Universiteit Leiden

Master Computer Science

Evaluating Llama 3.2 Vision for the Visual Entailment Task: A Study of Zero-shot, Few-shot, and Fine-tuned Approaches

Name:	Elena Pitta
Student ID:	s3840220
Date:	28/05/2025
Specialisation:	Data Science
1st supervisor:	Dr. Tessa Verhoef
2nd supervisor:	Tom Kouwenhoven
3rd supervisor:	Dr. Gijs Wijnholds

Master's Thesis in Computer Science

Leiden Institute of Advanced Computer Science (LIACS)
Leiden University
Niels Bohrweg 1
2333 CA Leiden
The Netherlands

Abstract

This study explores the capabilities of the Llama 3.2 Vision 11B model regarding the Visual Entailment (VE) task within zero-shot, few-shot, and fine-tuning settings. We investigate various factors that might affect the performance of the model, including the prompt design, the number, and the selection strategy that are related to in-context examples for the few-shot inference, and also the order of class labels. We also conducted experiments using randomly cropped images and black images to evaluate the performance of the model when it has limited vision. To evaluate the reasoning of the model, we conduct explanation-based experiments. Results indicate that three-shot inference improves the performance over the zero-shot baselines. However, additional examples introduce more noise than they provide benefits. Additionally, the order of the labels in the prompt is a critical factor that influences the predictions. The black-images experiment reveals the model's tendency to hallucinate, and most of the time it entails the hypotheses. Fine-tuned model achieves an accuracy of 83.3% on the e-SNLI-VE dataset for the VE task, outperforming the state-of-the-art OFA-X model. Additionally, the explanation evaluation demonstrates that the fine-tuned model provides semantically meaningful explanations with a BERTScore F1-score of 0.8916. Further research should examine larger multimodal models and integrate Chain-of-Thought prompting to further improve VE performance.

Contents

List of Figures

List of Tables

1 Introduction

In recent years, there have been advances in **A**rtificial **I**ntelligence (AI) and deep learning, which have improved the development of **N**atural **L**anguage **P**rocessing (NLP) and **C**omputer **V**ision (CV). While these domains were traditionally separate, the emergence of multimodal learning has unified them, allowing systems to interpret, reason, and produce meaning from combined textual and visual input. This has led to tasks such as **I**mage **C**aptioning (IC) [36],**V**isual **Q**uestion **A**nswering (VQA) [1], as well as **V**isual **E**ntailment (VE), the focus of this thesis. VE is an extension of the **T**extual **E**ntailment (TE) task, where the goal is to determine if the hypothesis to the given premise is contradicted, entailed, or neutral [35]. However, the VE introduces a significant challenge, the system must reason and align with more than one modality to reach a reasonable conclusion because the premise is an image rather than a text.

The introduction of the SNLI-VE and e-SNLI-VE datasets provided a crucial benchmark for evaluating this reasoning task. Despite this, VE models tend to underperform compared to their text-only models due to the difficulty of managing both modalities. Previous work, such as EVE [35], OFA [33], FM3 [8], and OFA-X [26], has attempted to address this particular challenge with different approaches, including multimodal transformers.

Recent developments in **M**ultimodal **L**arge **L**anguage **M**odels (MLLMs), such as Flamingo [2], Gemini [29], and Llama 3.2 Vision [13], have demonstrated promising zero-shot and few-shot generalization on a broad range of vision-language tasks. These models can be applied to new and unseen tasks using only in-context examples because they have been pre-trained on large and diverse datasets. MLLMs, despite their advantages, can be sensitive to minor changes in input, such as the order of labels or examples. Thus, questions regarding the extend to which an MLLM, and specifically the Llama 3.2 Vision, can perform VE without fine-tuning and what factors affect its performance in zero- and few-shot settings arise.

The main aim of the project is to understand the capabilities and limitations of the Llama 3.2 Vision model when we perform the VE task and investigate the factors that affect its performance. Thus, our research questions are:

- How does Llama 3.2 Vision perform on the visual entailment task in a zero-shot inference?
- What is the impact on zero-shot inference of having an incomplete or absent dataset?
- What is the impact of few-shot inference on the model's accuracy, and how does performance differ on different numbers of examples?
- How does the order of class labels in the prompt affect model predictions?
- What is the impact of varying the order of examples in a few-shot inference on model performance?
- To what extent does fine-tuning improve model performance compared to zero-shot and few-shot inference?

To answer these questions, we split the experiments into three stages:

- Zero-shot inference with multiple settings (prompt formulation, explanation generation, and limited visual input) to asses the performance.
- Few-shot inference using three and six in-context examples with multiple variations (prompt formulation, ordering of class labels, and example selection).
- Fine-tuning using QLoRa to evaluate whether the Llama 3.2 Vision can surpass the performance of the baseline models and the state-of-the-art models.

The outline of the thesis is structured as follows. An overview of the key concepts in our research is given in Section 2. Section 3 explores the related work on VE models. Section 4 provides the details of the zero-shot, few-shot, and fine-tuning. Section 5 includes the datasets, the baseline models, and the experimental details. Section 6 covers the results from the zero-shot, few-shot, and fine-tuned experiments, and Section 7 covers the limitations and future work. Finally, Section 8 provides the conclusions of the study.

2 Background

In this chapter, we will cover the fundamental concepts that form the background knowledge for the project. We will discuss the topics of Multimodal Learning, LLMs, MLLMs, and VE.

2.1 Multimodal Learning

Multimodal learning is a research field that aims to process and relate information through the integration of multiple modalities such as linguistic, visual, acoustic, and tactile [3] [19]. The importance of multimodal learning lies in its ability to map information across different modalities, thereby enhancing understanding and reasoning [3] [19]. Although integrating multiple modalities often leads to more robust and reliable systems, it is necessary to note that multimodal learning can also introduce difficulties, including the difficulty of aligning with human cognitive processes [4]. There are numerous applications of multimodal learning, including audio-visual speech recognition, video summarization, and healthcare applications such as automatic assessment of depression and anxiety [3].

However, there are still a number of difficulties and challenges. A key challenge is representation [3] [19], which occurs due to the difficulty in learning how to represent and summarize multimodal data in a way that takes advantage of the complementarity and redundancy of multiple modalities [3]. Beyond these, another major problem is generating a coherent output that corresponds to cross-modal interactions. Furthermore, transference [19] or different co-learning [3]is a core challenge, which particularly is the ability to transfer knowledge between different modalities.

2.2 Large Language Models

In the decade of 2010, **R**ecurrent **N**eural **N**etworks (RNNs) were widely used for many natural language applications, including machine translation,

text generation, and text classification [21]. However, RNNs suffer from limitations such as vanishing gradients and long-term dependencies, limiting their effectiveness [15].

The invention of the transformer architecture [31] was a milestone for the development of LLMs and concretely the introduction of the attention mechanisms that capture contextual information across the entire input sequence simultaneously [23]. LLMs are mainly transformer-based language models that have billions of parameters. In addition, LLMs are much larger in model size than regular **L**anguage **M**odels (LMs) and, most importantly, have stronger language understanding and generation abilities [21]. Also, LLMs have a major effect on various fields such as education [15] [23], healthcare [15] [23], and finance [15] [23]. Some examples are summarization of academic papers and generation of research hypotheses [23], analysis and generation of patient information leaflets [23], and sentiment analysis of financial reports and news [23].

LLMs can be grouped into three main categories: encoder-only, decoder-only, and encoder-decoder [21] [23]. The encoder-only family is advantageous for tasks such as sentence classification and **N**amed **E**ntity **R**ecognition (NER), where understanding the whole sentence is necessary [21]. One of the most famous encoder-only models is BERT (**B**idirectional **E**ncoder **R**epresentations from **T**ransformers) [11]. Next, the decoder-only family is great for text generation [23], and the GPT models are notable examples of this category. Most of the LLMs fall into this category. Finally, the encoder-decoder family is the most appropriate for summarization and translation [21] [23], and a remarkable example is T5 [27].

However, LLMs come with important limitations. First, LLMs are probabilistic, suggesting that even with the same prompt, the answer will probably be different [21] [15]. This variability can be beneficial for tasks that require creativity, however, it is a challenge for applications that require consistent and deterministic responses. In addition, LLMs are computationally expensive since they need costly GPUs (**G**raphical **P**rocessing **U**nits) for their training, which also has an environmental impact [21] [15]. Moreover, LLMs can produce hallucinations and untruthful answers [21]. Also, LLMs, according to their training data, may inherit biases, for example, on race or gender [15]. A final limitation is the lack of grounding in the real world. However,

there is active research in this area, with studies exploring how LLMs could achieve real-world understanding [22]. LLMs are trained primarily on textual data, which traditionally limits their ability to directly understand other modalities. This constraint has inspired the creation of MLLMs to resolve the gap through the integration of various modalities.

