

WHY CONTEXT MATTERS IN VQA & REASONING: SEMANTIC INTERVENTIONS FOR VLM INPUT MODALITIES

Kenza Amara *^{1,2}, **Lukas Klein** *^{1,3,4}, **Carsten Lüth**^{3,4}, **Paul Jäger**^{3,4}, **Hendrik Strobelt**⁵, **Mennatallah El-Assady**¹

¹ ETH Zurich, ² ETH AI Center, ³ German Cancer Research Center (DKFZ),

⁴ Helmholtz Imaging, ⁵ IBM Research

kenza.amara@ai.ethz.ch, lukas.klein@dkfz.de

ABSTRACT

The various limitations of Generative AI, such as hallucinations and model failures, have made it crucial to understand the role of different modalities in Visual Language Model (VLM) predictions. Our work investigates how the integration of information from image and text modalities influences the performance and behavior of VLMs in visual question answering (VQA) and reasoning tasks. We measure this effect through answer accuracy, reasoning quality, model uncertainty, and modality relevance. We study the interplay between text and image modalities in different configurations where visual content is essential for solving the VQA task. Our contributions include (1) the Semantic Interventions (SI)-VQA dataset, (2) a benchmark study of various VLM architectures under different modality configurations, and (3) the Interactive Semantic Interventions (ISI) tool. The SI-VQA dataset serves as the foundation for the benchmark, while the ISI tool provides an interface to test and apply semantic interventions in image and text inputs, enabling more fine-grained analysis. Our results show that complementary information between modalities improves answer and reasoning quality, while contradictory information harms model performance and confidence. Image text annotations have minimal impact on accuracy and uncertainty, slightly increasing image relevance. Attention analysis confirms the dominant role of image inputs over text in VQA tasks. In this study, we evaluate state-of-the-art VLMs that allow us to extract attention coefficients for each modality. A key finding is PaliGemma's harmful overconfidence, which poses a higher risk of silent failures compared to the LLaVA models. This work sets the foundation for rigorous analysis of modality integration, supported by datasets specifically designed for this purpose. The code is available at <https://gitlab.com/dekfsx1/si-vlm-benchmark> and the tool and dataset are hosted at <https://gitlab.com/dekfsx1/isi-vlm>.

1 INTRODUCTION

Vision-Language Models (VLMs) have shown remarkable performance across various NLP tasks by combining visual and textual information. Models like LLaVA (Liu et al., 2023a), GPT4-Vision (OpenAI, 2024), and PaliGemma (Beyer et al., 2024) use the text interface of Large Language Models (LLMs) alongside CLIP-style image encoders (Radford et al., 2021), making them well-suited for multi-modal tasks such as content summarization (Moured et al., 2024), text-guided object detection (Dorkenwald et al., 2024), and visual question answering (VQA) (Yue et al., 2024b).

The core principle of VLMs lies in the integration and interplay of the two modalities, which significantly amplifies the effectiveness of the models. However, the extent to which each modality influences the final answer and reasoning of VLMs remains unclear within the research community, with varying and sometimes conflicting conclusions (Gat et al., 2021; Frank et al., 2021; Chen et al., 2024). Prior research has highlighted the dominance of textual data (Parcalabescu & Frank, 2023), but such studies often operate in settings where each modality, i.e., image or text, could independently be used to solve a task. This definition of multimodality, focusing on two independent but informative inputs, limits our understanding of how these modalities interact. Establishing a dataset and a tool to benchmark and examine the role of input modalities in model predictions within a multimodal setting would therefore significantly benefit the community. This approach would not only aid in understanding and mitigating model failures, such as hallucinations but also guide researchers toward practical solutions for improving model performance and self-consistency.

To this end, we present a benchmark, investigating the effect of semantic interventions on text and image modalities, alongside a carefully curated dataset and an interactive tool. Specifically, we investigate the role of the text modality

*Contributed Equally

as a context guide in VQA tasks. We make the image the necessary data source to answer the question while using the text in different configurations to affect the model’s answer and reasoning. Initially, we develop the **Interactive Semantic Interventions (ISI) Tool** to perform inter-modality interventions, testing their effect on the model behavior and observing which interventions lead to model failure. Based on the results, we curate the **Semantic Interventions (SI)-VQA dataset**. It contains 100 cautiously crafted examples with controlled interventions in both the input image and context. Each instance includes an image, the same image annotated, a complementary or contradictory textual context, and a question with a ground truth Yes/No answer, which can only be inferred from the image. This dataset allows for a wide range of image-text combinations. Using the SI-VQA Dataset, we establish a **comprehensive benchmark** to assess the impact of various interventions on VLMs across multiple evaluation criteria. We experiment with different configurations where text serves as either complementary or contradictory context to the image, or as annotations directly on the image. We then analyze the model’s performance, uncertainty, and attention attribution across these modalities. Specifically, we examine the latest open-source VLMs including LLaVA 1.5 (Liu et al., 2023a), LLaVA-Vicuna (Liu et al., 2024a), LLaVA-NeXT (Liu et al., 2024b), and PaliGemma (Beyer et al., 2024).

Our study demonstrates that integrating complementary contextual information into VQA models enhances both their answer accuracy and reasoning quality, while contradictory information significantly degrades performance and confidence, comparable to the absence of visual input. Additionally, we find that these models inherently prioritize visual data over textual context; efforts to rebalance attention toward textual information—such as adding image annotations, textual descriptions, or prompt engineering—yield mixed results, with prompt engineering improving accuracy without altering attention distribution and textual descriptions unexpectedly reducing accuracy and increasing uncertainty. Using the AUGRC score to measure the frequency of silent failures, our experiments reveal the harmful overconfidence of the PaliGemma 3B model compared to the LLaVA models. These findings provide valuable insights for the AI community by highlighting the critical role of context in VQA models, guiding future developments to optimize attention between visual and textual modalities, and assessing the risk of silent model failures.

In summary, our contributions are:

- A large-scale benchmark to evaluate the semantic interplay between image and textual context based on diverse evaluation criteria and several state-of-the-art VLMs, meant to facilitate silent failure detection.
- Ablation studies regarding rebalancing attention to specific modalities and the effect of model quantization.
- The SI-VQA Dataset, a well-curated dataset that enables the exploration of various combinations of image and text modalities. The benchmark and dataset are hosted at: <https://gitlab.com/dekfsx1/si-vlm-benchmark>.
- The ISI Tool, a tool designed to enable users to investigate how VLMs respond to semantic changes and interventions across image and text modalities, with a focus on identifying potential model failures. The tool is hosted at: <https://gitlab.com/dekfsx1/isi-vlm>.

2 RELATED WORK

VQA Datasets Since the introduction of the Visual Question Answering (VQA) task in Antol et al. (2015), numerous datasets have been created to support research in this area, such as GQA (Hudson & Manning, 2019), Visual7W (Zhu et al., 2016), Visual Genome (Krishna et al., 2017), CLEVR (Johnson et al., 2017), and VQA 2.0 (Goyal et al., 2017). However, these datasets often lack reasoning, are restricted to textual-only modalities, or suffer from small-scale and limited domain diversity. ScienceQA (Lu et al., 2022) emerged as the first large-scale VQA dataset to incorporate textual context alongside textual explanations for model reasoning. It features multiple-choice questions across various scientific domains, with each question annotated with lectures and explanations. More diverse multimodal datasets, such as SEED (Li et al., 2023), MMBench (Liu et al., 2023b), and MMMU (Yue et al., 2024a), have since been introduced, combining text and image data. However, as shown by Chen et al. (2024), these datasets do not ensure that all evaluation samples require visual content for correct answers. To address this limitation, we propose a new curated VQA dataset designed such that answers can only be derived from the image, with complementary or contradictory context provided as additional information to influence the model’s answer and reasoning.

Modalities in Vision-Language Tasks Several methods have been proposed to investigate the extent to which VLMs leverage both visual and textual information. *Annotation and foiling approaches* introduce text annotation in images and mistakes in image descriptions and test whether the VLM predictions change. Shekhar et al. (2019); Parcalabescu et al. (2022) test VLMs’ sensitivity to discrepancies between images and captions and found that models often overlook such inconsistencies. (Gat et al., 2021) exchange images and captions with other instances and note a consistent decrease in accuracy, with textual input proving to be more influential than visual content. Building on these findings,

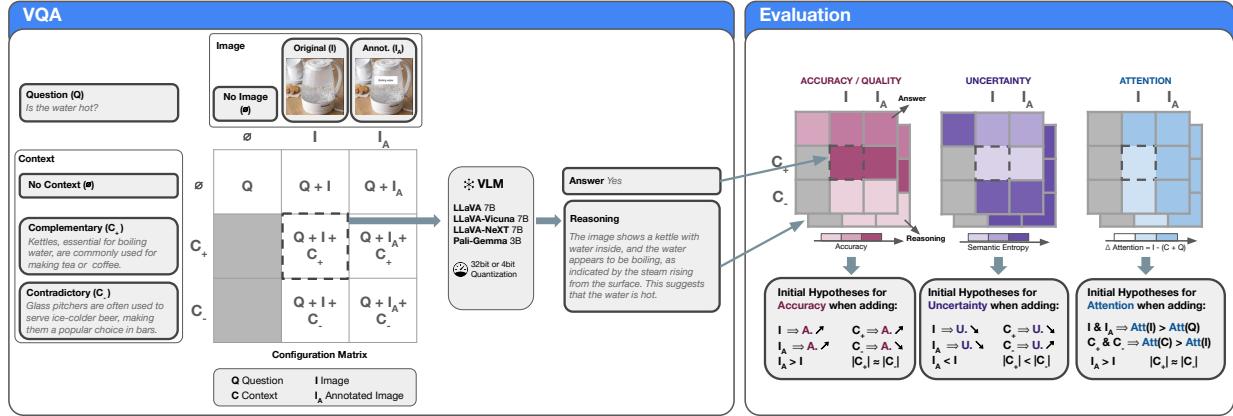


Figure 1: The SI-VQA framework examines the influence of various modality configurations on **answer accuracy and reasoning quality**, model **uncertainty**, and **attention attribution**. Seven different configurations are tested, combining inputs such as the question (Q), image (I), annotated image (I_A), and either complementary (C_+) or contradictory context (C_-): (Q) , $(Q+I)$, $(Q+I+C_+)$, $(Q+I+C_-)$, $(Q+I_A)$, $(Q+I_A+C_+)$, and $(Q+I_A+C_-)$. For each configuration, the VLM is assessed first on its answer and then on its reasoning. Furthermore, we establish prior assumptions regarding how each modality is expected to impact the model’s behavior.

we incorporate textual annotations into images in our study to explore how textual data, shown in previous work to be more significant than pixels, can support model reasoning. *Ablation methods* investigate model behavior when parts of the input are removed or masked (Bugliarello et al., 2021; Cafagna et al., 2021). Frank et al. (2021) occludes parts of the image or masks the text, and finds that the visual modality matters more than the text. In line with this, we introduce semantic perturbations, creating scenarios where the text complements the image in various ways. *Attention-based methods* correlate high attention scores with high feature importance, though this connection remains debated. (Serrano & Smith, 2019; Jain & Wallace, 2019) questioned the validity of attention as an indicator of importance, while (Wiegrefe & Pinter, 2019) challenged their arguments, asserting that attention can serve as an explanation, albeit not the definitive one. Following mechanistic interpretability, our paper also analyzes how attention scores are attributed across modalities in different text-image configurations. We view attention attribution as one key factor in understanding the intrinsic roles of each modality in VLM performance. Early work on the impact of modalities in VQA has produced mixed results. While some studies argue that VLMs rely more on text (Gat et al., 2021), others emphasize the importance of the visual modality (Frank et al., 2021). More recently, Chen et al. (2024) showed that visual data might be unnecessary to answer several questions, suggesting among others unintentional data leakage through LLM and VLM training. Our paper builds on this debate by proposing a rigorous framework for conducting semantic interventions on both modalities to assess their respective contributions to VQA tasks.

