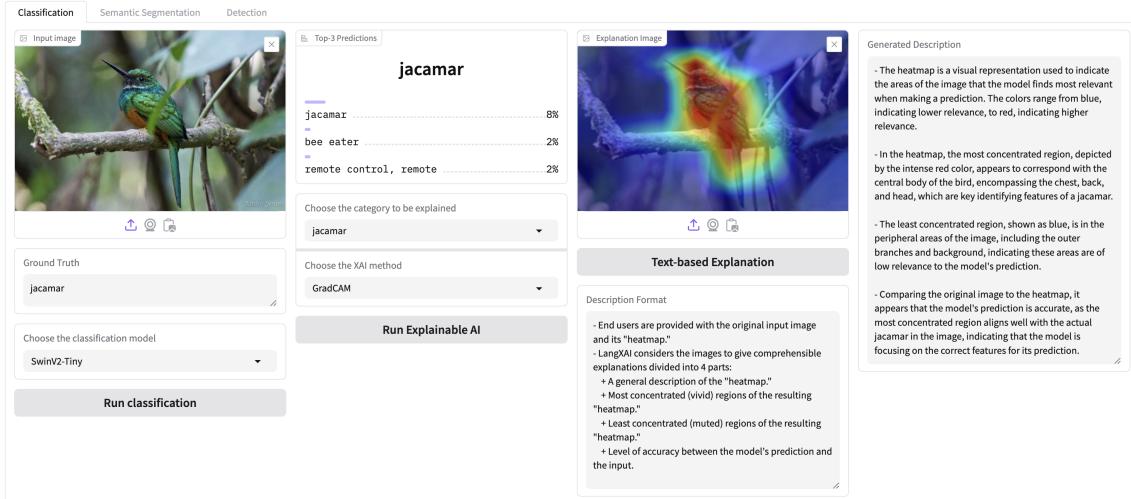


Figure 1: The interface of LangXAI showcases how it operates to make AI decisions in the classification task, which is designed straightforwardly with guidance so end-users can comprehend and monitor end-to-end explanations.

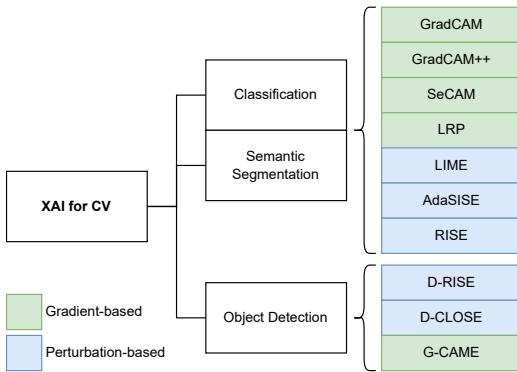


Figure 2: Classification of XAI methods by CV tasks and their mechanisms.

perturbation-based methods have been utilized, namely AdaSISE [Sudhakar *et al.*, 2021], RISE [Petsiuk *et al.*, 2018], D-RISE [Petsiuk *et al.*, 2021], and D-CLOSE [Truong *et al.*, 2023].

## 2.2 Large Vision Models (LVMs)

The advancements in LLMs have given rise to LVMs, which blend language understanding and reasoning with visual perception [Peng *et al.*, 2023]. Initial LVM strategies involve refining visual encoders based on language embeddings or visual-to-text conversion [Wu *et al.*, 2024]. Models such as Flamingo [Alayrac *et al.*, 2022] and PaLM-E [Driess *et al.*, 2023] exemplify the former approach, while techniques for the latter method have been proposed in [Hu *et al.*, 2023] and [Shao *et al.*, 2023]. In this research, we employed a new member of the LVMs family, which is GPT-4 Vision [OpenAI, 2023]. This LVM exhibits robust performance across diverse tasks [Yang *et al.*, 2023] and demonstrates a strong alignment with human evaluators [Zhang *et al.*, 2023].

It is worth noting that, despite the presence of several end-user-centered XAI frameworks and patterns [Jin *et al.*, 2019;

Jin *et al.*, 2021; Schoonderwoerd *et al.*, 2021], contemporary research lacks an attempt to develop a unified framework for providing comprehensive and trustworthy end-to-end explanations to end-users for the saliency-map explanations [Chang *et al.*, 2023; Clement *et al.*, 2024]. To address this gap, this paper aims to introduce a novel XAI framework that leverages the capabilities of LVMs to simplify user-system interaction and offer highly reliable interpretations.

## 3 Framework

Our framework is divided into two main parts, aiming to provide end users with a straightforward approach to comprehending complex image decision explanations. The procedural details of the framework are depicted in Figure 3.