2.3 Multimodal Large Language Models

MLLMs are LLMs that have extended their capabilities to deal with various types of data, such as image, video, and audio, in addition to text [7]. In other words, MLLMs are LLMS empowered with multimodal capabilities. The need for the development of the MLLMs is derived from the complex real-world task that no longer suffices for the unimodal systems [32]. MLLMs can inherit notable features from LLMs, such as robust language generation and transfer learning abilities [17]. The advantage of MLLMs is that they combine data from different modalities and achieve a more comprehensive understanding and production of information [32], leading to richer and more detailed output. The architecture of MLLMs consists of three main components: a multimodal input decoder, a pre-trained LLM, and a multimodal output decoder [32]. Specifically, the Llama 3.2 Vision model, where we focus on this research, exploits a combination of the Llama 3.1 8B text model with a separately trained vision adapter [12]. MLLMs have broadened their applications to domains including text-to-video generation, image captioning, and text-to-speech. Two state-of-the-art models with exceptional performance are Gemini [29] and GPT-4V [37].

Despite MLLMs' success, these models continue to encounter a number of difficulties. Interpretability is a major challenge. Specifically, it is difficult to understand how different modalities are combined and what the contribution is of each modality to the final decision [32]. Furthermore, the enormous amount of data makes MLLMs prone to the risk of serious security problems, such as bias and data leakage [32]. Beyond these, the high rates of MLLMs' tendency to hallucinate are an important challenge that threatens their reliability [7] [38] [6]. In addition, reducing their computation load is an essential need due to the high computational demand [6].

2.4 Visual Entailment

VE is a novel multimodal task, which is an extension of the traditional TE task. Prior to VE, TE is studied in the field of NLP and particularly in the domain of **N**atural **L**anguage **I**nference (NLI).

In the TE task, given a text Premise P and a text Hypothesis H, the goal is to determine whether the Premise P implies Hypothesis H [35]. The output of the model is a label among the three classes: *Entailment*, *Contradiction*, and *Neutral* based on the relation derived from the text pair (P, H) [35]. If there is sufficient evidence in P to draw the conclusion that H is true, then entailment holds. Wherever H contradicts P, a contradiction is identified. If not, the relation is neutral, suggesting that there is not enough data in P to infer anything from H. The difference between the TE and the VE is the replacement of the text Premise with a real-world image. Therefore the extended task is converted to a multimodal task because of the visual premise and the text hypothesis. Figure 1 indicates an example of VE, showing the image premise and three different text hypotheses, which results in one of the three classes each time.

VE is an important task, and there are several applications to which it can be applied. One application of the VE is fake news detection [35], where social media can verify whether an image entails an article. Moreover, court cross-examination [35] can use VE, where the visual evidence needs to verify the witness statements. Beyond these, VE can be used for e-commerce product verification, so that marketplaces can detect fraud if a product does not match with its description.

However, as mentioned in Subsection 2.1, multimodal tasks have to address several challenges. One example of a challenge is representation because of the difficulty of capturing the semantic overlap between images and text. Another challenge is alignment, where an image may represent ambiguous objects or the actions described in the text hypothesis. In addition, the subjectivity in annotations is also a significant challenge for VE, because two annotators might label differently a pair with the same image and the same hypothesis.

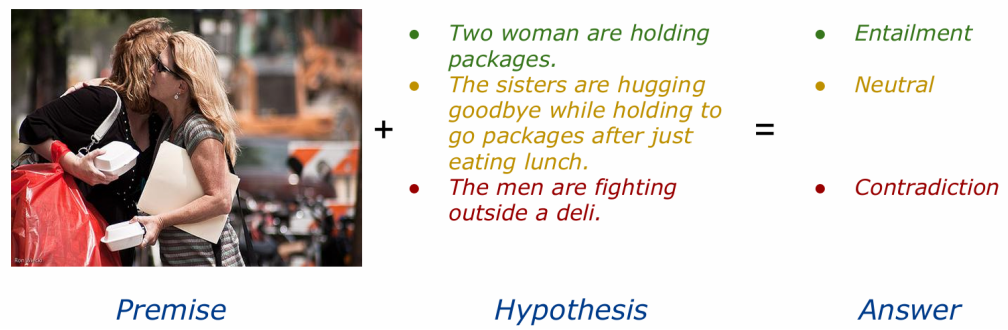


Figure 1: Example of VE task, showing an image premise and three different hypotheses resulting in three labels [35].

3 Related Work

This section covers the key papers on VE work that led to innovations and motivated further exploration in this area. The VE task builts upon the traditional TE, introducing a multimodal challenge at the intersection of computer vision and natural language understanding.

The VE task was introduced by Xie et al (2019) in the paper *Visual Entailment Task for Visually-Grounded Language Learning* [35]. They presented a dataset combining a TE corpus with an image dataset. They also proposed a model called **E**xplainable **V**isual **E**ntailment model (EVE), which uses attention mechanisms to find the inner relationships in both image and text feature spaces [35].

A major advancement in this field came with the OFA model (**O**ne **F**or **A**ll) [33]. OFA is a sequence-to-sequence learning framework and unifies various unimodal and cross-modal tasks, including the VE task. OFA achieves the state-of-the-art performance for the VE task on the SNLI-VE dataset with an accuracy of 91.2% on the test set. SNLI-VE is the most commonly used dataset for VE evaluation. More details about the dataset can be found in Subsection 5.1.

Extending this, OFA-X [26] is a proposal multitask framework that predicts not only the labels but also explanations. OFA-X is a fine-tuned version of

the OFA model and achieved the state-of-the-art performance for the VE task on the e-SNLI-VE dataset with an accuracy of 80.9% on the test set. The dataset was constructed by merging two datasets, which results in a visual entailment task with accompanying natural language explanations [18]. More information on the datasets will follow in Subsection 5.1.

One different direction is the **f**ew-shot learning for a **m**ultimodal **m**ultitask **m**ultilingual (FM3) framework [8]. This approach adapts to new tasks, such as VE, with little supervision through the use of frozen large language models with multimodal inputs. The model leverages the ability of transfer learning for few-shot learning and achieves a high accuracy with little labeled data on the SNLI-VE dataset.

Perhaps the boldest perspective comes from an approach in which the proposal model CLOSE (**C**ross moda**L** transfer **O**n **S**emantic **E**mbeddings) can achieve a comparable performance, without images, using only textual input [14]. For the VE task, the model uses the SNLI dataset for the training (it uses a text premise instead of an image), while for the evaluation, the SNLI-VE dataset was used, which combines vision and language. Despite not using images, CLOSE achieves similar performance to the image model. This suggests that the SNLI dataset may contain sufficient evidence to conclude the relationship without relying heavily on visual information and raises questions about whether a visual grounding is required.

The knowledge from these previous works directly influenced the design of our experiments. Inspired by OFA and FM3, we adopted a prompt-based few-shot setup to investigate how effective a model is without direct supervision. In addition, the idea of explanation generation in OFA-X led us to design an experiment to analyze the explanations from the model, helping assess its interpretability and reasoning. Finally, the innovative approach of the CLOSE model and its findings led us to test different experiments with limited vision to explore the extent of the visual dependency.

4 Methodology

This chapter discusses the methodology to evaluate the performance of the Llama 3.2 Vision model in the VE task. We first explain why the specific model was selected and then introduce three different approaches: zero-shot inference, few-shot inference, and fine-tuning.

4.1 Llama 3.2 Vision

Llama 3.2 Vision[1] is a powerful multimodal large language model released in September 2024 [24]. The model is available in two sizes: 11B and 90B parameters and for this study, we focus on the smaller version. The model can be used to achieve a variety of tasks, including visual recognition, image reasoning, captioning, and answering general questions about an image [24] [12]. The architecture of the model is based on the combination of the Llama 3.1 8B with a separately trained vision adapter [12]. Specifically, the model utilizes a vision adapter that integrates visual information into the text-based architecture of Llama 3.1 through cross-attention layers [2]. During the training phase, the text model was frozen in order to preserve text-only performance [12]. The model was trained on 6 billion image-text pairs with a diverse data mixture [12].

The reason for selecting this specific model among a variety of multimodal models is its competitive performance on standard benchmarks for multimodal tasks. It outperforms many previous open-source models in image understanding and reasoning. In addition, the selection of the smaller model is because the smaller size reduces hardware constraints while keeping the strong reasoning capabilities. Also, the large amount of data that is trained can be beneficial for the zero-shot inference, as it has probably been exposed to various visual-linguistic relationships.

