NL-EYE: ABDUCTIVE NLI FOR IMAGES

Mor Ventura[1], Michael Toker[1], Nitay Calderon[1], Zorik Gekhman[1,2], Yonatan Bitton[2], and Roi Reichart[1]

[1]Department of Data Science, Technion - IIT,
{mor.ventura,tok,nitay,zorik}@campus.technion.ac.il
[2]Google Research

ABSTRACT

Will a Visual Language Model (VLM)-based bot warn us about slipping if it detects a wet floor? Recent VLMs have demonstrated impressive capabilities, yet their ability to infer outcomes and causes remains underexplored. To address this, we introduce NL-EYE, a benchmark designed to assess VLMs' visual abductive reasoning skills. NL-EYE adapts the abductive Natural Language Inference (NLI) task to the visual domain, requiring models to evaluate the plausibility of hypothesis images based on a premise image and explain their decisions. NL-EYE consists of 350 carefully curated triplet examples (1,050 images) spanning diverse reasoning categories: physical, functional, logical, emotional, cultural, and social. The data curation process involved two steps—writing textual descriptions and generating images using text-to-image models, both requiring substantial human involvement to ensure high-quality and challenging scenes. Our experiments show that VLMs struggle significantly on NL-EYE, often performing at random baseline levels, while humans excel in both plausibility prediction and explanation quality. This demonstrates a deficiency in the abductive reasoning capabilities of modern VLMs. NL-EYE represents a crucial step toward developing VLMs capable of robust multimodal reasoning for real-world applications, including accident-prevention bots and generated video verification.[1]

1 INTRODUCTION

Abductive reasoning refers to the ability to infer and predict plausible outcomes or causes given a context scene Peirce et al. (1934); Fann (2012); Douven (2021). This reasoning skill is crucial for Visual Language Models (VLMs), as they are likely to become increasingly integrated into our daily lives (Yildirim et al., 2024; Anwar et al., 2024; Chiang et al., 2024; Shah et al., 2023). These models will be required to accurately monitor and interpret daily life scenes and correctly infer plausibility to prevent accidents and provide timely advice. For instance, would a bot warn us from slipping on a wet floor when there is no warning sign? or would it infer a missing pacifier as a cause of a crying baby?

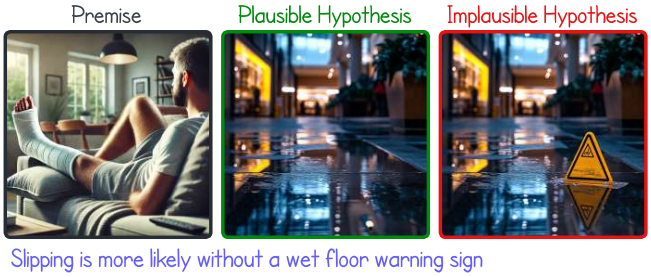

Figure 1: NL-EYE evaluates the abductive reasoning capabilities of VLMs. The main setup involves a premise image and two hypothesis images, where the model is tasked with inferring which hypothesis is more plausible, and to provide an explanation for its choice.

Although this capability is critical, previous work has mainly evaluated VLMs in a *single scene* setting — such as visual entailment or detecting improbable events like a fire in a closed jar — or in *sequential scenes*, such as next-frame prediction Xie et al. (2019); Fu et al. (2022); Hessel et al. (2022); Fu et al. (2024); Ganz et al. (2024); Yarom et al. (2024); Kadiyala et al. (2024). Consequently, it remains unclear to what extent existing VLMs are capable of abductive reasoning.

To address this, we introduce NL-EYE, a benchmark designed to evaluate *visual abductive reasoning* capabilities of VLMs across *multiple images*. NL-EYE is inspired by the textual abductive NLI task Bhagavatula et al. (2019) and applies it to the visual domain. In NL-EYE, a VLM is presented with a premise image and one or two hypothesis images. It then needs to infer how likely (plausible) a hypothesis image is to result from or lead to the premise image. The plausibility evaluation can be either done individually or in comparison to an alternative hypothesis. For instance, in Figure 1, the VLM needs to infer that, given the broken leg in the context image, it is more likely that the man slipped on the wet floor which lacked a warning sign (i.e., selecting hypothesis image 1).

[1]Data and code are available on the project page: https://venturamor.github.io/NLEye/.