### 3.1 Block 1: Saliency Map Extraction with XAI

The first part of our framework focuses on generating saliency maps using XAI methods from various CV models tailored for different tasks. The process involves uploading an image and selecting the desired task, with specific models assigned accordingly: Swin Transformer v2 [Liu *et al.*, 2022] for classification, DeepLabv3-ResNet50 and ResNet101 [Chen *et al.*, 2017] for semantic segmentation, and Faster R-CNN [Ren *et al.*, 2015] and YOLOX [Ge *et al.*, 2021] for object detection. Following the image analysis, users can specify the predicted class and choose the XAI method to generate saliency maps, which highlights the areas of interest for the model’s decision-making process.

### 3.2 Block 2: Text-based Explanation with LVM

In the second part, we integrate various data to aid the LVM in generating text-based explanations for end-users unfamiliar with AI and CV. The GPT-4 Vision [OpenAI, 2023] serves as the core LVM in our framework, leveraging information such as the input image, ground truth, model’s top-1 prediction, and saliency map. We employ a structured prompt for each task, starting with presenting the image and saliency

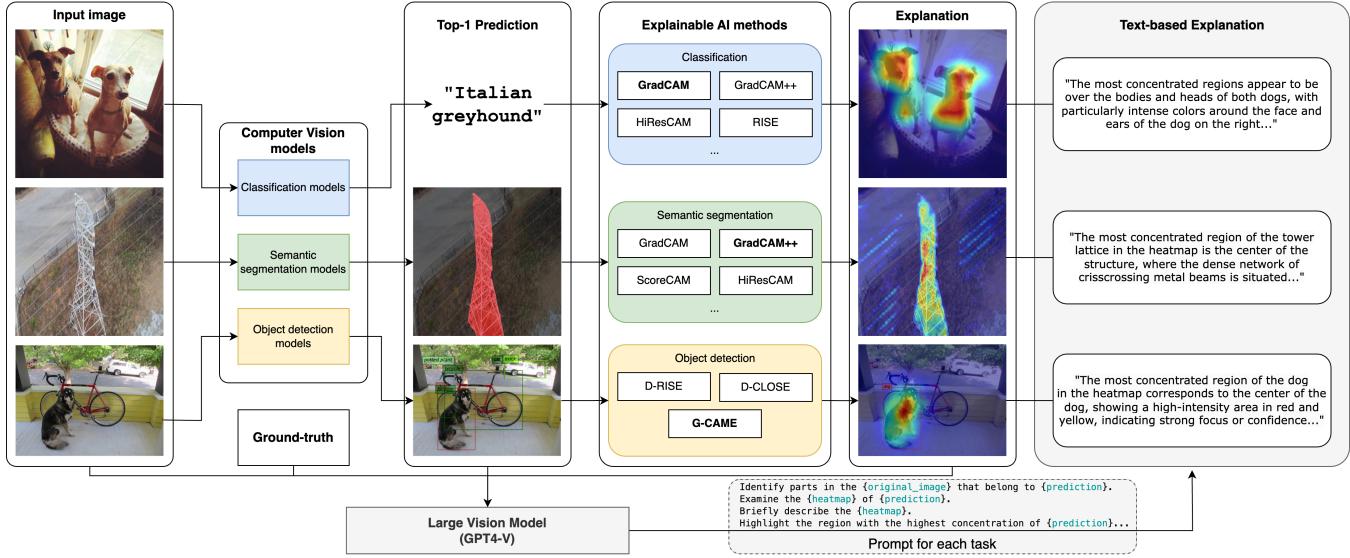


Figure 3: Our framework is split into two parts for explaining decisions made by AI models in CV tasks. The first part (in white blocks) generates saliency maps, where the XAI method in bold is used to generate the saliency map. The second part (in gray blocks) combines the input image, saliency map, ground truth, and prediction to provide a text-based explanation under prompts for each task.

map to help the LVM identify focal areas. We then combine the saliency map with the model’s prediction to verify the accuracy. In the end, we compare the model’s prediction with the ground truth to determine the reliability and assess potential confusion by background or other objects. This comprehensive process ensures explanations are both coherent and indeed based on the model’s visual analysis of the image.

## 4 Evaluation

In this section, we evaluate how well LangXAI’s text-based explanations align with expert interpretations in the field of XAI. We ensure a comprehensive assessment by having a domain expert who has knowledge in AI and XAI context to review and label 5 samples for each task. The datasets chosen for analysis are tailored to each task: ImageNetv2 [Recht *et al.*, 2019] for image classification, TTPLA [Abdelfattah *et al.*, 2020] for semantic segmentation, and MS-COCO 2017 [Lin *et al.*, 2014] for object detection. During the evaluation stage, we employ various metrics, including BLEU [Papineni *et al.*, 2002], METEOR [Banerjee and Lavie, 2005], ROUGE-L [Lin, 2004], and BERTScore [Zhang *et al.*, 2019], to comprehensively measure the performance of the LVM. Among these metrics, we prioritize BERTScore as it evaluates semantic similarity at a deeper level compared to surface lexical matches, providing a more nuanced reflection of expert interpretations. On the other hand, when assessing ROUGE-L and BERTScore, we particularly focus on their precision results.