[1] https://ollama.com/library/llama3.2-vision
[2] https://ollama.com/x/llama3.2-vision

```
"""Perform a visual entailment classification.
           You are provided with two inputs:

           1. **Premise**: An image described as follows (attached below).
           2. **Hypothesis**: A text description.

           Your task is to classify the relationship between the Premise (image) and Hypothesis (text) into one of the following three categories:
           - Entailment: The image provides enough evidence to conclude that the Hypothesis is true.
           - Contradiction: The image contradicts the Hypothesis.
           - Neutral: The image does not provide enough information to determine the truth of the Hypothesis.

           Provide a single classification in your response: one of Entailment, Contradiction, or Neutral. Do not include explanations, commentary, or any additional text in your response."""
```

Figure 2: Prompt 1 for zero-shot inference

4.2 Zero-shot Inference

Zero-shot inference describes a model's ability to perform a task without having been trained on it. Concretely, zero-shot is a way to evaluate the model on its generalization ability when applied to tasks that have not yet been seen. In this approach, Llama 3.2 Vision is assessed in the VE task using only its multimodal reasoning and pre-trained knowledge.

For zero-shot inference, the model is provided with a pair of image premise and text hypothesis and must categorize the relationship between them into one of the three classes: entailment, neutral, or contradiction. The output of the model is based on its prior knowledge of visual and natural language concepts, which it obtained during pre-training on a huge number of image-text pairs.

Llama 3.2 Vision is an instruction model, thus, we designed two prompts, where we describe the task and explain each class. The two prompts are shown in Figures 2 and 3. The two prompts are similar, with the only difference being the order of the classes. Specifically, in the first prompt, the order is entailment followed by contradiction, and at the end, the neutral class, while in the second prompt, the order is contradiction followed by neutral and entailment last. The aim of designing similar prompts is to investigate whether the order of classes can affect the results, especially considering recent insights that LLMs are sensitive based on the arrangement of the choices in a multiple-choice prompt's format [28]. Furthermore, multiple small variations of the prompt wording are to determine how they influence classification accuracy.

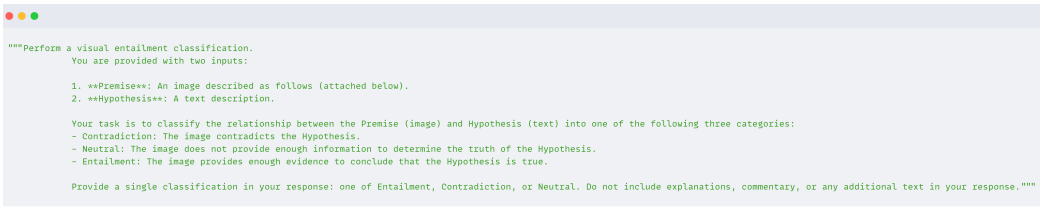


Figure 3: Prompt 2 for zero-shot inference

4.3 Few-shot Inference

Few-shot inference describes a model's ability to generalize to a task without having been trained on it but with only a small number of examples [30]. Unlike zero-shot inference, where the model exploits its pre-trained knowledge and reasoning, in few-shot inference the model also exploits a limited set of labeled examples before making predictions. This technique improves the classification accuracy by allowing the model to identify specific patterns with limited supervision.

For the VE task, a few pairs of image-hypothesis and their labels are demonstrated. The examples are selected in a way that represents all the possible relationships (entailment, contradiction, neutral) to guarantee that the model has been presented with a balance of reasoning patterns. For the experiments, the following strategies are examined:

- **Three-shot and six-shot inferences:** The model is tested with different numbers of examples to examine the effects of additional demonstrations on performance.
- **Variation in class order in the prompt:** The model is tested in different orderings of label classes in the prompt to assess if predictions are impacted by presentation bias.
- **Variation in-context examples order:** The model is tested with the ordering of label classes in the in-context examples to assess if predictions are impacted by presentation bias.
- **Diversity of examples:** The model utilizes two different example selection strategies to check the generalizability of the model.

4.4 Fine-tuning

The process of fine-tuning involves adjusting the weights of a pre-trained model to make it specific to a given task. In contrast to zero-shot and few-shot inference, fine-tuning integrates external knowledge into the model during training, while the former depend on the model's in-context learning capabilities. Usually, for the fine-tuning phase, the training on task-specific data requires only a few epochs, which permits the model to adjust its parameters while preserving the fundamental knowledge gained during pre-training. Additionally, fine-tuning requires a lot less computing power than starting training from scratch.

For the fine-tuning of the Llama 3.2 Vision model on the e-SNLI-VE dataset, we utilized Unsloth [3], a fast and light-weight framework designed to train large models. In order to dramatically reduce the compute and memory requirements, we exploit the technique QLoRA (**Q**uantized **Lo**w-**R**ank **A**daptation) [10]. QLoRA combines LoRA [16] with 4-bit quantization of the model weights, which supports effective fine-tuning of very large models on minimum hardware [9]. The use of 4-bit precision significantly reduces computing cost and memory. LoRA itself preserves most of the original model unchanged while fine-tuning a small set of additional adapter weight metrics (in 16-bit precision)[9]. The number of parameters that must be changed during training is significantly reduced with this method.

5 Experiments

In this chapter, we conducted several experiments in order to evaluate the model. Before the presentation of the results, the dataset used for the experiments will be examined, the evaluation metrics and the baselines will be outlined. Through the experiments, we will be able to draw conclusions on the main questions. In particular, we aim to address the following key questions:

[3] https://unsloth.ai/blog/vision

- How does Llama 3.2 Vision perform on the visual entailment task in a zero-shot inference?
- What is the impact on zero-shot inference of having an incomplete or absent dataset?
- What is the impact of few-shot inference on the model's accuracy, and how does performance vary with different numbers of examples?
- How does the order of class labels in the prompt affect model predictions?
- What is the impact of varying the order of examples in a few-shot inference on model performance?
- To what extent does fine-tuning improve model performance compared to zero-shot and few-shot inference?

5.1 Dataset

The most common dataset used for the VE task is SNLI-VE (**S**tanford **N**atural **L**anguage **I**nference Corpus - **V**isual **E**ntailment) [4]. Specifically, this dataset is a combination of the SNLI (**S**tanford **N**atural **L**anguage **I**nference Corpus) and Flickr30k (image captioning dataset), where the premises from the SNLI are replaced with the corresponding images from Flickr30k [35]. This is feasible because the textual premises in SNLI are the caption sentences of those photos [18].

Although the SNLI-VE dataset is the most common dataset for the VE task, recent research documented that 39% of the neutral labels in the validation and test sets were incorrectly labeled [18]. This is mainly due to the replacement of the text premise with the image premise, which thus led to labeling errors, as an image typically contains more information than a single caption describing it [18]. Thus, the e-SNLI-VE (**E**xplainable **SNLI** - **V**isual **E**ntailment) dataset was created, a merger of the SNLI-VE and the e-SNLI (**E**xplainable **SNLI**), which yields a visual entailment task with explanations

[4] https://github.com/maximek3/e-ViL/tree/main/data

pairID,Flickr30kID,hypothesis,gold_label

3539960792.jpg#4r1e,3539960792.jpg,The person has a piece of athletic equipment.,entailment

Figure 4: Example of the dataset

in natural language. The specific dataset has better quality annotations due to hand-relabeled validation and test sets. Moreover, the e-SNLI-VE dataset has over 430k instances. Table 1 shows how the dataset splits and the number of each class in the sets, and Figure 4 shows an example of a dataset sample. The dataset demonstrates a class imbalance, with contradiction being the most frequent class, followed by entailment with a slightly smaller number of occurrences, and neutral with the fewest cases (Table 1).

Split	**Train**	**Dev**	**Test**
# Images	29,783	1,000	1,000
# Entailment	131,023	5,254	5,218
# Neutral	125,902	3,442	3,801
# Contradiction	144,792	5,643	5,721
# Total Labels	401,717	14,339	14,740

Table 1: Overview of the e-SNLI-VE dataset.

5.2 Baselines

This section outlines the baselines that were utilized in the experiments to compare the performance of Llama 3.2 Vision.

5.2.1 State-of-the-art model

The state-of-the-art model on the VE task with the e-SNLI-VE dataset is the OFA-X model with an accuracy of 80.9% [26]. OFA-X builds upon the OFA model (**O**ne **F**or **A**ll), which is a generative transformer pre-trained on a diverse set of multimodal and unimodal tasks, and fine-tuned on specific vision-language tasks. Concretely, OFA-X leverages the weights of the OFA model and then fine-tunes on targeted datasets, including the e-SNLI-VE. In addition, the original OFA model is the state-of-the-art model on the SNLI-VE, achieving 91.0% accuracy on the validation set and 91.2% on the test set [33]. The difference in performance between OFA-X in e-SNLI-VE (80.9%) and OFA in SNLI-VE (91.2%) demonstrates differences in the quality of the dataset, specifically the existence of mislabeled cases in SNLI-VE, which probably affects the performance of the model.

5.2.2 Zero-shot Experiment on Llama 3.2 Vision

In addition to the OFA-X, we conducted a zero-shot evaluation using the Llama 3.2 Vision on the e-SNLI-VE dataset. The results from this zero-shot experiment serve as a baseline for comparing the effectiveness of zero-shot inference with different settings, few-shot inference, and fine-tuning.