Multimodal XAI Multimodal Explainable AI (MXAI) is a branch of XAI that includes a range of XAI techniques tailored to address the unique challenges posed by multimodal data inputs, tasks, and architectures. Recent reviews study the new challenges and differences of those methods as compared to traditional XAI approaches (Joshi et al., 2021; Rodis et al., 2023). In the multimodal context, MXAI leverages the complementary explanatory strengths of different modalities offering richer insights. In specific scenarios, language can provide a deeper understanding and clarification of concepts, while in others, the visual modality may be more informative (Park et al., 2018). MXAI is used as a tool to interpret VQA tasks that might require higher-order reasoning and a deep understanding of semantic context (Antol et al., 2015).

3 SI-VQA DATASET AND ISI TOOL

This section introduces the dataset and tool developed to examine the role of modalities in VQA & Reasoning. They are designed to explore semantic interventions on input modalities for thorough post-hoc interpretability of VLMs.

3.1 SI-VQA DATASET

The SI-VQA dataset is a closed-question VQA dataset consisting of 100 samples. Each sample consists of an image, question, and ground truth answer (Yes/No) pair, as well as one text annotated image, one contradictory context, and one complementary context. The image is the necessary and sufficient element to get the question right as it contains the critical information. The context is always in relation to the ground truth answer, aiming to either confuse or help the model without stating an explicit answer to the question, as each question can only be answered through the image. Thus, the model has a very limited ability to leverage prior knowledge to answer the question and must utilize the image input. The annotations on the image give text-written hints to the ground truth answer. We only study the impact of “positive” annotations, i.e., text that informs the model about key elements in the image to correctly answer the question. See Figure 1 for an exemplary sample. The images are open-source and from the MMMU Benchmark (Yue et al., 2024b). The selected images of SI-VQA Dataset span a variety of fields, including geography, history, art and design, sport, and biology, along with everyday objects and landscapes. They encompass diverse formats, such as natural photographs, cartoons, sketches, and paintings. We carefully crafted all other modalities for the dataset, focusing on quality not quantity.

We design seven scenarios for VLM interpretability analysis, creating seven modality configurations: question (Q), question + image (Q+I), question + image + complementary context (Q+I+C₊), question + image + contradictory context (Q+I+C₋), question + annotated image (Q+I_A), question + annotated image + complementary context (Q+I_A+C₊), question + annotated image + contradictory context (Q+I_A+C₋). They can be inferred using the 3x3 matrix in Figure 1. For the baseline configuration Q (question-only), the image corresponds to a black image. In this case, the model’s answer accuracy is at random (See Appendix subsection H.1) proving the necessity of visual content in SI-VQA Dataset (Chen et al., 2024). We also conducted the baseline experiment using noise-only images and with the images entirely removed, observing similar results in both cases (See Appendix C).

3.2 ISI TOOL

To provide an intuitive and agile way to explore modality interplay in the context of multimodal interpretability, we developed the ISI Tool. We used this interactive tool to design the SI-VQA Dataset. It is designed to enable researchers and VLM users to investigate how VLMs respond to semantic changes and interventions across image and text modalities, with a focus on identifying potential model failures in the context of VQA. Specifically, it allows the perturbation of images, the addition of personalized shapes and annotations, and the arbitrary adaptation of text inputs. Users can upload their own images and questions or choose from 100 preloaded samples with semantic intervention presets from our ISI-VQA Dataset. More details about the tool and its features are available in Appendix E.

4 EXPERIMENT METHODOLOGY

Our benchmark pipeline illustrated in Figure 1 investigates the contribution of each modality toward the performance of several VLMs. Performance is measured in terms of output quality, model uncertainty, and attention attribution toward the input elements of the SI-VQA Dataset.

4.1 VISION-LANGUAGE MODELS

The VLMs selected for this study are state-of-the-art models for Visual Question Answering (VQA) tasks. We excluded models where extracting attention coefficients for each modality is not feasible, such as Flamingo (Alayrac et al., 2022), which employs gated cross-attention between text and image. The final architectures chosen for evaluation are LLaVA 1.5, LLaVA-Vicuna, LLaVA-NeXT, and PaliGemma (Beyer et al., 2024), with weights provided from HuggingFace (Wolf et al., 2020). LLaVA-Vicuna is a version of LLaVA 1.5 leveraging the Vicuna LLM, a conversation-fine-tuned version of LLaMA. LLaVA-Vicuna and LLaVA-NeXT both utilize dynamic high resolution for the image input (Liu et al., 2024a), increasing visual reasoning and optical character recognition (OCR) capabilities. Although PaliGemma is intentionally designed for pre-training followed by fine-tuning, we employ it in this study within a zero-shot setting. We do not present any reasoning results for the PaliGemma model as it mainly generates a default response “Sorry, as a base VLM I am not trained to answer this question.” when it is asked to explain its answer due to its strong safety fine-tuning (Beyer et al., 2024). Each LLaVA architecture comprises 7B parameters, while PaliGemma consists of 3B parameters, and all can be quantized to reduce VRAM usage and improve computational efficiency. Our ablation study in Appendix F shows, however, that while 32bit models have lower uncertainty in VQ answering and reasoning, answer accuracy is not substantially worse for even 4bit quantized models.

4.2 ANSWER & REASONING EVALUATION

We assess the model’s performance on the VQA task using the SI-VQA Dataset by evaluating both its answer and reasoning. To measure the VQ answering quality, we use the accuracy metric by comparing the model’s binary Yes/No response to the closed question with the ground truth provided in the dataset. Evaluating reasoning is more complex, as no ground truth exists for the explanations. Without a reference rationale, we assess reasoning based on the quality of argumentation and the truthfulness of statements. To this end, we useo as an external evaluator (Zheng et al., 2024; Thakur et al., 2024). The model is prompted once for each sample to rate the reasoning from 0 to 10, considering an evaluation prompt as well as the question, image, answer, and reasoning. While the quality scores seem very reasonable to us, we observe a bias toward the score number "8" (see subsection H.1), a behavior also observed in other studies using LLMs as a judge (Thakur et al., 2024). See Appendix B for the evaluation prompts and hyperparameters.

4.3 MODEL UNCERTAINTY

$$SE(x) = - \sum_c p(c|x) \log p(c|x) = - \sum_c \left(\left(\sum_{s \in c} p(s|x) \right) \log \left[\sum_{s \in c} p(s|x) \right] \right) \quad (1)$$

For quantifying model uncertainty, we employ semantic entropy $SE()$ (Farquhar et al., 2024), which calculates entropy based on the sum of token likelihoods $p()$ between the sets c of semantically similar clustered sentences s (see Equation 1). For semantic clustering, we use the DeBERTa (He et al., 2021) entailment model. During uncertainty computation, the number of sampled outputs and the sampling temperature T are set to 10 and 0.9 respectively. High $SE()$ means high uncertainty and low confidence in the outputs.

For VLMs, uncertainty quantification is especially important for detecting model failures, including hallucinations and silent failures. Silent failures are instances where the model generates incorrect information with high confidence, making these errors increasingly difficult to detect (Bender et al., 2021; Jaeger et al., 2023). The Area Under the Generalized Risk Coverage curve (AUGRC) metric (Traub et al., 2024) evaluates the extent to which a model makes incorrect predictions with high confidence, where high confidence is characterized by low semantic entropy, specifically below a defined threshold, τ :

$$\text{AUGRC} = \int_0^1 P(Y_f = 1, SE(x) \leq \tau) dP(SE(x) \leq \tau) \quad (2)$$

With Y_f as the binary failure indicator, i.e., $Y_f = 1$ indicates a wrong prediction by the model. We refer to Traub et al. (2024) for a detailed explanation. The AUGRC therefore directly measures the harmful overconfidence of a model, i.e., when the model is confident and wrong. A high AUGRC indicates that the model has a tendency for silent failures.

4.4 ATTENTION ATTRIBUTION

Attributing attention values to the different modalities serves as one indicator of each modality’s contribution to the final answer and reasoning (Wiegreffe & Pinter, 2019). We adopt a mechanistic approach, treating attention as a VLM interpretability measure to evaluate the roles of the question, image, and context. By aggregating attention across layers and heads, we derive relevance scores for each token’s contribution to the answer or reasoning (Parcalabescu & Frank, 2023; 2024). To determine which input modality is most relevant to the prediction, we sum the relevance scores of their respective input tokens. Finally, to standardize the scores for comparability, we compute the relative relevance score for each sample. A detailed explanation of attention aggregation and implementation is provided in Appendix A.

5 EXPERIMENT RESULTS

5.1 INITIAL HYPOTHESES

Based on the seven modality configurations and the results of previous related work, we define three prior hypotheses regarding the anticipated outcomes when including a new type of input. As discussed in subsection 4.1, only the VQ answering results are shown for PaliGemma.

Hypothesis 1 (Including Image): In the SI-VQA dataset, images are essential for VQA tasks, as they serve as the only modality providing the necessary information to answer the questions. Therefore, we hypothesize a significant positive impact on accuracy and a reduction in model uncertainty when the image is incorporated alongside the question.

Furthermore, the natural image is expected to receive greater attention compared to the baseline black or random pixel images in the question-only configuration. The impact on reasoning quality, however, remains in our opinion uncertain.

Hypothesis 2 (Including Context): We hypothesize that adding complementary context will improve model accuracy, confidence, and reasoning quality. This additional information is expected to facilitate more detailed and precise rationalization. Conversely, contradictory context is anticipated to decrease model accuracy, confidence, and reasoning quality, reflecting the model’s doubt. Based on previous research, we also expect the model to exhibit a higher attention toward text compared to the image.

Hypothesis 3 (Including Image Annotations): Compared to configurations including the natural image, those utilizing the annotated image are expected to enhance the model’s ability to answer and reason with greater accuracy and confidence. Additionally, we anticipate increased attention toward the annotated image relative to the natural image.

In the following sections, we will first present the benchmark results before comparing them to the prior hypotheses we have just put forward.