The evaluation results from Table 1 demonstrate varying performance levels of the LVM across different tasks: image classification, semantic segmentation, and object detection. It is worth noting that BERTScore metrics consistently yield high scores across all tasks, which indicates a robust semantic alignment between the model’s explanations and expert

Task	BLEU	METEOR	ROUGE-L	BERTScore
<b>Classification</b>	0.2971	0.5122	0.5196	0.9341
<b>Semantic Segmentation</b>	0.2552	0.4741	0.4714	0.8594
<b>Object Detection</b>	0.2754	0.4904	0.4911	0.9093

Table 1: The performance of LVM across tasks is measured using BLEU, METEOR, ROUGE-L, and BERTScore metrics. These metrics evaluate how closely the model’s text-based explanations align with expert interpretations in the context of XAI. Higher scores indicate better performance.

interpretations. These results also demonstrate a deep comprehension of the model for the tasks at hand. Specifically, the image classification task receives the highest evaluation scores, likely owing to its straightforward nature of identifying depicted subjects in images. On the other hand, semantic segmentation and object detection demand more intricate explanations. They delve into object positions, the background, surrounding contextual information, and interactions within images, posing challenges in conveying them clearly to non-expert users.

## 5 Conclusion

This paper presents LangXAI, a novel framework combining XAI with advanced vision models to create textual explanations for visual tasks. Our framework can enhance purely visual explanations with natural language and therefore could help narrow the knowledge gap for users with limited AI expertise, highlighted by an average BERTScore of 0.9, indicating its potential to improve AI system transparency and reliability.

## References

- [Abdelfattah *et al.*, 2020] Rabab Abdelfattah, Xiaofeng Wang, and Song Wang. Ttpla: An aerial-image dataset for detection and segmentation of transmission towers and power lines. In *Proceedings of the Asian Conference on Computer Vision*, 2020.
- [Alayrac *et al.*, 2022] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. Flamingo: a visual language model for few-shot learning, 2022.
- [Banerjee and Lavie, 2005] Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72, 2005.
- [Cao *et al.*, 2023] Quoc Hung Cao, Truong Thanh Hung Nguyen, Vo Thanh Khang Nguyen, and Xuan Phong Nguyen. A novel explainable artificial intelligence model in image classification problem, 2023.
- [Chang *et al.*, 2023] Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Wei Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qiang Yang, and Xing Xie. A survey on evaluation of large language models, 2023.
- [Chattopadhyay *et al.*, 2018] Aditya Chattopadhyay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, March 2018.
- [Chen *et al.*, 2017] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017.
- [Clement *et al.*, 2024] Tobias Clement, Truong Thanh Hung Nguyen, Mohamed Abdelaal, and Hung Cao. Xai-enhanced semantic segmentation models for visual quality inspection. *arXiv preprint arXiv:2401.09900*, 2024.
- [Draelos and Carin, 2020] Rachel Lea Draelos and Lawrence Carin. Use hirescam instead of grad-cam for faithful explanations of convolutional neural networks. *arXiv preprint arXiv:2011.08891*, 2020.
- [Driess *et al.*, 2023] Danny Driess, Fei Xia, Mehdi S. M. Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, Wenlong Huang, Yevgen Chebotar, Pierre Sermanet, Daniel Duckworth, Sergey Levine, Vincent Vanhoucke, Karol Hausman, Marc Toussaint, Klaus Greff, Andy Zeng, Igor Mordatch, and Pete Florence. Palm-e: An embodied multimodal language model, 2023.
- [Ge *et al.*, 2021] Zheng Ge, Songtao Liu, Feng Wang, Zeming Li, and Jian Sun. Yolox: Exceeding yolo series in 2021. *arXiv preprint arXiv:2107.08430*, 2021.
- [Hu *et al.*, 2023] Yushi Hu, Hang Hua, Zhengyuan Yang, Weijia Shi, Noah A Smith, and Jiebo Luo. Promptcap: Prompt-guided task-aware image captioning, 2023.
- [Jin *et al.*, 2019] Weina Jin, Sheelagh Carpendale, Ghassan Hamarneh, and Diane Gromala. Bridging ai developers and end users: An end-user-centred explainable ai taxonomy and visual vocabularies. *Proceedings of the IEEE Visualization, Vancouver, BC, Canada*, pages 20–25, 2019.
- [Jin *et al.*, 2021] Weina Jin, Jianyu Fan, Diane Gromala, Philippe Pasquier, and Ghassan Hamarneh. Euca: The end-user-centered explainable ai framework. *arXiv preprint arXiv:2102.02437*, 2021.
- [Lin *et al.*, 2014] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. *CoRR*, abs/1405.0312, 2014.
- [Lin, 2004] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004.
- [Liu *et al.*, 2022] Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, et al. Swin transformer v2: Scaling up capacity and resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12009–12019, 2022.
- [Nguyen *et al.*, 2023a] Quoc Khanh Nguyen, Truong Thanh Hung Nguyen, Vo Thanh Khang Nguyen, Van Binh Truong, and Quoc Hung Cao. G-came: Gaussian-class activation mapping explainer for object detectors. *arXiv preprint arXiv:2306.03400*, 2023.
- [Nguyen *et al.*, 2023b] Truong Thanh Hung Nguyen, Van Binh Truong, Vo Thanh Khang Nguyen, Quoc Hung Cao, and Quoc Khanh Nguyen. Towards trust of explainable ai in thyroid nodule diagnosis. *arXiv preprint arXiv:2303.04731*, 2023.
- [OpenAI, 2023] OpenAI. Gpt-4v(ision) system card, sep 2023.
- [Papineni *et al.*, 2002] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002.
- [Paralič *et al.*, 2023] Ján Paralič, Michal Kolářík, Zuzana Paraličová, Oliver Lohaj, and Adam Jozefík. Perturbation-based explainable ai for ecg sensor data. *Applied Sciences*, 13(3), 2023.
- [Peng *et al.*, 2023] Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei.