5.3 Evaluation metrics

The performance of the model in the VE task is assessed using multiple evaluation metrics to understand its strengths and weaknesses. Specifically,

the metrics of accuracy, precision, recall, F1-score, and balanced accuracy were used for each experiment [5].

Accuracy
This metric calculates the proportion of correctly classified instances [5].

$$Accuracy = \frac{Number\ of\ correct\ predictions}{Total\ number\ of\ predictions} \tag{1}$$

Precision
This metric measures the proportion of correctly predicted positive instances out of all instances predicted as positive. We use the *weighted precision* because the dataset is imbalanced. This metric averages the precision for each class, weighted by the number of true instances in that class [5].

$$P_i = \frac{True\ Positives_i}{True\ Positives_i + False\ Positives_i} \tag{2}$$

Let n_i be the number of true instances for class i, and P_i be the precision for class i. The weighted precision is then calculated as:

$$Weighted\ Precision = \sum_{i=1}^{3} w_i \cdot P_i \tag{3}$$

where the weight w_i is defined as:

$$w_i = \frac{n_i}{\sum_{j=1}^{3} n_j} \tag{4}$$

Recall
This metric calculates the percentage of actual positive instances that the

[5] https://scikit-learn.org/stable/modules/model_evaluation.html#classification-metrics

model correctly identifies. We use the *weighted recall* because the dataset is imbalanced. This metric averages the recall for each class, weighted by the number of true instances in that class [5].

$$R_i = \frac{True\ Positives_i}{True\ Positives_i + False\ Negatives_i} \tag{5}$$

Let n_i be the number of true instances for class i, and R_i be the recall for class i. The weighted precision is then calculated as:

$$Weighted\ Recall = \sum_{i=1}^{3} w_i \cdot R_i \tag{6}$$

where the weight w_i is defined as equation 4.

F1-score
This metric is the harmonic mean of the metrics of precision and recall. We use the *weighted f1-score* because the dataset is imbalanced. This metric averages the F1-score for each class, weighted by the number of true instances in that class [5].

$$F1 = 2 \times \frac{Precision_i\ \times Recall}{Precision\ +\ Recall} \tag{7}$$

$$Weighted\ F1 = \sum_{i=1}^{3} w_i \cdot F1_i \tag{8}$$

where the weight w_i is defined as equation 4.

Balanced Accuracy
This metric is used for multiclass classification to deal with the imbalanced accuracy [5].

$$Balanced\ Accuracy = \frac{1}{3}\sum_{i=1}^{3} R_i \tag{9}$$

5.4 Experimental details

In this section, the details of the experiments conducted will be discussed. All the experiments were performed using the Llama 3.2 Vision model on the e-SNLI-VE dataset. While the e-SNLI-VE dataset provides explanations, in the majority of experiments, they were ignored and focused only on classification. For the few experiments in which explanations were considered, they are explicitly mentioned. It is also important to note that all the results are based on a single run due to high computational costs and time constraints. The experiments are divided into four main categories: Zero-shot, Three-shot, Six-shot, and Fine-tuning.

5.4.1 Zero-shot Inference

In order to achieve a deterministic output and discourage diversity, the temperature parameter is set to zero (`temperature=0`). For zero-shot inference, we conducted multiple variations of experiments to assess different factors that impact the performance of the model. All experiments were performed twice, one for each prompt described in the chapter of Methodology (Figure 2 and Figure 3).

To further investigate the understanding of visual information we were inspired by the CLOSE model and designed two experiments with limited and removed visual input. Using randomly cropped versions of the original images as premises, we aim to investigate performance with restricted vision. In addition, using black images as premises, we aim to evaluate performance when visual information is completely absent. We hypothesize that these settings will significantly affect the performance of the model. In particular, the conducted experiments are:

- Baseline zero-shot experiments (Prompt 1 and Prompt 2)
- Zero-shot with explanations (Prompt 1 and Prompt 2)
- Zero-shot with Six Prompts (Each sample evaluated in one out of 6 prompts, which are all possible orderings of the three-class labels in the prompt)
- Zero-shot with six prompts per sample (Each sample evaluated using all possible orderings of the three-class labels in the prompt)
- Zero-shot with randomly cropped images (Prompt 1 and Prompt 2)
- Zero-shot with black images as premise (Prompt 1 and Prompt 2)

5.4.2 Three-shot Inference

For the three-shot inference, the model is provided with three labeled examples before making the prediction. Labeled examples are selected from the training set to ensure that the model has never seen the examples in the test set. The experiments are divided into two main categories: Random selection of three examples and Individual selection of three examples (manually chosen examples). For the individual selection, the examples selected have the same premise (image), different hypotheses, and different classes. In both cases, we have made sure that one example from each category is provided. Furthermore, to investigate whether the order of the examples affects the performance of the model, we have performed experiments with a different class as the first example, keeping the same three examples but in a different order. Similarly to the zero-shot inference, we set `temperature=0`. Thus, the experiments conducted are:

- Three-shot with **random** selection of the examples: Contradiction label as a first example (Prompt 1 and Prompt 2)
- Three-shot with **random** selection of the examples: Entailment label as a first example (Prompt 1 and Prompt 2)

- Three-shot with **random** selection of the examples: Neutral label as a first example (Prompt 1 and Prompt 2)
- Three-shot with **individual** selection of three examples: Contradiction label as a first example (Prompt 1 and Prompt 2)
- Three-shot with **individual** selection of three examples: Entailment label as a first example (Prompt 1 and Prompt 2)
- Three-shot with **individual** selection of three examples: Neutral label as a first example (Prompt 1 and Prompt 2)

5.4.3 Six-shot Inference

For the six-shot inference, the model is provided with six labeled examples, randomly selected from the training set. This experiment investigates whether additional examples improve model performance. We have made sure that two examples from each category are provided. For this experiment we select a fair order of the examples, which translates that the class of the first example is different from the class of the last example, and the order of the first 3 examples is not the same as the order of the last three examples. Specifically, the order of the examples' classes is: Contradiction, Neutral, Entailment, Neutral, Contradiction, Entailment. In addition, as in the zero-shot and three-shot inference, we set `temperature=0`. Specifically, the conducted experiments are:

- Six-shot with fair order (Prompt 1 and Prompt 2)

5.4.4 Fine-tuning

As we have mentioned in Section 4.4, for the fine-tuning we used the technique QLoRA, which combines LoRA with 4-bit quantization of the model weights. Table 2 shows some of the main parameters used [6]. We conducted

[6] `https://docs.unsloth.ai/get-started/beginner-start-here/lora-parameters-encyclopedia`

Parameter	Value
num_train_epochs	1
learning_rate	2e-4
warmup_steps (Gradually increases learning rate at the start of training.)	5
r (Rank of decomposition)	8
lora_alpha (Scaling factor for weight updates.)	16
lora_dropout (Dropout rate to prevent overfitting.)	0
evaluation_strategy	epoch
weight_decay (Penalizes large weight updates to prevent overfitting.)	0.01
seed	3407

Table 2: Fine-tuning parameters

two experiments where the main difference is in the existence of explanations. Specifically, the experiments conducted are:

- Fine-tuning without the explanations (Prompt 1)
- Fine-tuning with the explanations (Prompt 1)

6 Results

This section presents results from experiments described in the previous chapter. The experiments highlight the performance of the Llama 3.2 Vision on the VE task, focusing on variations in the order of class labels in the prompt, the number of examples, and the order of examples. All experiments were evaluated in the test set, which contains 14740 instances.

6.1 Zero-shot inference (Baselines)

Table 3 demonstrates the results for prompt 1 (Figure 2) and prompt 2 (Figure 3) for the zero-shot inference. Recall that the only difference in these two prompts is the classes' order, and they are not provided with in-context examples.

In general, prompt 1 achieved an accuracy of 44.5%, while prompt 2 achieved a slightly lower accuracy of 41.3%. These results confirm the modest ability of the model to perform visual entailment in a zero-shot setting. The other metrics (F1, recall, precision, and balanced accuracy) also show that prompt 1 outperforms prompt 2. Notable is that the accuracy is always equal to recall because we use weighted recall [7].

In addition, focusing on the metrics per class for each prompt, we can conclude that the model over-predicts the entailment class due to high recall. Both prompts have high precision but low recall for contradiction, indicating that, while correct predictions are mostly accurate, the model rarely selects this class. The neutral class has the worst per-class results in both prompts. Although the weighted metrics were used for the overall performance to ensure fairness among the imbalanced dataset, however, the fewer instances of the neutral class and the ambiguity that can occur have an impact on the ability of the model to correctly classify that class. Interestingly, the distribution of predictions among the classes differs between the two prompts. Concretely, prompt 1 predicts 57.1% the entailment class, while prompt 2 increases the percentage to 82.7%. Therefore, contradiction and neutral classes are predicted much less often than in prompt 1. This observation suggests that the order of class labels within the prompt significantly affects the predictions of the model. The high rate of entailment predictions in prompt 2 may reflect a recency effect, where the last reported class label becomes more noticeable in the model.