5.2 ANSWER & REASONING EVALUATION

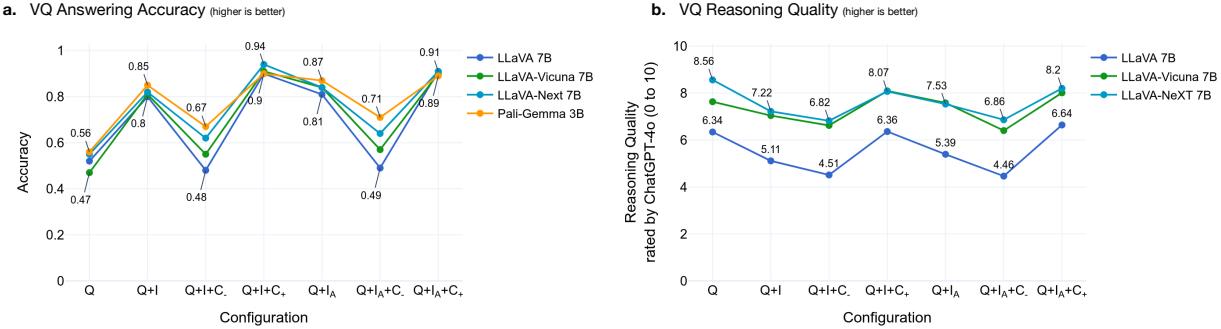


Figure 2: Quality of VLM answers and reasoning in the seven modality configurations of question (Q), image (I), annotated image (I_A), complementary (C_+) and contradictory context (C_-). Answer accuracy is measured using the ground truth labels of our SI-VQA Dataset and reasoning quality is evaluated using the external scoring of GPT-4o as a judge. A significant drop of accuracy in the answer and reasoning is observed for all models when adding contradictory context, i.e., $Q+I+C_-$ and $Q+I_A+C_-$. Results for PaliGemma 3B are only displayed for answering (see subsection 4.1).

Figure 2 presents the results of the VQA accuracy (a.) and the quality of the VLM VQ reasoning (b.), as judged by GPT-4o. Similar patterns emerge across all models. The answer accuracy is low in the question-only baseline, where the model lacks sufficient information to provide correct answers. However, when given just a question and a black image, the models consistently offer strong reasoning quality, consistently justifying their response by acknowledging the absence of image information. Incorporating complementary context into the I+Q configuration enhances both answer accuracy and reasoning quality by providing additional details necessary for a correct response and a well-supported rationale. In contrast, the introduction of contradictory context significantly degrades response accuracy. The decline in accuracy is the smallest for PaliGemma, whereas for LLaVA, it drops to a level comparable to the question-only configuration. Additionally, reasoning quality declines as the models are misled by the conflicting information. When exchanging the natural image with an annotated image, we observe no change in accuracy or reasoning quality, even for architectures optimized for OCR.

When comparing the VLM architectures, a notable discrepancy emerges in their handling of contradictory information, with models responding differently to contradictions (configuration $Q+I+C_-$). Surprisingly, PaliGemma demonstrates the most robustness in managing contradictions and achieving the highest accuracy scores in five out of seven configurations, despite having less than half the parameters of the LLaVA models and not being explicitly fine-tuned for VQA tasks. LLaVA-NeXT ranks second in accuracy but does not fully leverage its enhanced OCR capabilities when the annotated image is included. In terms of reasoning abilities, the conversational fine-tuned VLMs produce substantially higher-quality reasoning compared to the standard LLaVA 1.5 model.

Answer Accuracy & Reasoning Quality

While answer accuracy is low in the question-only baseline, models still provide strong reasoning quality by acknowledging missing image data. Complementary context improves accuracy and reasoning quality, but contradictory context significantly degrades performance. We observe the strongest decline for LLaVA and smallest in PaliGemma. Replacing the natural image with an annotated one shows no effect on accuracy or reasoning quality.

5.3 MODEL UNCERTAINTY

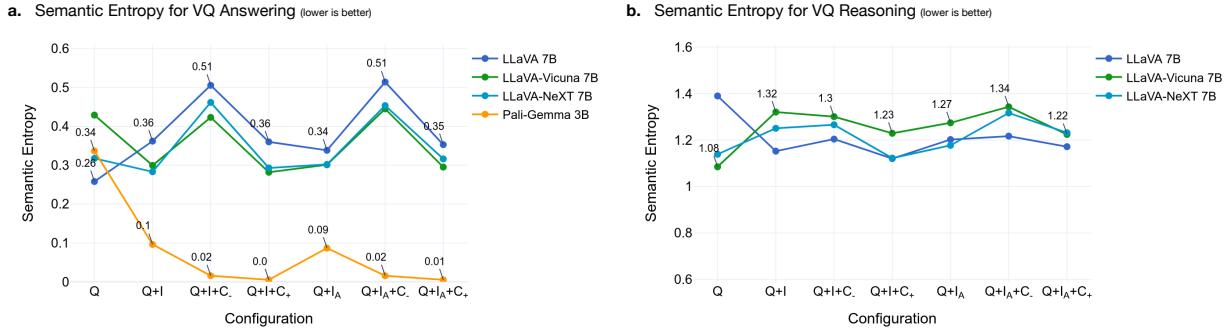


Figure 3: VLM uncertainty when generating answers and reasonings in the seven modality configurations of question (Q), image (I), annotated image (I_A), complementary (C_+) and contradictory context (C_-). Uncertainty is measured using the semantic entropy—the lower the entropy, the more confident the model. PaliGemma 3B shows extreme confidence overall in its answers. However, no reasoning results for PaliGemma are provided (see subsection 4.1). C_- negatively impacts the certainty of LLaVA models when generating answers.

Figure 3 displays the model uncertainty measured through semantic entropy for both the answer (a.) and the reasoning (b.). For all models, the absence of image-based information (configuration Q) results in similar levels of uncertainty in both VQ answering and reasoning. In addition, we observe for all models that image text annotations have almost no impact on the model uncertainty compared to the configurations including the natural images.

For PaliGemma, adding image and context information significantly reduces uncertainty in VQ answering, making the model much more confident in its predictions. It seems that providing additional context, regardless of its content, leads the model to be more self-assured. This intriguing pattern shows large overconfidence in PaliGemma, which does not always have to be beneficial as it can, e.g., lead to silent failures, where the model is extremely confident in its wrong predictions (Bender et al., 2021; Jaeger et al., 2023).

For all LLaVA models, we observe overall an inverse relationship between answer uncertainty and reasoning uncertainty, with LLaVA 1.5 exhibiting the highest uncertainty in VQ answering but the lowest in VQ reasoning.

When the image is added, LLaVA-Vicuna and LLaVA-NeXT show reduced uncertainty in VQ answering but increased uncertainty in VQ reasoning, as the models, in the question-only configuration, only acknowledge the absence of the image and therefore reason with high confidence. Complementary context slightly decreases model uncertainty, indicating a marginal increase in confidence for both VQ answering and reasoning. This effect is minor though, as shown in Figure 3 b. where all LLaVA models exhibit nearly identical semantic entropies for $I+Q$ and $Q+I+C_+$, as well as for $Q+I_A$ and $Q+I_A+C_+$. Contradictory contextual information, on the other hand, significantly increases uncertainty in the model answers. Its effect on reasoning is also particularly pronounced in LLaVA 1.5 but remains relatively minor for LLaVA-Vicuna and LLaVA-NeXT. Thus, the LLaVA models appear to be slightly influenced by reinforcing information sources but are more easily unsettled by contradictory ones.

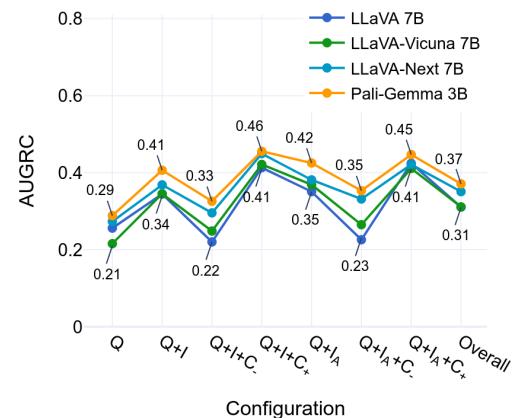


Figure 4: AUGRC evaluating the ability to detect silent failures through semantic entropy for each model and configuration (lower is better).

Model Failure Detection through Semantic Entropy Interpreting VLMs’ behavior through the lense of model uncertainty is crucial for identifying and understanding failures, including hallucinations and silent failures. Specifically, for PaliGemma, silent failures cannot be dismissed due to the model’s extreme overconfidence, despite its prediction accuracy being comparable to that of LLaVA models. To quantify this harmful overconfidence—when the model confidently predicts incorrect answers with very low uncertainty—we employ the AUGRC metric, as detailed in Equation 2. Figure 4 displays the AUGRC values across all model architectures (where lower is better). Our results confirm that PaliGemma performs the worst, validating our hypothesis regarding its harmful overconfidence. Additionally, we observe that in cases of high uncertainty, such as with the $Q+I+C_-$ and $Q+I_A+C_-$ configurations, AUGRC is low, indicating fewer silent failures. In these scenarios, the contradictory context reduces the likelihood of confident incorrect answers, meaning the models become more uncertain about their mistakes, thereby making them more trustworthy.

Uncertainty

PaliGemma demonstrates high overconfidence, leading to silent failures, as evidenced by the AUGRC metric. LLaVA models display an inverse relationship between VQ answering and reasoning uncertainty, probably due to more detailed reasoning in more advanced models. Contradictory context significantly increases the VQ answering uncertainty for LLaVA models but has only a minor effect on VQ reasoning.

5.4 ATTENTION ATTRIBUTION

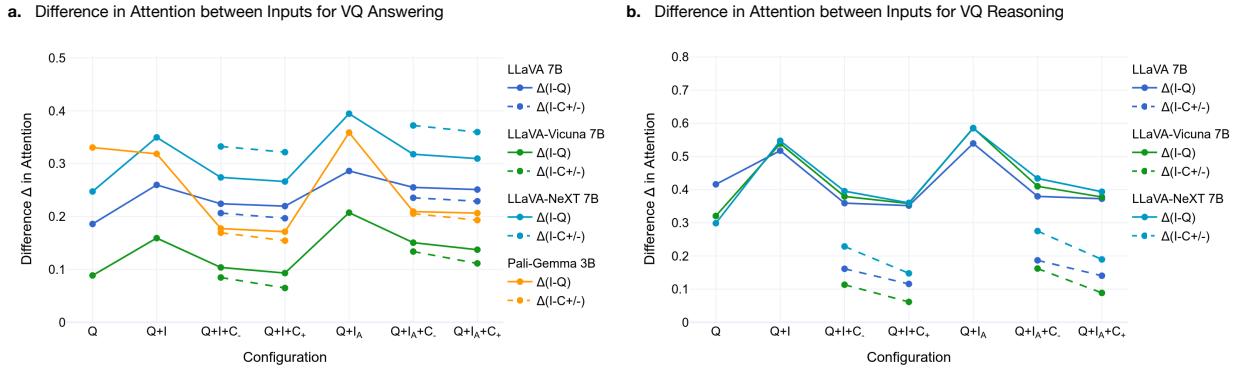


Figure 5: Difference in attention attribution between the image and the question (solid line) and between the image and the context when present (dashed line). The computation of modality attention attribution is described in Appendix A. The image (I) always gets the highest attention attribution compared to text modalities (Q, C₋, C₊). No reasoning results for PaliGemma are provided (see subsection 4.1).