- Kosmos-2: Grounding multimodal large language models to the world, 2023.
- [Petsiuk *et al.*, 2018] Vitali Petsiuk, Abir Das, and Kate Saenko. Rise: Randomized input sampling for explanation of black-box models, 2018.
- [Petsiuk *et al.*, 2021] Vitali Petsiuk, Rajiv Jain, Varun Manjunatha, Vlad I. Morariu, Ashutosh Mehra, Vicente Ordonez, and Kate Saenko. Black-box explanation of object detectors via saliency maps, 2021.
- [Recht *et al.*, 2019] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *International Conference on Machine Learning*, pages 5389–5400, 2019.
- [Ren *et al.*, 2015] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015.
- [Rodrigues *et al.*, 2024] Caroline Mazini Rodrigues, Nicolas Bouthy, and Laurent Najman. Transforming gradient-based techniques into interpretable methods, 2024.
- [Schoonderwoerd *et al.*, 2021] Tjeerd AJ Schoonderwoerd, Wiard Jorritsma, Mark A Neerincx, and Karel Van Den Bosch. Human-centered xai: Developing design patterns for explanations of clinical decision support systems. *International Journal of Human-Computer Studies*, 154:102684, 2021.
- [Selvaraju *et al.*, 2017] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 618–626, 2017.
- [Shao *et al.*, 2023] Zhenwei Shao, Zhou Yu, Meng Wang, and Jun Yu. Prompting large language models with answer heuristics for knowledge-based visual question answering. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14974–14983, 2023.
- [Sudhakar *et al.*, 2021] Mahesh Sudhakar, Sam Sattarzadeh, Konstantinos N. Plataniotis, Jongseong Jang, Yeonjeong Jeong, and Hyunwoo Kim. Ada-sise: Adaptive semantic input sampling for efficient explanation of convolutional neural networks, 2021.
- [Truong *et al.*, 2023] Van Binh Truong, Truong Thanh Hung Nguyen, Vo Thanh Khang Nguyen, Quoc Khanh Nguyen, and Quoc Hung Cao. Towards better explanations for object detection. *arXiv preprint arXiv:2306.02744*, 2023.
- [Wu *et al.*, 2024] Tong Wu, Guandao Yang, Zhibing Li, Kai Zhang, Ziwei Liu, Leonidas Guibas, Dahua Lin, and Gordon Wetzstein. Gpt-4v(ision) is a human-aligned evaluator for text-to-3d generation, 2024.
- [Yang *et al.*, 2023] Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. The dawn of Imms: Preliminary explorations with gpt-4v(ision), 2023.
- [Zhang *et al.*, 2019] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*, 2019.
- [Zhang *et al.*, 2023] Xinlu Zhang, Yujie Lu, Weizhi Wang, An Yan, Jun Yan, Lianke Qin, Heng Wang, Xifeng Yan, William Yang Wang, and Linda Ruth Petzold. Gpt-4v(ision) as a generalist evaluator for vision-language tasks, 2023.