[7] https://scikit-learn.org/stable/modules/generated/sklearn.metrics.recall_score.html#sklearn.metrics.recall_score

Results Prompt 1		**Results Prompt 2**	
Metric	Value	Metric	Value
Accuracy	0.445	Accuracy	0.413
F1	0.409	F1	0.319
Precision	0.536	Precision	0.515
Recall	0.445	Recall	0.413
Balanced Acc.	0.437	Balanced Acc.	0.388
Metrics per class Prompt 1			
Metric	Entailment	Neutral	Contradiction
F1	0.657	0.232	0.299
Precision	0.532	0.205	0.760
Recall	0.859	0.266	0.186
Predictions	8418 (57.1%)	4917 (33.4%)	1403 (9.5%)
Metrics per class Prompt 2			
Metric	Entailment	Neutral	Contradiction
F1	0.587	0.051	0.254
Precision	0.419	0.086	0.887
Recall	0.979	0.036	0.148
Predictions	12197 (82.7%)	1585 (10.8%)	957 (6.5%)

Table 3: Results for zero-shot inference (Baselines).

6.2 Three-shot Inference

For the three-shot inference, 6 experiments were conducted, each tested with the two prompts (prompt 1 and prompt 2), resulting in 12 evaluations. With these experiments, we want to explore how the selection strategy and the ordering of the selected examples affect the model's performance. The selection methods we apply are random selection and individual selection. In random selection, the examples are selected randomly from the training set, and in individual selection, the three examples are selected from the training set so that they have the same premise image but differ in hypothesis and class label. In both strategies, we ensure that one example for each class

is selected. Moreover, in order to asses the impact of example ordering we varied the class label of the first example in each group.

Tables 4 and 5 show the examples selected for the experiments. Tables 6, 7, and 8 highlight the results for random selection with contradiction, entailment, and neutral as the first example, respectively. In addition, Tables 9, 10, and 11 indicate the results for individual selection with contradiction, entailment, and neutral as the first example, respectively.

Number	**ImageID**	**Hypothesis**	**Label**	**Explanation**
319663	1065831604.jpg	The cyclists were going straight.	Contradiction	One group is turning, the other is going straight.
273112	38138101.jpg	A construction worker is welding.	Entailment	Someone who welds on-site is welding by definition.
255488	3636796219.jpg	Guys are playing baseball for charity.	Neutral	Not all guys play for charity.

Table 4: Randomly selected examples for the three-shot inference.

According to the results, we can see that, compared to the zero-shot, in all three-shot experiments, prompt 2 outperforms prompt 1, especially when a contradiction appears as the first example, the model is more robust. Furthermore, the experiments with random selection have better accuracy and F1-scores compared to individual selection, suggesting that the diversity of premises may provide a richer context for the model and contribute positively. When the model is exposed to many diverse images during the few-shot inference, it can learn more distinguishing features. This means that contextual diversity improves generalization and allows the model to identify more extensive semantic patterns. On the other hand, in individual selection, where the premise is the same and differs only in the hypothesis, the model strug-

Number	ImageID	Hypothesis	Label	Explanation
145101	4453784684.jpg	Two men are wrestling each other.	Contradiction	If the wrestler is female, she cannot be one of the two men.
145100	4453784684.jpg	The spectators look upon a female wrestler.	Entailment	The spectators were being diagnosed by doctor.
145099	4453784684.jpg	The doctor is worried about the wrestler.	Neutral	It is not known what the doctor is thinking, so one cannot infer that the doctor is worried.

Table 5: Individually selected examples for the three-shot inference.

gles to understand how to classify correctly and overpredicts the neutral class as the safest option. These findings might seem counterintuitive, however, a diversity of premises is more beneficial than only one premise is given because random selection helps the model to avoid overfitting to the small set of specific characteristics.

Regarding the order of the three in-context examples, we can infer that it has a considerable influence on the outcome. Experiments demonstrate that the first example in the few-shot setting has a large impact on the predictions of the model. When the class of contradiction is the first, the model performs the best, particularly in the experiment with prompt 2 (which also has the class of contradiction as the first in order in the prompt) and random selection (Table 6, 2nd column). Placing contradiction first in the in-context examples

Results Prompt 1		**Results Prompt 2**	
Metric	Value	Metric	Value
Accuracy	0.474	Accuracy	0.487
F1	0.449	F1	0.426
Precision	0.464	Precision	0.469
Recall	0.474	Recall	0.487
Balanced Acc.	0.442	Balanced Acc.	0.448
Metrics per class Prompt 1			
Metric	Entailment	Neutral	Contradiction
F1	0.587	0.139	0.527
Precision	0.476	0.182	0.641
Recall	0.766	0.113	0.448
Predictions	8387 (56.9%)	2354 (16%)	3998 (27.1%)
Metrics per class Prompt 2			
Metric	Entailment	Neutral	Contradiction
F1	0.591	0.053	0.523
Precision	0.441	0.156	0.702
Recall	0.895	0.032	0.416
Predictions	10583 (71.8%)	768 (5.2%)	3389 (23%)

Table 6: Results for three-shot inference with random selection and Contradiction as first example.

may cause a primacy bias that helps mitigate the model's strong bias toward predicting the entailment class in the corresponding zero-shot scenario.

When comparing the results of zero-shot inference with those of three-shot inference, we can observe an improvement in the performance when adding in-context examples. Concretely, the best accuracy and F1-score for zero-shot is 44.5% (prompt 1) and 40.9%, while the best performance for three-shot (prompt 2, random selection and contradiction as the first example) is 48.7% and 42.6%, respectively. The improvement in the balance by class and F1 score for the three-shot inference, particularly for the contradiction class, suggests a more robust understanding of the task, although the increase in

accuracy may seem modest. Specifically, the model significantly overpredicts the entailment class in zero-shot results (prompt 2 yields it in over 80% of cases). On the other hand, three-shot inference mitigates this bias and produces more balanced class predictions, especially when random selection is used. This shift suggests that examples with few shots lead the model to a more nuanced decision process rather than to a dominant class. Thus, it reduces the strong class prediction bias present in zero-shot inference. Moreover, the order of the class labels in the prompt seems to have a less severe effect on the prediction when the model has been given three in-context examples, indicating that few-shot learning provides a stabilizing influence on class prediction.

Results Prompt 1		**Results Prompt 2**	
Metric	Value	Metric	Value
Accuracy	0.412	Accuracy	0.429
F1	0.370	F1	0.380
Precision	0.486	Precision	0.494
Recall	0.412	Recall	0.429
Balanced Acc.	0.392	Balanced Acc.	0.404
Metrics per class Prompt 1			
Metric	Entailment	Neutral	Contradiction
F1	0.554	0.146	0.351
Precision	0.418	0.174	0.756
Recall	0.821	0.126	0.229
Predictions	10243 (69.5%)	2760 (18.7%)	1731 (11.7%)
Metrics per class Prompt 2			
Metric	Entailment	Neutral	Contradiction
F1	0.559	0.121	0.389
Precision	0.414	0.176	0.777
Recall	0.859	0.092	0.260
Predictions	10825 (73.4%)	2003 (13%)	1911 (13%)

Table 7: Results for three-shot inference with random selection and Entailment as first example.

Results Prompt 1		**Results Prompt 2**	
Metric	Value	Metric	Value
Accuracy	0.421	Accuracy	0.426
F1	0.372	F1	0.377
Precision	0.516	Precision	0.528
Recall	0.421	Recall	0.426
Balanced Acc.	0.411	Balanced Acc.	0.413
Metrics per class Prompt 1			
Metric	Entailment	Neutral	Contradiction
F1	0.594	0.216	0.273
Precision	0.456	0.216	0.770
Recall	0.850	0.215	0.166
Predictions	9723 (66%)	3781 (25.7%)	1235 (8.4%)
Metrics per class Prompt 2			
Metric	Entailment	Neutral	Contradiction
F1	0.594	0.204	0.292
Precision	0.453	0.212	0.806
Recall	0.862	0.197	0.179
Predictions	9928 (67.4%)	3544 (24%)	1268 (8.6%)

Table 8: Results for three-shot inference with random selection and Neutral as first example.