This section examines how attention is distributed across the three inputs—question, image, and context—across the seven configurations. Figure 5 shows the average difference between attention to the image and the question, as well as the attention to the image and the context. Since all differences are positive, the image consistently receives the highest average attention in both VQ answering (Figure 5a.) and reasoning (Figure 5b.). The attention distribution across different inputs is similar among the models, with LLaVA-Vicuna showing the highest attention to textual inputs and LLaVA-NeXT focusing more on the image. Both answering and reasoning exhibit higher attention to the natural image compared to the black baseline image in the question-only configuration. Further, attention to the image decreases when context is added and the annotated image receives more attention than the natural image. In VQ answering, attention to the question and context is nearly equal, whereas, in VQ reasoning, the model shows significantly higher attention to the context, almost equal to the attention given to the image. Detailed figures are provided in Appendix subsection H.3. Overall, no strong correlation is observed between attention attribution and accuracy (see Appendix subsection H.4).

Attention Attribution

The image modality receives the highest attention compared to the question and context, with its relevance further increasing when annotated. In VQ answering, attention to the question and context is nearly equal, whereas in VQ reasoning, the model allocates significantly more attention to the context. LLaVA-Vicuna pays the highest attention to textual inputs and LLaVA-NeXT to the image.

5.5 REBALANCING MODALITY IMPORTANCE

Given the observed high attention allocated to the image modality, it raises the question of how the results might change if we intervene to direct more attention toward the text modalities. Specifically, we aim to investigate how the model’s performance changes when either the information in the image is described with text or the attention of the model is guided toward the text via prompt engineering.

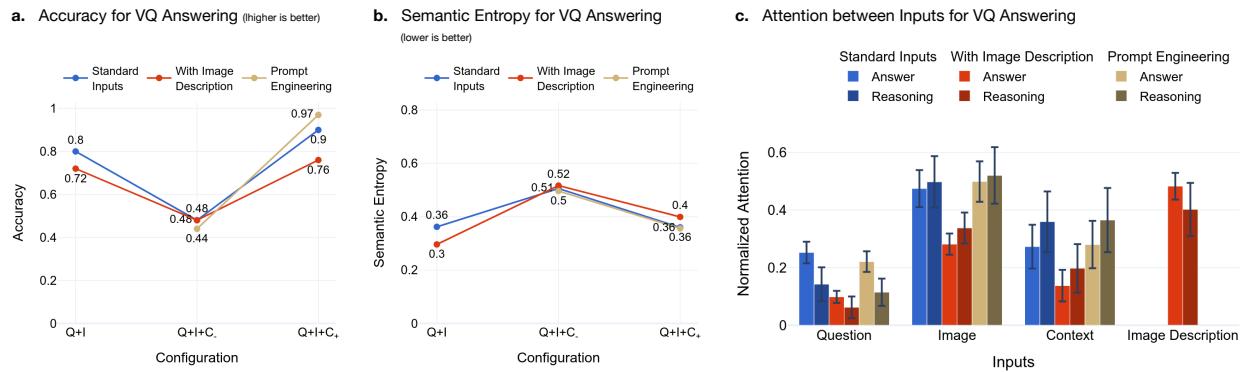


Figure 6: Two strategies to adjust modality importance: incorporating the image’s textual description into the input and modifying the prompt to shift more attention toward the context. The impact of these changes is evaluated based on a. answer accuracy, b. model uncertainty, and c. attention attribution. The hereby experiments were conducted with the LLaVA 7B model, see Appendix I for the PaliGemma results.

Textual Description of the Image *Does adding a text description of the image greatly improve model performance and confidence?* We initially hypothesized that augmenting the model’s input with a textual description of the image would enhance its accuracy and reduce uncertainty, based on the premise that key features necessary for answering the question—already present in the image—would be more accessible to the LLM decoder in text form. However, as illustrated for LLaVA in Figure 6, the results reveal a surprising decrease in answer accuracy and an increase in model uncertainty for configuration $I+Q+C_+$. We observe similar results for PaliGemma (see Appendix I), with minor increases in uncertainty. We argue that these findings indicate the models are already proficient at extracting essential information from the image alone and that the addition of textual information introduces confusion due to redundancy or potential inconsistencies in the image description. Moreover, the high attention allocated to the image description in Figure 6c. underscores the model’s sensitivity to textual inputs, which may inadvertently dominate visual cues.

Prompt Engineering *Can prompt engineering help VLMs re-balance their attention toward the context?* We modified the initial prompt to direct the model’s attention more toward the textual context, which typically receives less emphasis in the standard setting (see Appendix B for implementation details). Given that the complementary context provides information intended to guide the model toward the correct answer, while the contradictory context aims to mislead it, we expected an increase in accuracy for the $Q+I+C_+$ configuration and a decrease for the $Q+I+C_-$ configuration. As shown in Figure 6 a, we observed a decrease in accuracy for $Q+I+C_-$, whereas prompt engineering to emphasize the complementary context resulted in improved answer accuracy. Unexpectedly, these changes are not reflected in the attention attribution: Figure 6c. indicates that prompt engineering does not alter the attention distribution across modalities, as the context does not receive increased attention. This lack of correlation between attention and model performance highlights the necessity for cautious interpretation of attention mechanisms in model predictions (see section 2). Unlike the LLaVA results, PaliGemma exhibited a significant increase in uncertainty (see Appendix I), highlighting the large influence of the image in the standard inputs.

6 DISCUSSION & CONCLUSION

Evaluation of Hypotheses Our findings reveal notable insights into the role of each modality in VQA and reasoning tasks. Specifically, we compare all results with our initial hypotheses from subsection 5.1.

Hypothesis 1 (Including Image): As expected, introducing the image results in a significant increase in answer accuracy across all VLMs. However, it unexpectedly leads to a decrease in reasoning quality, as in the question-only setting, the models simply acknowledge the absence of the image. We observe a similar pattern in model uncertainty: it decreases for VQ answering but increases in reasoning. As hypothesized, the natural image indeed receives more attention compared to the black baseline image.

Hypothesis 2 (Including Context): Consistent with our expectations, the inclusion of complementary context enhances both accuracy and reasoning quality, while contradictory context has a strongly negative effect. However, in VQ answering, the complementary context does not reduce model uncertainty, whereas the contradictory context significantly increases it. In VQA reasoning, contradictory context does not affect uncertainty, and complementary context only slightly decreases it. Generally, the impact of adding context is much stronger in the VQ answering than in the reasoning task. Interestingly, contradictory context can sometimes be beneficial, as it helps to minimize the occurrence of silent failures. Additionally, the models continue to show higher attention toward the image than the context, not supporting our prior hypothesis.

Hypothesis 3 (Including Image Annotations): Surprisingly, the image text annotations play a minimal role in enhancing model performance. Although the models exhibit increased attention toward annotated images, the positive impact of these annotations on performance metrics and uncertainty reduction is nearly negligible.

We also investigate methods to guide the model to favor one modality over another to observe the effect on VLM performance. While adding redundant textual information can overwhelm the model and decrease accuracy, prompt engineering can improve predictions without, however, strong changes in attention distribution.

Limitations The SI-VQA dataset contains exactly 100 instances. Although relatively small, it has been meticulously handcrafted, with each instance carefully designed to meet key criteria: the question can only be answered using the image; the context provides additional global information either reinforcing the image’s consistency or misleading the model; and the annotations are simplified concepts written directly on the image to aid in accurately interpreting the scene. In this work, we also limit our study to seven different modality configurations with specific interventions. For those interested in more specific or advanced interventions, we refer to our ISI Tool, which allows testing of almost all possible forms of interventions. In future research, it would be interesting to replicate this study in the reverse scenario—where text is the primary content required for answering, and the image serves as contextual information—to compare whether similar effects of primary and secondary modalities are observed. We employ semantic entropy as our unique uncertainty measure, considering other measures for free-form text generation, such as token entropy (Kadavath et al., 2022; Lindley, 1956) or self-expressed uncertainty (Lin et al., 2022; Liao & Wortman Vaughan, 2024), to be significantly less suitable. They either only consider local, token-level uncertainty or rely on the model’s potentially biased self-assessment, neither adequately reflecting the overall semantic uncertainty. We validated the expected behavior of semantic entropy across the different configurations using the AUGRC metric in Appendix G.

Conclusion This study is the first to systematically examine the role of contextual information in VQA, evaluating the results based on diverse metrics and distinguishing between answering and reasoning tasks. Leveraging the well-curated SI-VQA dataset and the ISI tool—our interactive, ready-to-use interface—our work aims to provide a deeper understanding of VLM behavior and the influence of each modality. Moreover, we show that our results can also be used to understand and detect model failure in free-form text generation, and setting the stage for future analyses of modality integration across various VLM tasks and the development of VL datasets tailored for this objective.

REFERENCES

- Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.
- Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pp. 2425–2433, 2015.
- Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big?. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pp. 610–623, 2021.
- Lucas Beyer, Andreas Steiner, André Susano Pinto, Alexander Kolesnikov, Xiao Wang, Daniel Salz, Maxim Neumann, Ibrahim Alabdulmohsin, Michael Tschannen, Emanuele Bugliarello, Thomas Unterthiner, Daniel Keysers, Skanda Koppula, Fangyu Liu, Adam Grycner, Alexey Gritsenko, Neil Houlsby, Manoj Kumar, Keran Rong, Julian Eisenschlos, Rishabh Kabra, Matthias Bauer, Matko Bošnjak, Xi Chen, Matthias Minderer, Paul Voigtlaender, Ioana Bica, Ivana Balazevic, Joan Puigcerver, Pinelopi Papalampidi, Olivier Henaff, Xi Xiong, Radu Soricut, Jeremiah Harmsen, and Xiaohua Zhai. PaliGemma: A versatile 3B VLM for transfer, July 2024.
- Emanuele Bugliarello, Ryan Cotterell, Naoaki Okazaki, and Desmond Elliott. Multimodal pretraining unmasked: A meta-analysis and a unified framework of vision-and-language BERTs. *Transactions of the Association for Computational Linguistics*, 9:978–994, 2021. doi: 10.1162/tacl_a_00408. URL <https://aclanthology.org/2021.tacl-1.58>.
- Michele Cafagna, Kees van Deemter, and Albert Gatt. What vision-language models see when they see scenes. *arXiv preprint arXiv:2109.07301*, 2021.
- Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. Are we on the right way for evaluating large vision-language models? *arXiv preprint arXiv:2403.20330*, 2024.
- Michael Dorkenwald, Nimrod Barazani, Cees G. M. Snoek, and Yuki M. Asano. Pin: Positional insert unlocks object localisation abilities in vlms, 2024. URL <https://arxiv.org/abs/2402.08657>.
- Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630, June 2024. ISSN 0028-0836, 1476-4687. doi: 10.1038/s41586-024-07421-0.
- Stella Frank, Emanuele Bugliarello, and Desmond Elliott. Vision-and-language or vision-for-language? on cross-modal influence in multimodal transformers. *arXiv preprint arXiv:2109.04448*, 2021.
- Itai Gat, Idan Schwartz, and Alex Schwing. Perceptual score: What data modalities does your model perceive? *Advances in Neural Information Processing Systems*, 34:21630–21643, 2021.
- Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6904–6913, 2017.
- Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*, 2021. URL <https://openreview.net/forum?id=XPZIaotutsD>.
- Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6700–6709, 2019.
- Paul F. Jaeger, Carsten T. Lüth, Lukas Klein, and Till J. Bungert. A call to reflect on evaluation practices for failure detection in image classification, 2023. URL <https://arxiv.org/abs/2211.15259>.