Beyond *plausiblity prediction*, NL-EYE facilitates the evaluation of the models' capability to provide faithful explanations. This allows us to explore whether they are correct for the right reasons rather than relying on shallow heuristics McCoy et al. (2019). For example, a valid explanation for the broken leg scene would suggest that the presence of a warning sign would have made the man more alert, thereby potentially preventing the accident. In contrast, a shallow explanation might suggest that the man was simply resting on a cozy rainy day.

Each NL-EYE example features a premise image alongside two hypothesis images, annotated with a gold label indicating the index of the more plausible hypothesis. The examples also include a gold explanation detailing why the chosen hypothesis is more plausible than the alternative. Each example is categorized into one of six *reasoning categories* – physical, logical, emotional, functional, cultural, and social – and includes temporal annotations that specify whether the hypotheses occur *before*, *after*, or *simultaneously* with the premise, and whether the time duration between the premise and hypothesis scenes is *short* or *long*. This rich annotation aids in diagnosing current VLMs and highlights their strengths and weaknesses. Figure 2 presents a detailed example.

To create NL-EYE, we collected a large pool of high-quality textual scenes created by experienced human annotators. The resulting scenes were then provided to professional designers who utilized Midjourney and DALL-E (Ramesh et al., 2021) to synthesize the corresponding images. The designers are also tasked with categorizing each example and creating the explanation that is used as the gold label. The image generation process was iterative, requiring multiple attempts to ensure consistency between the textual descriptions and the visual scenes, as well as visual coherence among the images within the same triplet. This process resulted in a total of 1,050 generated images, yielding 350 image triplets. Overall, NL-EYE is characterized by carefully curated examples, offering high quality both in terms of the scenarios and the consistency and quality of the images.

The first analysis is *human evaluation* where annotators select the more plausible hypothesis and explain their choice. Our results indicate that humans successfully identify the more plausible hypothesis in 85% of the cases. Furthermore, in our assessment of the quality of the human-generated explanations, we find that in 94% of the cases where the correct hypothesis was selected, the humans also provided a valid explanation. This demonstrates that humans perform reasonably well on the NL-EYE tasks.

Next, we design a comprehensive study to evaluate the abductive reasoning abilities of modern VLMs. We take multiple measures to ensure the robustness of our evaluation, including addressing sensitivity to the order in which hypotheses are presented and exploring various input strategies, such as feeding the model three separate images or presenting it with a single combined-image that composites all three. Since real-world scenarios may not always provide two alternatives, we also evaluate the model's ability to assign a plausibility score to a single hypothesis, in addition to comparing two candidates. We have also developed a framework that utilizes a text-based baseline that processes textual descriptions of visual scenes. Specifically, we compare the results with gold descriptions and with the captions of the images as generated by the VLMs. Lastly, evaluating model-generated explanations is challenging, as comparing generated text to a single reference (gold) explanation can be limiting and may not capture the variety and validity of possible correct answers. To address this, we adopt the evaluation proposed by Bitton-Guetta et al. (2023): human annotators are presented with an image triplet where the correct hypothesis is already labeled and select valid explanations from a provided set.

Our results show that while humans perform well on NL-EYE, VLMs struggle, with most models failing to surpass a random baseline in the *plausibility prediction* task. Even when identifying the plausible hypothesis, VLMs fail to provide accurate explanations in over 50% of cases, revealing a major weakness in their abductive reasoning. Furthermore, our text-based experiments indicate that these models often succeed in textual reasoning even when they fail to reason over images. Interestingly, when we prompt the VLMs to generate image captions, the resulting captions prove ineffective for solving the task. Consequently, we hypothesize that the VLMs reasoning is hindered by inaccurate visual interpretations. We also find that these models are sensitive to the order in which the hypotheses are presented and to the input format (three separate images vs a single combined-image). This sensitivity is concerning, as it raises the possibility that the models may not genuinely understand the underlying concepts, potentially relying on superficial cues to make decisions.

To summarize, we introduce NL-EYE a carefully curated benchmark designed to test the abductive reasoning abilities of VLMs across various categories and temporal relations. We then conduct a comprehensive study evaluating modern VLMs on NL-EYE and find notable deficiencies in their abductive reasoning capabilities. We believe NL-EYE represents a crucial step toward enhancing the reasoning abilities of VLMs, moving them closer to truly understanding complex, real-world scenarios and providing more reliable and interpretable outputs.

2 THE NL-EYE BENCHMARK

2.1 TASKS

Our objective is to explore and benchmark the abductive reasoning capabilities of modern VLMs. Unlike much of the previous work in NLP, our focus is on reasoning solely based on visual inputs: premise and hypothesis images.