6.3 Six-shot Inference

For the six-shot inference, the model is provided with six in-context examples instead of three as in the previous experiments. This experiment was motivated by the observation that increasing the number of examples can help the model better generalize the task and improve the performance [5]. The in-context examples are selected randomly from the training set, and it is ensured that we have 2 examples of each class. However, we performed only one six-shot experiment, and we did not test all the possible permutations of the class ordering as we did in the three-shot settings. This decision was

Results Prompt 1		Results Prompt 2	
Metric	Value	Metric	Value
Accuracy	0.399	Accuracy	0.451
F1	0.396	F1	0.454
Precision	0.503	Precision	0.482
Recall	0.399	Recall	0.451
Balanced Acc.	0.404	Balanced Acc.	0.432
Metrics per class Prompt 1			
Metric	Entailment	Neutral	Contradiction
F1	0.284	0.327	0.543
Precision	0.633	0.244	0.556
Recall	0.183	0.498	0.531
Predictions	1508 (10.2%)	7766 (52.7%)	5465 (37.1%)
Metrics per class Prompt 2			
Metric	Entailment	Neutral	Contradiction
F1	0.429	0.260	0.607
Precision	0.567	0.226	0.574
Recall	0.345	0.306	0.644
Predictions	3173 (21.5%)	5153 (35%)	6413 (43.5%)

Table 9: Results for three-shot inference with individual selection and Contradiction as the first example.

made because the number of possible permutations grows rapidly when we have six in-context results. Figure 12 presents the selected examples. After the random selection, to have a fair order, the examples are manually shuffled so that the class's order of the first three examples is different from the order of the last three examples, and also that the first and last examples are not in the same class.

The results of the six-shot inference are shown in Table 13. Looking at the results, prompt 2 outperforms slightly prompt 1, achieving an accuracy of 36.5% and 35.0% respectively. The metrics per class show that the model overpredicts the neutral class for both prompts, and it is the dominant class

Results Prompt 1		**Results Prompt 2**	
Metric	Value	Metric	Value
Accuracy	0.291	Accuracy	0.299
F1	0.268	F1	0.278
Precision	0.489	Precision	0.486
Recall	0.291	Recall	0.299
Balanced Acc.	0.323	Balanced Acc.	0.334
Metrics per class Prompt 1			
Metric	Entailment	Neutral	Contradiction
F1	0.323	0.322	0.182
Precision	0.362	0.224	0.781
Recall	0.292	0.575	0.103
Predictions	4207 (28.5%)	9778 (66.3%)	753 (5.1%)
Metrics per class Prompt 2			
Metric	Entailment	Neutral	Contradiction
F1	0.300	0.338	0.217
Precision	0.367	0.232	0.762
Recall	0.253	0.623	0.127
Predictions	3596 (24.4%)	10193 (69.1%)	951 (6.5%)

Table 10: Results for three-shot inference with individual selection and Entailment as first example.

with over 70% classified as neutral. This leads to a high recall for the neutral class but lower precision, which means that from the neutral predictions, only a few are correctly classified. The other classes have high recall and low precision, indicating that while correct predictions are mostly accurate, the model rarely selects these classes.

When comparing the results of six-shot with those of three-shot, we can extract some important conclusions. Firstly, the performance does not consistently improve with more in-context examples. Although it was expected that performance would improve and the model would generalize better as the number of in-context examples increased [5], this did not occur in our

Results Prompt 1		**Results Prompt 2**	
Metric	Value	Metric	Value
Accuracy	0.300	Accuracy	0.308
F1	0.206	F1	0.224
Precision	0.624	Precision	0.620
Recall	0.300	Recall	0.308
Balanced Acc.	0.367	Balanced Acc.	0.373
Metrics per class Prompt 1			
Metric	Entailment	Neutral	Contradiction
F1	0.171	0.414	0.100
Precision	0.728	0.265	0.768
Recall	0.097	0.950	0.053
Predictions	692 (4.7%)	13648 (92.6%)	397 (2.7%)
Metrics per class Prompt 2			
Metric	Entailment	Neutral	Contradiction
F1	0.194	0.414	0.125
Precision	0.708	0.266	0.774
Recall	0.112	0.938	0.068
Predictions	826 (5.6%)	13413 (90.1%)	501 (3.4%)

Table 11: Results for three-shot inference with individual selection and Neutral as first example.

six-shot experiment. The best performance of six-shot (36.5%) is actually lower than the best performance of three-shot (48.7%). This suggests that improved performance in few-shot experiments would not always reflect a more in-depth understanding of the task. Instead, it may be a result of random biases that inflate metrics such as accuracy, including overfitting to the dominant order. Secondly, in a six-shot experiment, there is class bias because the majority of the predictions are classified as neutral, which is also observed in the individual selection of the three-shot experiments.

Compared to zero-shot, six-shot inference has a slightly more balanced performance per class, as reflected by the increase in F1-score on the classes of

Number	ImageID	Hypothesis	Label	Explanation
319663	1065831604.jpg	The cyclists were going straight.	Contradiction	One group is turning, the other is going straight.
255488	3636796219.jpg	Guys are playing baseball for charity.	Neutral	Not all guys play for charity.
273112	38138101.jpg	A construction worker is welding.	Entailment	Someone who welds on-site is welding by definition.
109227	1001633352.jpg	Four people are in a life or death situation.	Neutral	Not everyone jumping from the top of the stairs is in a life or death situation.
39031	4597029194.jpg	An old man playing video games on his Laptop	Contradiction	A man can be either playing an accordion or playing video games.
130144	3426964258.jpg	Some people in a group are holding multicolored flags.	Entailment	Some people in a group are people, and green, white, and tan striped flags are multicolored flags.

Table 12: Randomly selected examples for the six-shot inference.

contradiction and neutral, but results in lower overall accuracy. Specifically, zero-shot achieves an accuracy of 44.5% while six-shot achieves an accuracy of 36.5%. This indicates that additional in-context examples can introduce noise, especially if the model overfits patterns in the examples that do not generalize well to unseen test instances. Conversely, while zero and three-shot experiments lead to increased accuracy, this enhanced performance may not always be due to the right reasons.

Results Prompt 1		Results Prompt 2	
Metric	Value	Metric	Value
Accuracy	0.350	Accuracy	0.365
F1	0.319	F1	0.356
Precision	0.565	Precision	0.559
Recall	0.350	Recall	0.365
Balanced Acc.	0.398	Balanced Acc.	0.402
Metrics per class Prompt 1			
Metric	Entailment	Neutral	Contradiction
F1	0.240	0.405	0.333
Precision	0.573	0.268	0.754
Recall	0.152	0.828	0.214
Predictions	1381 (9.4%)	11738 (79.6%)	1621 (11%)
Metrics per class Prompt 2			
Metric	Entailment	Neutral	Contradiction
F1	0.373	0.383	0.322
Precision	0.523	0.262	0.788
Recall	0.289	0.715	0.203
Predictions	2883 (20%)	10386 (70.5%)	1471 (10%)

Table 13: Results for six-shot inference with random selection and fair order.

6.4 Zero-shot Inference

In this subsection, we conducted five experiments to investigate whether the performance observed in zero-shot baseline experiments reflects a true understanding of the VE task or whether biases lead to it. These experiments aim to test the reliability of the model.

Specifically, we implemented a zero-shot experiment where each sample was evaluated using one of the six permutations of the three class labels. Instead of using the same prompt structure for all test samples, in this experiment, each sample was assigned to one specific label ordering. Then, we conducted

a zero-shot experiment in which each example was evaluated with all six class orderings. In addition, we investigated whether adding an explanation to the output would give us a better justification for the class selection. Furthermore, we performed an experiment where we replaced the premise images with randomly cropped versions. Finally, we perform an experiment where we replace the premise images with black images.

Table 14 shows the results for the zero-shot with 6 prompts. Each instance was evaluated in one of six prompts, which are all possible orderings of the three-class labels in the prompt. This experiment aimed to mitigate the positional bias observed in the previous experiments. The accuracy (40.8%) is decreased compared to the accuracy of the baseline models (44.5% and 41.3%). Looking at the metrics per class, most predictions are in the entailment class, and the contradiction class is selected less than 10%, as in the baseline models.

Results for 6 prompts			
	Metric	Value	
	Accuracy	0.408	
	F1	0.336	
	Precision	0.521	
	Recall	0.408	
	Balanced Acc.	0.396	
Metrics per class			
Metric	Entailment	Neutral	Contradiction
F1	0.622	0.171	0.184
Precision	0.472	0.168	0.801
Recall	0.911	0.174	0.104
Predictions	10066 (68.3%)	3928 (26.6%)	743 (5%)

Table 14: Results for zero-shot with 6 prompts. Each sample was evaluated in one of 6 prompts, which are all possible orderings of the three-class labels in the prompt.

Table 15 presents the results for the zero-shot experiment, in which each instance was evaluated in all the permutations of the class labels in the prompt.