Sarthak Jain and Byron C. Wallace. Attention is not Explanation. In Jill Burstein, Christy Doran, and Thamar Solorio (eds.), *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)*, pp. 3543–3556, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1357. URL <https://aclanthology.org/N19-1357>.

Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2901–2910, 2017.

Gargi Joshi, Rahee Walambe, and Ketan Kotecha. A review on explainability in multimodal deep neural nets. *IEEE Access*, 9:59800–59821, 2021.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, Deep Ganguli, Danny Hernandez, Josh Jacobson, Jackson Kernion, Shauna Kravec, Liane Lovitt, Kamal Ndousse, Catherine Olsson, Sam Ringer, Dario Amodei, Tom Brown, Jack Clark, Nicholas Joseph, Ben Mann, Sam McCandlish, Chris Olah, and Jared Kaplan. Language models (mostly) know what they know, 2022. URL <https://arxiv.org/abs/2207.05221>.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123:32–73, 2017.

Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*, 2023.

Q. Vera Liao and Jennifer Wortman Vaughan. AI Transparency in the Age of LLMs: A Human-Centered Research Roadmap. *Harvard Data Science Review*, (Special Issue 5), may 31 2024. <https://hdsr.mitpress.mit.edu/pub/aelql9qy>.

Stephanie Lin, Jacob Hilton, and Owain Evans. Teaching models to express their uncertainty in words, 2022. URL <https://arxiv.org/abs/2205.14334>.

D. V. Lindley. On a Measure of the Information Provided by an Experiment. *The Annals of Mathematical Statistics*, 27(4):986 – 1005, 1956. doi: 10.1214/aoms/1177728069. URL <https://doi.org/10.1214/aoms/1177728069>.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023a.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved Baselines with Visual Instruction Tuning, May 2024a.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, January 2024b. URL <https://llava-vl.github.io/blog/2024-01-30-llava-next/>.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? *arXiv preprint arXiv:2307.06281*, 2023b.

Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *The 36th Conference on Neural Information Processing Systems (NeurIPS)*, 2022.

Omar Moured, Jiaming Zhang, M. Saquib Sarfraz, and Rainer Stiefelhagen. Altchart: Enhancing vlm-based chart summarization through multi-pretext tasks, 2024. URL <https://arxiv.org/abs/2405.13580>.

OpenAI. GPT-4 Technical Report, March 2024.

Letitia Parcalabescu and Anette Frank. Mm-shap: A performance-agnostic metric for measuring multimodal contributions in vision and language models & tasks. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.acl-long.223. URL <http://dx.doi.org/10.18653/v1/2023.acl-long.223>.

Letitia Parcalabescu and Anette Frank. On measuring faithfulness or self-consistency of natural language explanations. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 6048–6089, 2024.

Letitia Parcalabescu, Michele Cafagna, Lilitta Muradjan, Anette Frank, Iacer Calixto, and Albert Gatt. VALSE: A task-independent benchmark for vision and language models centered on linguistic phenomena. In Smaranda Mureșan, Preslav Nakov, and Aline Villavicencio (eds.), Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 8253–8280, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.567. URL <https://aclanthology.org/2022.acl-long.567>.

Dong Huk Park, Lisa Anne Hendricks, Zeynep Akata, Anna Rohrbach, Bernt Schiele, Trevor Darrell, and Marcus Rohrbach. Multimodal explanations: Justifying decisions and pointing to the evidence. In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 8779–8788, 2018.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. URL <https://arxiv.org/abs/2103.00020>.

Nikolaos Rodis, Christos Sardianos, Georgios Th Papadopoulos, Panagiotis Radoglou-Grammatikis, Panagiotis Sarigiannidis, and Iraklis Varlamis. Multimodal explainable artificial intelligence: A comprehensive review of methodological advances and future research directions. arXiv preprint arXiv:2306.05731, 2023.

Sofia Serrano and Noah A. Smith. Is attention interpretable? In Anna Korhonen, David Traum, and Lluís Márquez (eds.), Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 2931–2951, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1282. URL <https://aclanthology.org/P19-1282>.

Ravi Shekhar, Ece Takmaz, Raquel Fernández, and Raffaella Bernardi. Evaluating the representational hub of language and vision models. In Simon Dobnik, Stergios Chatzikyriakidis, and Vera Demberg (eds.), Proceedings of the 13th International Conference on Computational Semantics - Long Papers, pp. 211–222, Gothenburg, Sweden, May 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-0418. URL <https://aclanthology.org/W19-0418>.

Aman Singh Thakur, Kartik Choudhary, Venkat Srinik Ramayapally, Sankaran Vaidyanathan, and Dieuwke Hupkes. Judging the judges: Evaluating alignment and vulnerabilities in llms-as-judges, 2024. URL <https://arxiv.org/abs/2406.12624>.

Jeremias Traub, Till J. Bungert, Carsten T. Lüth, Michael Baumgartner, Klaus H. Maier-Hein, Lena Maier-Hein, and Paul F Jaeger. Overcoming common flaws in the evaluation of selective classification systems, 2024. URL <https://arxiv.org/abs/2407.01032>.

Sarah Wiegreffe and Yuval Pinter. Attention is not explanation. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pp. 11–20, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1002. URL <https://aclanthology.org/D19-1002>.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierrick Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Huggingface’s transformers: State-of-the-art natural language processing, 2020. URL <https://arxiv.org/abs/1910.03771>.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhua Chen. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In Proceedings of CVPR, 2024a.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhua Chen. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In Proceedings of CVPR, 2024b.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena. In Proceedings of the 37th International Conference on Neural Information Processing Systems, NIPS '23, Red Hook, NY, USA, 2024. Curran Associates Inc.

Yuke Zhu, Oliver Groth, Michael Bernstein, and Li Fei-Fei. Visual7w: Grounded question answering in images. In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 4995–5004, 2016.

A MODALITY ATTENTION AGGREGATION

From the VLMs, we extract an attention matrix $A_t \in \mathbb{R}^{t \times t}$ for each output token t , where the matrix size grows with the number of predicted tokens. Each row represents a token as a query, and each column corresponds to a token as a key. Therefore, each row r contains the attention coefficients of each token i with respect to the token r : $A[r, i] = \alpha_{r,i} = q_r^\top k_i$. For each token t , we focus only on the attention of preceding tokens, represented by $v_t^\top = A[-1] \in \mathbb{R}^{1 \times t}$. v_t is normalized so that the attention coefficients of all preceding tokens to t sum to 1.

This process is repeated for all output tokens $t \in [1, T]$ where T is the output length, yielding all normalized attention vectors $v_t, t \in [1, T]$. These are averaged to produce the final attention vector for the VLM output— $V_A \in \mathbb{R}^{N_0}$ for answers and $V_R \in \mathbb{R}^{N_0+T_A+N_1}$ for reasoning. Here, N_0 is the size of the input prompt including image tokens, question, and eventually context, T_A is the answer length, and N_1 is the size of the second prompt asking for an explanation.

To calculate the attention given to different modalities, we sum the attention coefficients based on the token positions in the averaged attention vector V_A or V_R , resulting in normalized relevance scores: R_I for the image, R_Q for the question, and R_C for the context, which sum to 1, i.e., $R_I + R_Q + R_C = 1$. This is done by adding hooks into the LLaVA architecture to capture the start and end positions of the question, image, and context tokens. Due to dynamic high resolution, the number of image tokens can vary significantly even with minimal image perturbations.

Algorithm 1 Attention Attribution Computation

- 1: **Input:** Attention matrix $A_t \in \mathbb{R}^{t \times t}$ for each token t , Token positions for Image (I), Question (Q), and Context (C), Length of output T
 - 2: **Output:** Normalized relevance scores R_I, R_Q, R_C
 - 3: **for** $t \in [1, T]$ **do**
 - 4: Extract attention for preceding tokens: $v_t^\top = A[-1] \in \mathbb{R}^{1 \times t}$
 - 5: Normalize attention coefficients: $v_t \leftarrow \frac{v_t}{\sum_{i=1}^t v_t[i]}$
 - 6: **end for**
 - 7: Compute average attention vector: $V = \frac{1}{T} \sum_{t=1}^T v_t^T$
 - 8: Compute total attention for each modality:
 - 9: $R_I \leftarrow \sum_{i \in I} V[i]$
 - 10: $R_Q \leftarrow \sum_{i \in Q} V[i]$
 - 11: $R_C \leftarrow \sum_{i \in C} V[i]$
 - 12: Normalize relevance scores:
 - 13: $R_I \leftarrow \frac{R_I}{R_I + R_Q + R_C}$
 - 14: $R_Q \leftarrow \frac{R_Q}{R_I + R_Q + R_C}$
 - 15: $R_C \leftarrow \frac{R_C}{R_I + R_Q + R_C}$
 - 16: **Return:** R_I, R_Q, R_C
-

B HYPERPARAMETER AND EVALUATION PROMPTS

In reasoning quality evalution, GPT-4o is prompted to rate the reasoning from 0 to 10, using the prompt: “Rate the explanation’s quality from 0 to 10. Give 10 for detailed, well-argued, and correct explanations. Give 0 for a poorly reasoned, wrong, or single-word explanation based on the question and image. Don’t rate too harshly, use the full scale and output only the final score”. During uncertainty computation, the number of the sampled outputs and the sampling temperaturee T are set to 10 and 0.9 respectively. We use conditional probabilistic sampling. See Appendix A for the attention attribution implementation.

In the prompt engineering ablation study in subsection 5.5 we use the prompt “Answer the question only with Yes or No. Answer based on the context provided in the text.” for the answer, followed by “Explain your answer based on the context provided in the text.” for reasoning.

C SI-VQA DATASET BASELINE

In the baseline configuration of the SI-VQA Dataset, the image tokens carry no meaningful information. We experiment with several methods to remove image data: (1) replacing the image with black pixels, (2) adding random

Gaussian noise centered on (0,0,0), or (3) omitting the image tokens entirely. In (1) and (2), the image size is by default set to 256 × 256 pixels. Across our experiments, we observed similar results when using black pixels or noise. However, when no image input is provided, the model produces incoherent predictions and reasoning, often generating “\n”. Scenario (3) is out-of-distribution for any VLMs, leading to nonsensical outputs, as the model was trained exclusively in a multimodal context and has never encountered unimodal (text-only) input.

(1) Baseline configuration with black image

Question: Is there a yellow and black frog in this school book?

Ground truth answer: Yes



Answer: No

Reasoning: The image you provided is completely black, which does not allow for the identification of any objects or creatures, including a yellow and black frog.

(2) Baseline configuration with noisy image

Question: Is there a yellow and black frog in this school book?

Ground truth answer: Yes



Answer: No

Reasoning: The image is a close-up of a colorful, pixelated background. There is no frog visible in the image.

(3) Baseline configuration with no image token

Question: Is there a yellow and black frog in this school book?

Ground truth answer: Yes

Answer: 是 ◊] ◊] \n \n Is a \n\n

Reasoning: 10

D SI-VQA BENCHMARK EXAMPLES

Each instance of the SI-VQA Dataset was double-checked and validated by three individuals.

Boiling.jpg

Question: Is the water hot?

Ground truth answer: Yes

LLaVA 7B: No. The water is not hot.

LLaVA-Next 7B: No. The image you provided is too dark and lacks any discernible features to determine if the water is hot or not.

Question: Is the water hot?

LLaVA 7B: No. The water in the pitcher is cold.

LLaVA-Next 7B: No. The image shows a pitcher of water with ice cubes, which suggests that the water is cold.



Question: Is the water hot?

Complementary Context: Kettles, essential for boiling water, are commonly used for making tea, coffee, or instant meals, and come in various designs, including traditional stove top and modern electric models.

LLaVA 7B: Yes. The water in the kettle is hot.

LLaVA-Next 7B: No. The image shows a water kettle with water inside it, but it is not hot. The water appears to be at room temperature or slightly cooler, as indicated by the condensation on the outside of the kettle.



Question: Is the water hot?

Contradictory Context: Glass pitchers are often used to serve ice-cold beer, making them a popular choice in bars and restaurants for sharing drinks in a casual, communal setting.

LLaVA 7B: No. The water in the pitcher is not hot. It is ice-cold, which is a common feature of glass pitchers used for serving beer.

LLaVA-Next 7B: No. The image shows a glass pitcher filled with ice and water, which is not hot. The ice is used to keep the water cold, and the pitcher is designed to serve multiple people at once. This type of pitcher is commonly used for serving cold beverages, such as water or iced tea, in a social setting.



Question: Is the water hot?

LLaVA 7B: No. Boiling water is water that is heated to the point of boiling, which is the temperature at which water vaporizes. In this image, the water in the kettle is not hot, as it is described as "boiling water" and "cold water."

LLaVA-Next 7B: Yes. The image shows a water pitcher with the label "Boiling water" on it, which suggests that the water inside is being heated to boiling point. The presence of steam and the fact that the water is being heated in a pitcher designed for boiling water further supports the idea that the water is hot.



Question: Is the water hot?

Complementary Context: Kettles, essential for boiling water, are commonly used for making tea, coffee, or instant meals, and come in various designs, including traditional stove top and modern electric models.

LLaVA 7B: Yes. The water in the kettle is hot, as indicated by the steam rising from it.

LLaVA-Next 7B: No. The image shows a water kettle with a label that reads "Boiling water." However, the water in the kettle appears to be cold, as indicated by the condensation on the outside of the kettle. This suggests that the water has not yet reached its boiling point.



Question: Is the water hot?

Contradictory Context: Glass pitchers are often used to serve ice-cold beer, making them a popular choice in bars and restaurants for sharing drinks in a casual, communal setting.

LLaVA 7B: No. The water in the pitcher is not hot. It is ice-cold, which is a common feature of glass pitchers used for serving beer.

LLaVA-Next 7B: No. The image shows a water pitcher with the label "Boiling water" on it, but the water inside appears to be cold. The context of the image, which includes a wooden table and a few cookies, suggests that the water is being used for a hot beverage, such as tea or coffee. However, the water itself is not boiling at the moment the image was taken.



Painting_asia.jpg

Question: Is the painting of european descent?

Ground truth answer: No

LLaVA 7B: No. The painting is of African descent.

LLaVA-Next 7B: No. The image you provided is too dark and lacks detail to determine if it is a painting of European descent. The image is not clear enough to make an accurate assessment.



Question: Is the painting of european descent?

LLaVA 7B: No. The painting is of Asian descent.

LLaVA-Next 7B: No. The painting is not of European descent. It appears to be a traditional East Asian artwork, likely depicting a scene from a historical period in China or Japan, given the style of clothing and the architectural elements.



Question: Is the painting of european descent?

Complementary Context: Scholars and students engaged in study and teaching within an ancient Chinese educational setting. The intricate details of the architecture and the serene outdoor environment emphasize the importance of learning and intellectual pursuit in classical Chinese culture.

LLaVA 7B: Yes. The painting is of an ancient Chinese educational setting, featuring scholars and students engaged in study and teaching. The intricate details of the architecture and the serene outdoor environment emphasize the importance of learning and intellectual pursuit in classical Chinese culture.

LLaVA-Next 7B: No. The painting is not of European descent. It depicts a traditional Chinese scene, with scholars and students engaged in study and teaching within an ancient Chinese educational setting. The architecture and the serene outdoor environment are characteristic of Chinese art and culture, and the painting is not of European origin.



Question: Is the painting of european descent?

Contradictory Context: In an old German schoolhouse, students diligently engage in their lessons, surrounded by the timeless charm of wooden desks, slate chalkboards, and the warm glow of afternoon sunlight filtering through the tall, arched windows.

LLaVA 7B: Yes. Yes.

LLaVA-Next 7B: Yes. The painting depicts a scene that is characteristic of traditional European schools, with students seated at wooden desks, using slate chalkboards, and the presence of tall, arched windows suggests a European architectural style. The clothing of the students and the overall composition of the painting are consistent with European art and culture.



Question: Is the painting of european descent?

LLaVA 7B: No. The painting is of Chinese descent.

LLaVA-Next 7B: No. The painting is not of European descent. It depicts a traditional Chinese scene, with Chinese architecture, clothing, and people engaged in activities such as writing and reading. The style of the painting, including the clothing and the architectural elements, is indicative of a Chinese setting, which suggests that the painting is not of European descent.



Question: Is the painting of european descent?

Complementary Context: Scholars and students engaged in study and teaching within an ancient Chinese educational setting. The intricate details of the architecture and the serene outdoor environment emphasize the importance of learning and intellectual pursuit in classical Chinese culture.

LLaVA 7B: No. The painting is of Chinese descent, depicting a scene of scholars and students engaged in study and teaching within an ancient Chinese educational setting. The intricate details of the architecture and the serene outdoor environment emphasize the importance of learning and intellectual pursuit in classical Chinese culture.

LLaVA-Next 7B: No. The painting is not of European descent. It depicts a traditional Chinese scene, with Chinese architecture, clothing, and hairstyles, which are characteristic of East Asian art. The style of the painting, including the brushwork and the subject matter, are indicative of Chinese art and culture.



Question: Is the painting of european descent?

Contradictory Context: In an old German schoolhouse, students diligently engage in their lessons, surrounded by the timeless charm of wooden desks, slate chalkboards, and the warm glow of afternoon sunlight filtering through the tall, arched windows.

LLaVA 7B: No. The painting is of Chinese descent, depicting a scene from a Chinese school.

LLaVA-Next 7B: Yes. The painting depicts a traditional European schoolhouse, with students seated at wooden desks and chalkboards, and the architecture and style of the building suggest a European influence. The presence of tall, arched windows and the overall composition of the scene are characteristic of European art and architecture.

E ISI TOOL

The interactive tool can be used to analyze VLMs with the provided SI-VQA Dataset and follows a main pipeline that consists of three main steps: 1) Data & Model Selection 2) Interventions on Image, Context, and Question, and 3) Evaluation. Figure 7 gives an overview of this pipeline.

E.1 GENERAL INFORMATION

Users The application is catered toward researchers, developers, and other users with a basic understanding of VLMs, who are interested in interpreting model behavior through semantic interventions on VLMs. By enabling fast-paced iterations in a human-in-the-loop scenario, it allows the building of intuitions before scaling experiments in large-scale projects.

System Requirements ISI for VLMs is an interactive tool embedded in a locally hosted web application requiring a computer with sufficient VRAM for VLM inference. The minimum required VRAM for a 4bit-quantized LLAVA 7B model is around 8GB while LLava-Vicuna and LLava-Next require 12GB. The computation of the semantic entropy with the DeBERTa model requires an additional 7GB. The exact amount of VRAM depends on the amount of input tokens.

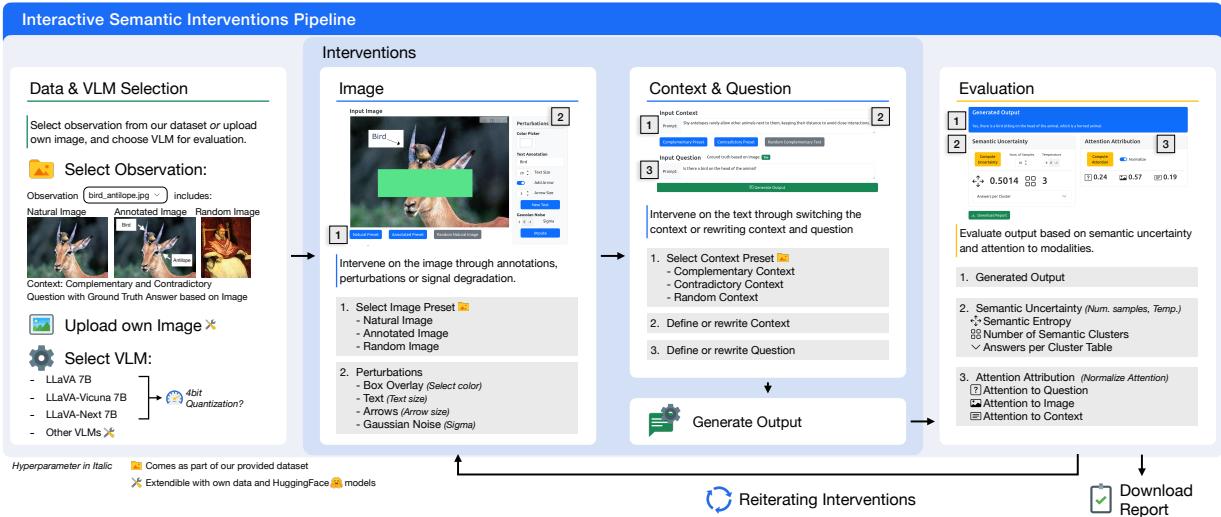


Figure 7: Illustration of the evaluation pipeline used in the ISI for VLMs to enable interactive exploration of VLM behavior under various scenarios. It consists of three main stages: 1) Data & VLM Selection: Users choose an observation either from the provided SI-VQA Dataset or upload their own, and select a VLM for evaluation. 2) Interventions on Image, Context & Question: The selected image can be altered through presets or perturbations, and the context or question can be edited or also switched with presets. 3) Evaluation: The output is analyzed for semantic uncertainty and attention attribution, allowing for iterative refinement of interventions.

E.2 INTERACTIVE SEMANTIC INTERVENTION PIPELINE

Data & Model Selection: As the first step, a user either chooses an observation from the SI-VQA Dataset or uploads their custom image. Each observation from the dataset includes an image, corresponding context, and a question with a ground truth answer, as well as the presets for the annotated images and contradictory and complementary context. The corresponding image, context, and question are displayed. In the next step, the user selects a VLM (LLava, LLava-Vicuna, LLava-Next) and the number of parameters (7B, 11B, 32B) in two separate drop-down menus. 4-bit quantization can be enabled to reduce the computational load and VRAM requirements.

Interventions on the Image: For interventions on the image, ISI allows the user two main functionalities. First, on the proposed SI-VQA Dataset the user can select for each observation three different image presets (natural image without modifications, annotated image with hand-crafted annotations, and random natural image from the dataset) by selecting the respective buttons. Second, ISI allows perturbing the image directly in the tool by overlaying boxes

with selectable colors, inserting and modifying text, adding directional arrows, and introducing Gaussian noise with adjustable noise values.