The overall accuracy shows how many predictions match with the ground truth, while the majority vote accuracy counts a prediction as correct only if at least four out of six outputs match the correct label. The drop in majority vote accuracy compared to the overall accuracy suggests that the model frequently changes predictions across different prompts, highlighting its sensitivity. Figure 5 shows the per-class performance. The high proportion of recall for entailment indicates that the model frequently predicts entailment. In contrast, contradiction achieves a high precision (80.70%) but suffers from a low recall (11.79%), implying that while the model is usually correct when predicting contradiction, it rarely chooses this class. The neutral class consistently shows the weakest performance across all metrics, suggesting that the model struggles to identify neutral relationships in the VE task. To explore the sensitivity of the prompt, Figure 6 illustrates the accuracy obtained for each of the six prompts. The accuracy fluctuates across prompts, confirming that the order of class labels in the prompt influences the model's predictions. Furthermore, Figure 7 reveals how often the model's prediction changes per sample across the six prompts. Almost half samples (7106) received the same prediction across all six prompts, which indicates that the model was fully consistent for those cases. However, 6647 samples had two different predictions, and 964 samples had three different predictions, confirming that the model was highly inconsistent for some cases. These results demonstrate the instability of the model's output under minimal modifications and are consistent with the drop in majority vote accuracy.

Results for 6 prompts each instance	
Overall Accuracy	0.410
Majority Vote Accuracy	0.337

Table 15: Accuracy for zero-shot inference with 6 prompts. Each sample was evaluated in 6 prompts.

For the next experiment, we modified the prompt slightly so that the model was required to explain its choice, except for the label. To analyze the reasoning behavior of the model, Figure 8 demonstrates the average length of the explanation per class for both prompts. In both cases, Neutral explanations are longer, indicating that the model over-explains these cases. The model requires more justification to explain the lack of implication or contradiction, and this is because neutral often contains ambiguity, which makes it

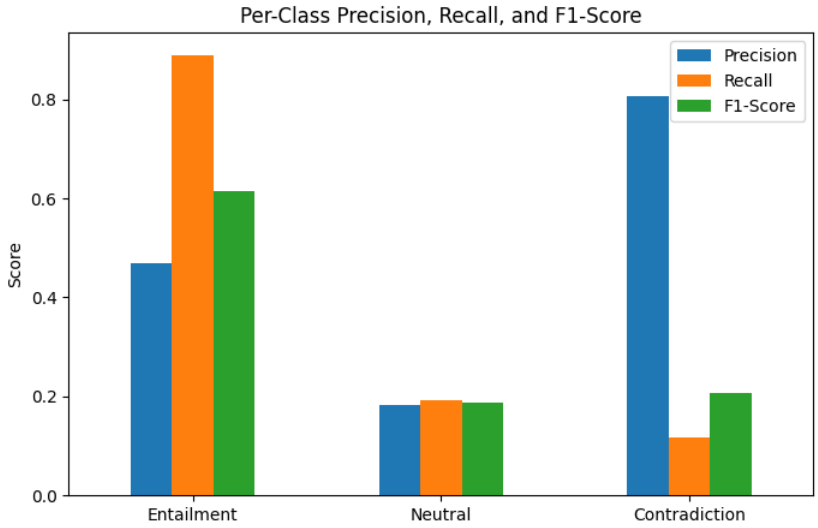


Figure 5: Results for performance per class for the zero-shot inference with 6 prompts for each instance.

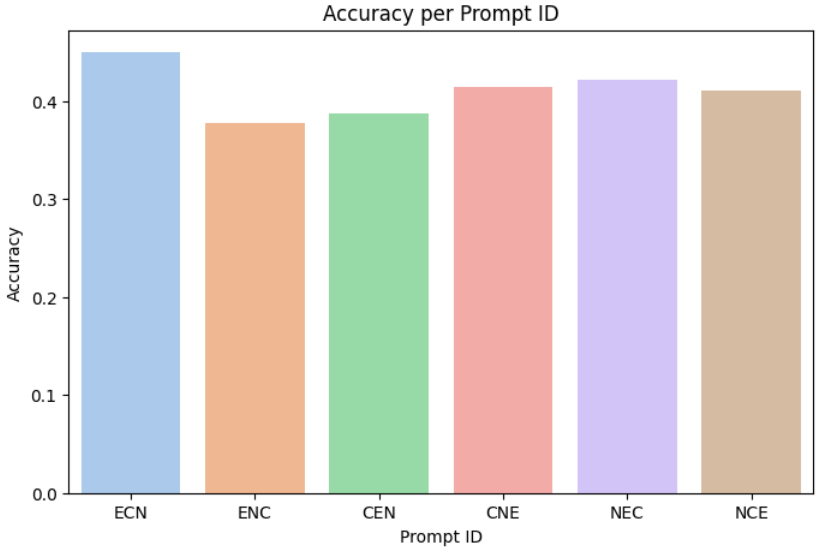


Figure 6: Prompt sensitivity

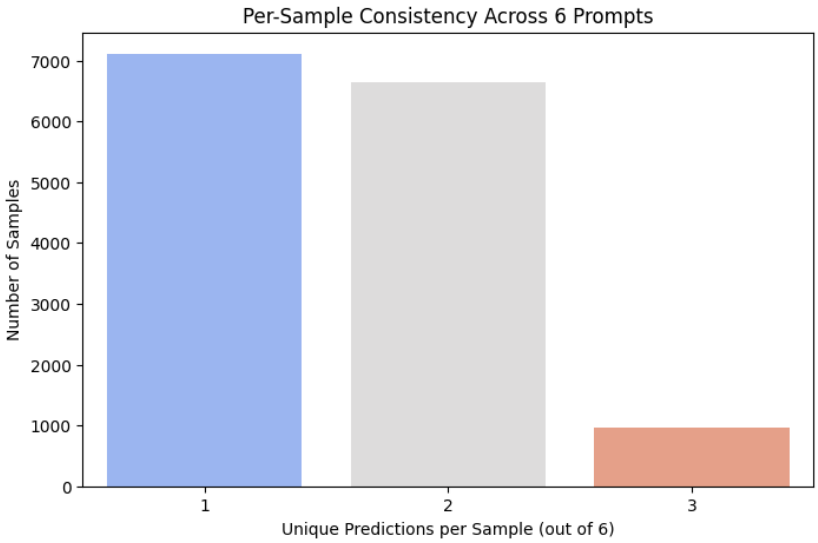
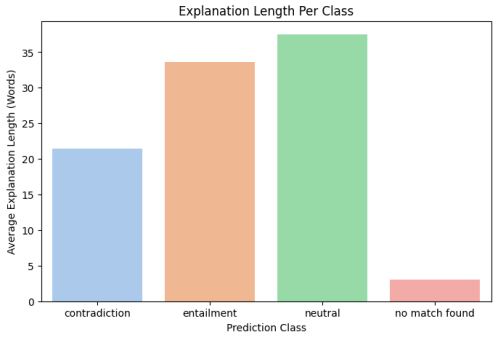


Figure 7: Results per sample consistency across six prompts.

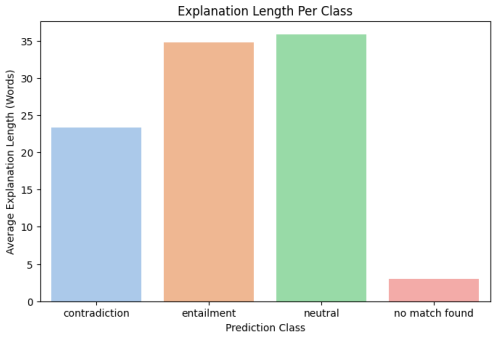
more difficult to justify them concisely. In addition, the prompts have similar length explanations for the classes of entailment and neutral. This suggests that for entailment and neutral, the model is robust to prompt variations. Figure 9 shows an example where the model incorrectly predicts neutral instead of contradiction, and the explanations are similar. Conversely, Figure 10 presents an example where the model predicts different outputs depending on the prompt. One prediction matches the true label, while the other does not. However, both explanations are logically correct, indicating that even when the prediction is incorrect, the model can still produce a reasonable logic.

For the next experiment, we randomly cropped versions of the original images and used them as premises. As shown in Table 16, overall performance decreases compared to baseline results. The decrease in accuracy was expected since the cropped images do not have all the details of the original images. Figure 11 presents an example of the original image and the new cropped image used as a premise.

The last experiment with zero-shot is to replace the original images with black images and ask for an explanation. The results from this experiment



(a) Explanations length per class for prompt 1



(b) Explanations length per class for prompt 2

Figure 8: Length per class for textual explanations

are very interesting because the model tends to hallucinate. Table 17 shows the performance of the model, which is worse than the performance of the baseline models, which is expected due to the removal of the model's visual ability to make accurate predictions. As shown from the metrics per class, the model overpredicts the entailment class while severely underpredicts the contradiction class (0.79% and 1.0% for prompt 1 and prompt 2, respectively). Figure 12 shows an example of the original image. Although the original image clearly shows two baseball players with baseball uniforms and a baseball bat, the model entails the hypothesis *"2 basketball players"*. The hallucination of the model is even clearer in the reasoning: *"The image shows two individuals wearing basketball jerseys, which is consistent with the description of basketball players."* This result shows that the model imagines visual evidence to support the hypothesis and justifies the high proportions of entailment class (84.17% and 89.95% for prompt 1 and prompt 2, respectively).