Interventions on Context & Question: To facilitate the user's ability to observe how various contexts and questions affect the model's performance two functionalities are supported. In the proposed SI-VQA Dataset, users can choose from three distinct context presets—complementary, contradictory, or random—by selecting the respective button, automatically updating the content in the text input fields. Additionally, the user can manually edit the context and question in these fields.

Evaluation The evaluation is designed to enable quantitative analysis of how interventions on image and text impact the behavior of the selected VLM. At the top, the current input is visualized to always relate the evaluation results to the correct input. Below, the generated output, semantic uncertainty, and attention attribution are shown. For computational reasons, each evaluation can be started separately.

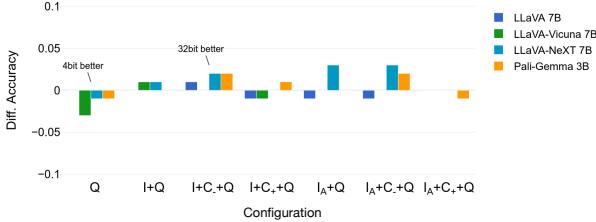
The semantic uncertainty tab allows users to evaluate model uncertainty by clustering sampled outputs according to their semantic meaning. This feature highlights the range of semantic differences in the outputs and calculates semantic entropy, providing a comprehensive view of the model's overall uncertainty. It displays key metrics, such as semantic entropy and the number of semantic clusters, which are influenced by adaptable hyperparameters like the number of samples and the sampling temperature. For deeper exploration, the "Answers per Cluster" dropdown provides a table displaying all sampled answers along with their assigned semantic clusters. This table enables users to examine the full range of generated outputs and understand the semantic similarities within each cluster. To evaluate the significance of each of the three inputs during generation, the attention attribution tab displays the absolute or relative attention assigned to the question, context, and image input tokens.

To provide a contextual understanding of the current observation, the tool additionally displays average values for attention attribution and semantic entropy across the entire SI-VQA Dataset based on each VLM architecture. These averages are shown when hovering over the relevant values.

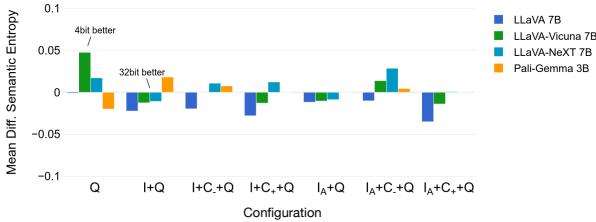
Export The results of one iteration can be exported as a PDF to facilitate the systematic collection of example cases for further analysis and to support the transition from initial qualitative insights to small-scale quantitative evaluation. After the analysis, users can download a comprehensive report that includes the image, context, question, detailed model setup, hyperparameters, and all computed evaluation metrics.

F 4BIT QUANTIZATION

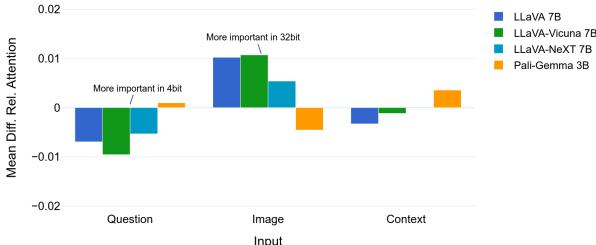
a. Difference 4bit/32bit Quantization VQ Answering Accuracy



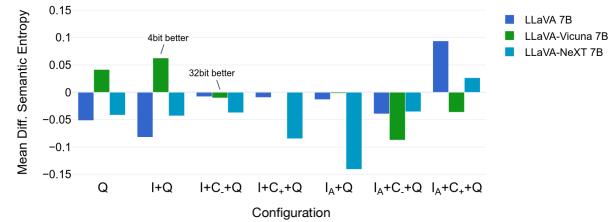
b. Difference 4bit/32bit Quantization Sem. Entropy Answering



d. Difference 4bit/32bit Quantization Attention Answering



c. Difference 4bit/32bit Quantization Sem. Entropy Reasoning



e. Difference 4bit/32bit Quantization Attention Reasoning

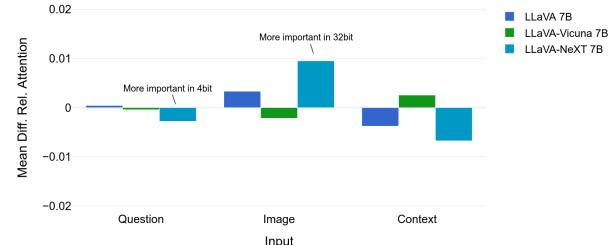


Figure 8: Difference in performance between the 32Bit and 4Bit quantized versions of the models for a. VQA accuracy, b. VQA semantic entropy, c. reasoning semantic entropy, d. VQA attention attribution, and e. reasoning attention attribution.

To quantify the effect of different model sizes, we additionally perform all experiments also with the same models but 4Bit quantized, as there are no, e.g., 3B parameter LLaVA versions. Figure 8 shows the difference in results for all five experiments between the 32Bit and 4Bit models. To our surprise, the difference in accuracy is not that large. Results for the question-only configuration are not meaningful as both models randomly guess. However, in terms of model uncertainty, the 32Bit model usually scores better. Mean differences in attention distribution are almost neglectable as they are at a maximum of 0.01 percentage points. The results show that for simple VQA, quantized models can achieve almost similar accuracy than their significantly larger 32Bit counterparts.

G VALIDATING SEMANTIC ENTROPY

Any assumptions made on the uncertainty of the models should be reflected in the AUGRC. We observe in Figure 4 that in the case of contradicting image and context ($Q+I+C_-$) the AUGRC goes down for all models, reducing the overconfidence in wrong classified samples, which is correctly captured by the semantic entropy. As the model makes more mistakes in configuration $Q+I+C_-$, the set size of wrong classified samples is larger. In the case of $Q+I$ or $Q+I+C_+$ the AUGRC rises again as there is no confidence reducing context. Additionally, the accuracy is higher in the case of $Q+I+C_+$, reducing the set size of wrong classified samples. This empirical evaluation shows we can use the AUGRC to quantify and evaluate the performance of semantic entropy for model failure detection. To our knowledge, this is the first time semantic entropy has been evaluated for failure detection.

H ADDITIONAL RESULTS

H.1 VQA ACCURACY AND QUALITY

a. VQA Confusion Matrices

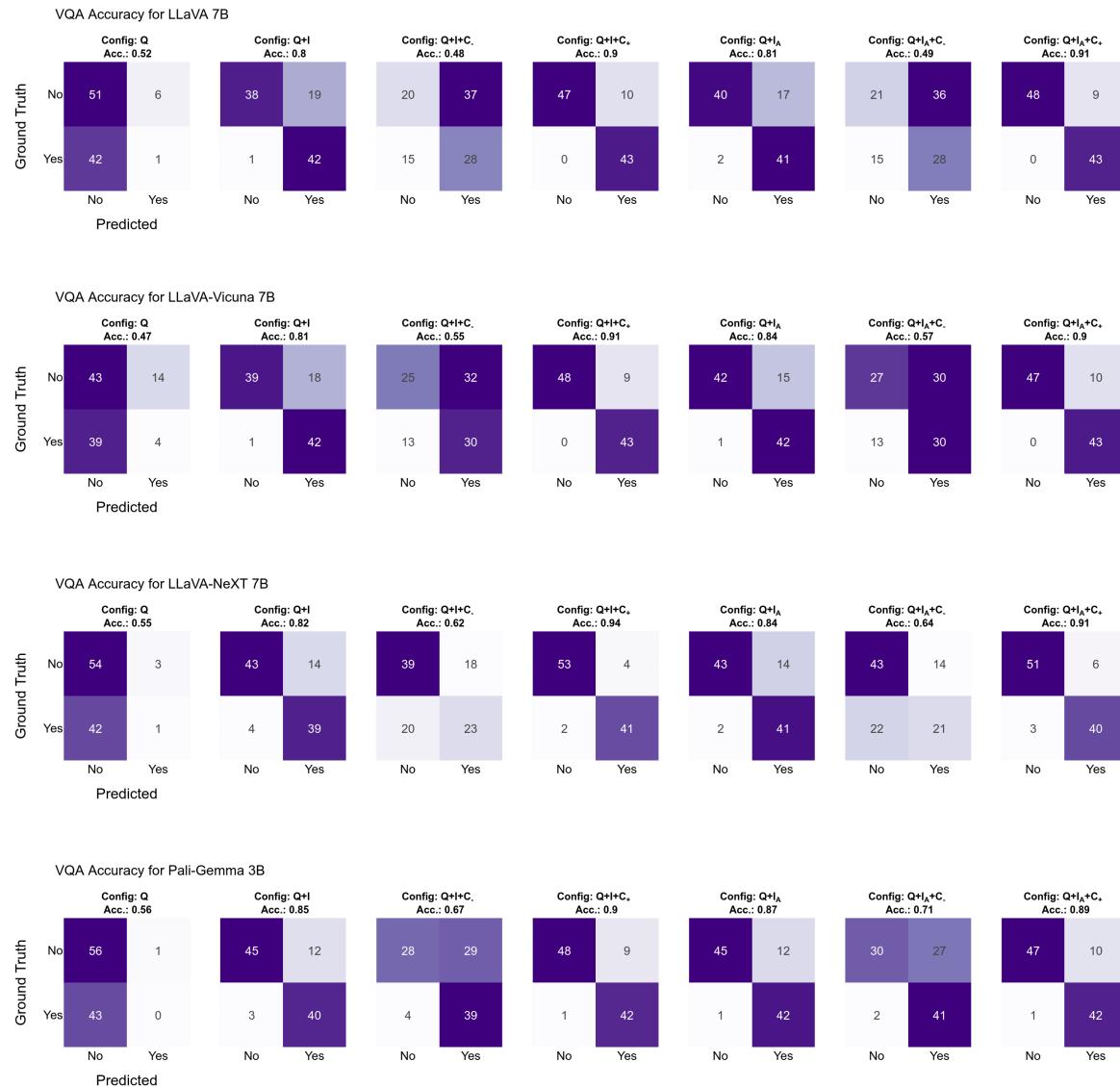


Figure 9: Confusion matrices and accuracy values for all model architectures and configurations.

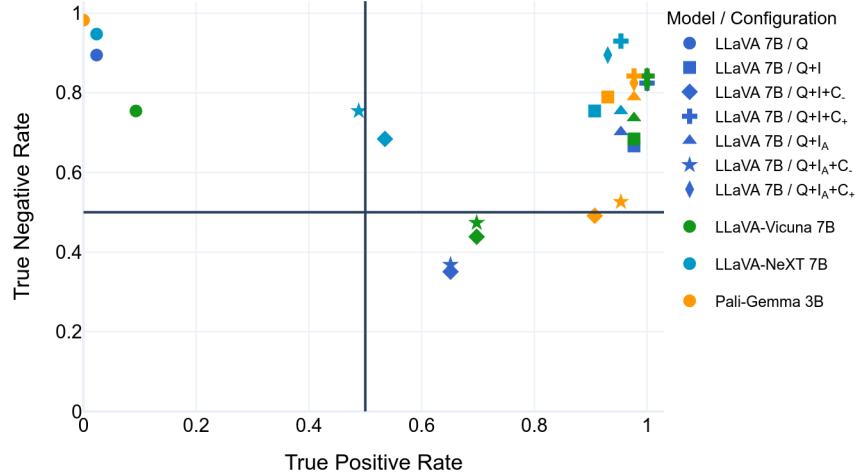


Figure 10: True Negative Rate versus the True Positive Rate in the VQA task of all model architectures and configurations. Values below 0.5 on one axis indicate worse than random guessing performance.

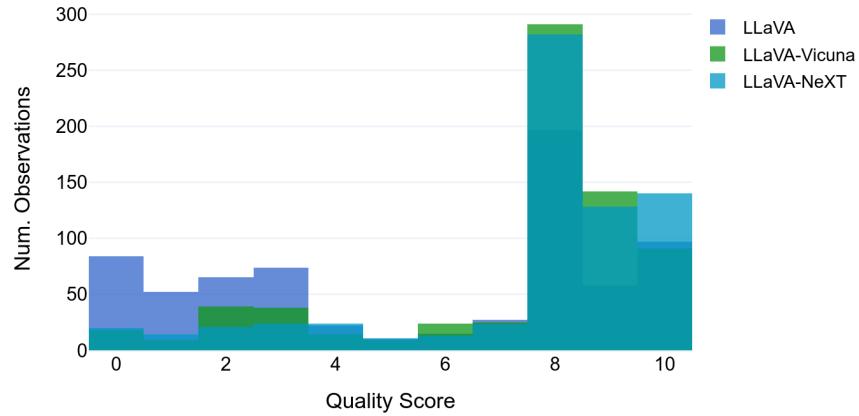
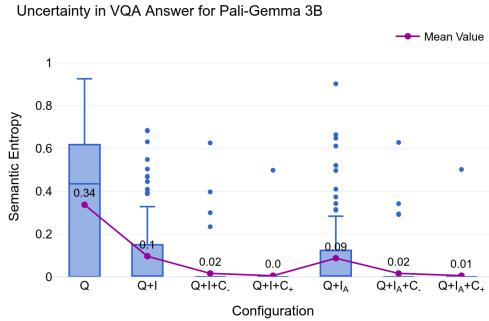
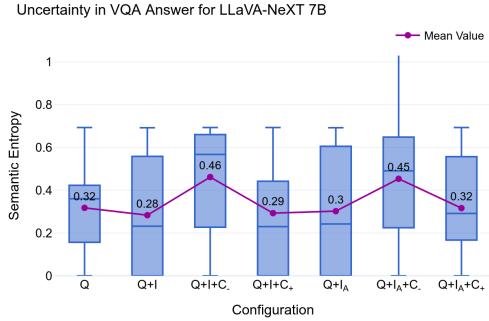
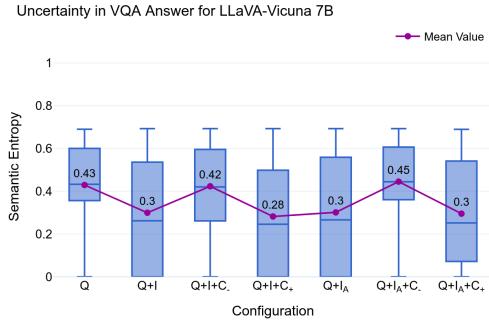
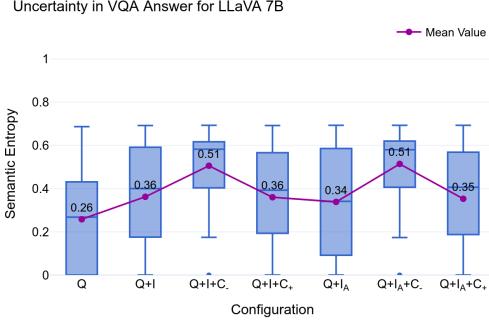


Figure 11: Distribution of reasoning quality scores by GPT-4o for all LLaVA architectures. We observe lower values for the standard LLaVA and a high bias towards the quality score of “8” for all models.

H.2 SEMANTIC ENTROPY

a. Semantic Entropy for VQ Answering (lower is better)



b. Semantic Entropy for VQ Reasoning (lower is better)

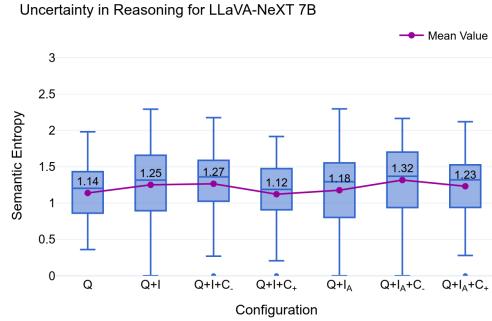
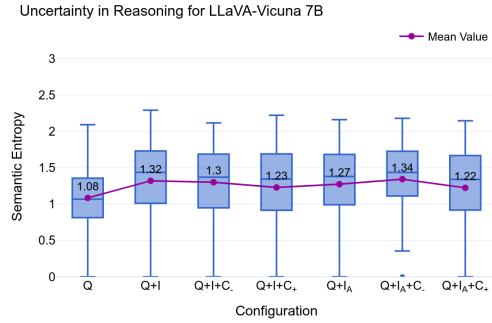
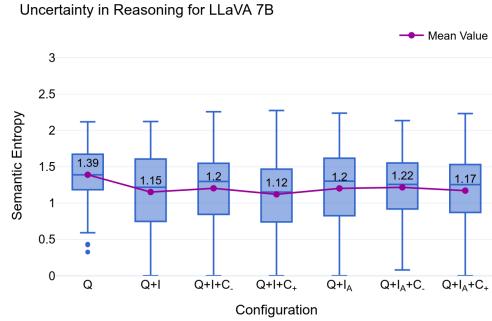


Figure 12: Distribution of semantic entropy for VQ answering and reasoning task for all model architectures and configurations.

H.3 ATTENTION

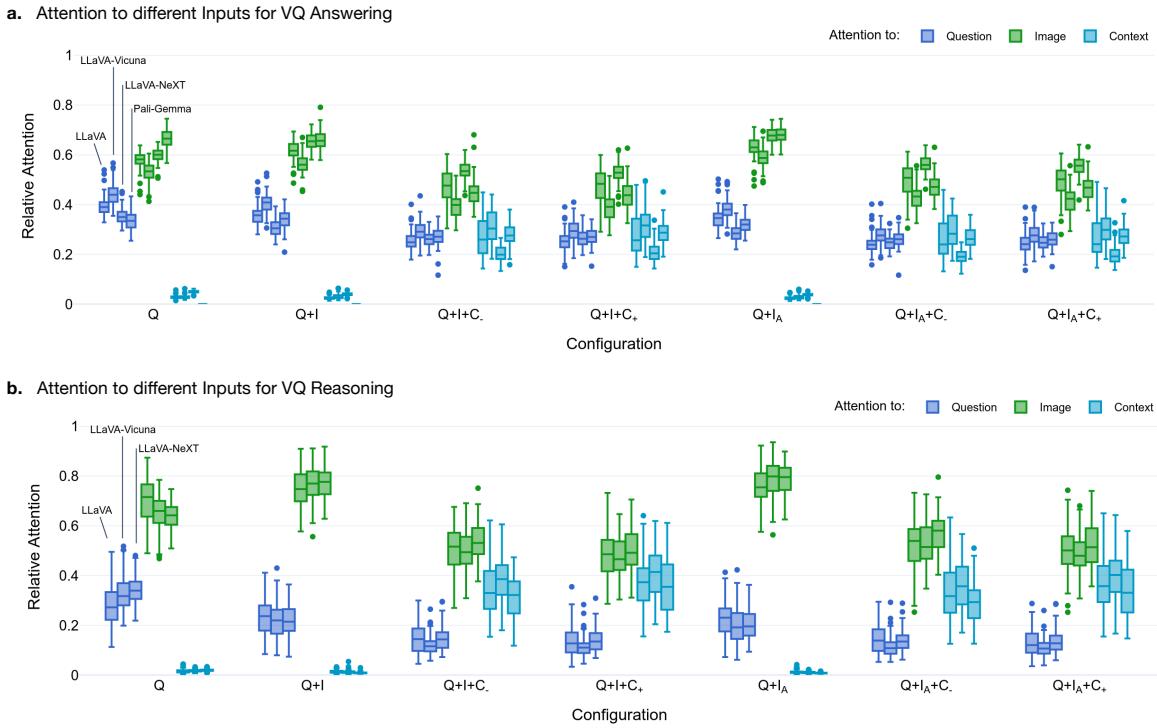


Figure 13: Attention attribution to the question (Q), image (I), and context ($C_{+/-}$) by the three LLaVA models and PaliGemma and the seven modality configurations, during the answering (a.) and the reasoning (b.) in the VQA tasks. Variance across samples is greater for the reasoning than the answering process.

H.4 PEARSON CORRELATION BETWEEN ATTENTION AND ACCURACY

Table 1: The Pearson correlation coefficients (PCC) between the attention to the different inputs and the accuracy per sample. Correlation is between -0.07 and 0.9 for all inputs and model architectures and can fluctuate between architectures.

Model Architecture:	LLAVA 7B	LLAVA-Vicuna 7B	LLAVA-NeXT 7B	PaliGemma 3B
Question	-0.034	-0.034	-0.074	-0.044
Image	0.087	0.002	-0.005	-0.014
Context	-0.036	0.016	0.045	0.026

I REBALANCING MODALITY IMPORTANCE FOR PALIGEMMA 3B

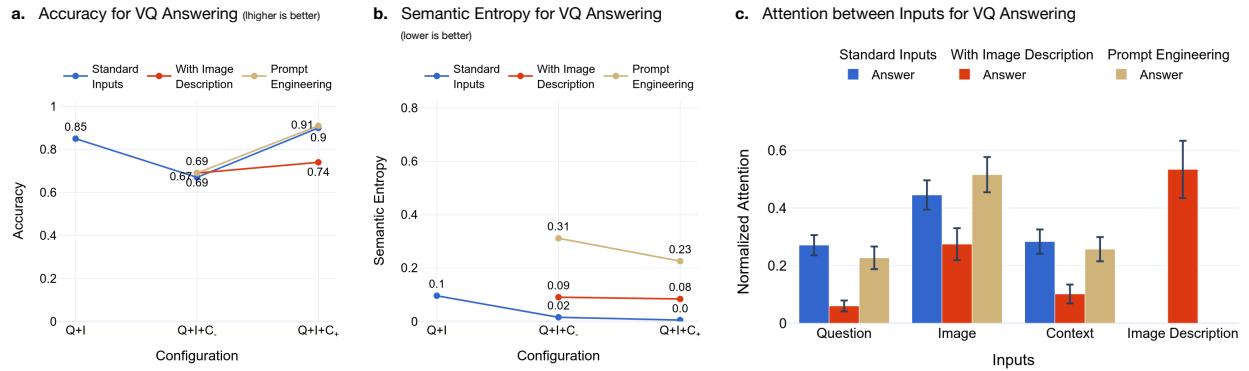


Figure 14: Two strategies to adjust modality importance: incorporating the image’s textual description into the input and modifying the prompt to shift more attention toward the context. The impact of these changes is evaluated based on a. answer accuracy, b. model uncertainty, and c. attention attribution. The hereby experiments were conducted with the PaliGemma model.