6.5 Fine-tuning

Finally, the last two experiments are fine-tuning the model using prompt 1, due to time constraints.

premise,hypothesis,label,prediction,explanation
Prompt 1:
flickr30k_images/3416050480.jpg,"A person is at a diner, ordering an omelette.",**contradiction,neutral**,"The image shows a horse and rider jumping over an obstacle, which does not provide enough information to determine the truth of the text description about someone ordering an omelette at a diner."

Prompt 2:
flickr30k_images/3416050480.jpg,"A person is at a diner, ordering an omelette.",**contradiction,neutral**,"The image shows a horse and rider jumping over an obstacle, which does not provide enough information to determine the truth of the Hypothesis about someone ordering an omelette at a diner."

Figure 9: Example of the experiment zero-shot inference with explanations.

premise,hypothesis,label,prediction,explanation
Prompt 1:
flickr30k_images/97162852.jpg,Woman wearing nothing sewing her thumb.,**contradiction,contradiction**,"** The image depicts a woman wearing a red and white plaid shirt, not ""nothing"", while operating a sewing machine. This contradicts the text description of the woman being naked."

Prompt 2:
flickr30k_images/97162852.jpg,Woman wearing nothing sewing her thumb.,**contradiction,neutral**,"** The image does not provide enough information to determine the truth of the Hypothesis, as it is unclear what the woman is actually doing or what she is wearing. The description ""sewing her thumb"" could be interpreted in various ways, and without more context or visual evidence, it is impossible to definitively classify the relationship between the image and the text."

Figure 10: Example 2 of the experiment of zero-shot inference with explanations.

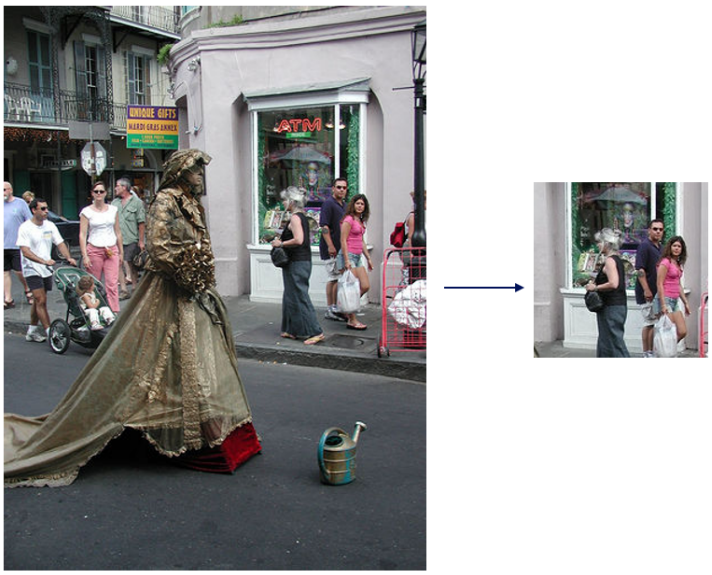


Figure 11: Example of the experiment zero-shot inference with the cropped versions of the original images.

premise,hypothesis,label,prediction,explanation
Prompt 1:
flickr30k_images/2725441885.jpg,2 basketball players,**contradiction,entailment**,"The image shows two individuals wearing basketball jerseys, which is consistent with the description of basketball players."

Prompt 2:
flickr30k_images/2725441885.jpg,2 basketball players,**contradiction,entailment**,"The image shows two individuals wearing basketball jerseys, which is consistent with the description of basketball players."

Figure 12: Example of the experiments with black premise. (The image is the original image.)

Results Prompt 1		Results Prompt 2	
Metric	Value	Metric	Value
Accuracy	0.344	Accuracy	0.380
F1	0.288	F1	0.321
Precision	0.533	Precision	0.552
Recall	0.344	Recall	0.380
Balanced Acc.	0.363	Balanced Acc.	0.386
Metrics per class Prompt 1			
Metric	Entailment	Neutral	Contradiction
F1	0.535	0.296	0.057
Precision	0.471	0.223	0.794
Recall	0.618	0.440	0.030
Predictions	6848 (46.46%)	7480 (50.75%)	214 (1.45%)
Metrics per class Prompt 2			
Metric	Entailment	Neutral	Contradiction
F1	0.572	0.268	0.128
Precision	0.459	0.227	0.853
Recall	0.760	0.328	0.069
Predictions	8638 (58.60%)	5502 (37.33%)	493 (3.14%)

Table 16: Results for zero-shot inference with randomly cropped images as premise.

The results for the first experiment without any explanation are presented in Table 18. As shown in Table 18, the model achieved a high overall accuracy of 83.3%, with an F1 score of 0.836, precision of 0.846, and recall of 0.833. These results indicate that the model generalizes well across the three classes. The most challenging class is the neutral, because the precision (0.679) in combination with the high recall (0.807) indicates that the model frequently misclassifies other classes as neutral. When compared to zero and few-shot experiments, the fine-tuned model shows a significant improvement in both general and class-specific performance. When comparing with the state-of-the-art model OFA-X, which achieved an accuracy of 80.9%, the Llama 3.2 Vision fine-tuned model performs competitively, and outperforms, achieving

Results Prompt 1		**Results Prompt 2**	
Metric	Value	Metric	Value
Accuracy	0.360	Accuracy	0.369
F1	0.250	F1	0.246
Precision	0.504	Precision	0.543
Recall	0.360	Recall	0.369
Balanced Acc.	0.353	Balanced Acc.	0.357
Metrics per class Prompt 1			
Metric	Entailment	Neutral	Contradiction
F1	0.520	0.211	0.031
Precision	0.369	0.290	0.769
Recall	0.878	0.165	0.016
Predictions	**12407 (84.17%)**	2163 (14.67%)	117 (0.79%)
Metrics per class Prompt 2			
Metric	Entailment	Neutral	Contradiction
F1	0.531	0.161	0.043
Precision	0.370	0.317	0.851
Recall	0.940	0.108	0.022
Predictions	**13259 (89.95%)**	1297 (8.80%)	148 (1.00%)

Table 17: Results for zero-shot inference with black images as premise.

83.3% accuracy.

The second experiment aims to evaluate the explanations of the model. Metrics like BLEU [25] (**Bi**ngual **E**valuation **U**nderstudy) and ROUGE [20] (**R**ecall-**O**riented **U**nderstudy for **G**isting **E**valuation) are traditional metrics used for evaluation of generated text, however fail to consider the lexical and syntactic diversity that preserves meaning [39]. This is because they focus on the n-gram overlap between the generated text and the reference text. Thus, explanations with similar meaning but different phrasing can receive low scores. On the other hand, BERTScore [39] is highly correlated with human evaluations and computes token similarity using contextual embeddings [39]. Table 19 shows the experiment results. According to the BERTScore,

the model achieves an F1-score of 0.8916, indicating that the generated explanations are semantically similar to the reference explanations, even if they differ in the exact words, which is justified by the high BERTScore and the low scores of BLUE and ROUGE.

Results for fine-tuned model (prompt 1)			
Metric	**Value**		
Accuracy	**0.833**		
F1	0.836		
Precision	0.846		
Recall	0.833		
Balanced Acc.	0.831		
Metrics per class			
Metric	**Entailment**	**Neutral**	**Contradiction**
F1	0.864	0.737	0.876
Precision	0.858	0.679	0.947
Recall	0.870	0.807	0.816
Predictions	5289 (35.88%)	4521 (30.67%)	4930 (33.45%)

Table 18: Results for the fine-tuned model with prompt 1.

Metric	**Recall**	**Precision**	**F1 Score**
BLEU Score		0.0802	
ROUGE-1	0.3486	0.4182	0.3582
ROUGE-2	0.1393	0.1582	0.1380
ROUGE-L	0.3059	0.3648	0.3134
BERTScore	**0.8869**	**0.8968**	**0.8916**

Table 19: Results for the evaluation of explanation.

6.6 Answering the Research Questions

After our experimentation, we can answer the questions we set at the beginning of the Experiments section.

- *How does Llama 3.2 Vision perform on the visual entailment task in a zero-shot inference?*
  In the zero-shot inference (baseline models), the performance of the model is limited, achieving an accuracy of 44.5% and 41.3% depending on the prompt and specifically on the class label order. The model showed a tendency to overpredict the entailment class.

- *What is the impact on zero-shot inference of having an incomplete or absent dataset?*
  The two experiments conducted (cropped images and black images) revealed a drop in performance. Both experiments struggle to correctly classify mainly the contradiction class. Also, in the experiment with black images, the model hallucinates to entail the hypothesis, and thus it overpredicts the entailment class.

- *What is the impact of few-shot inference on the model's accuracy, and how does performance vary with different numbers of examples?*
  The best few-shot results improve the performance (48.7%) over the zero-shot setting, and it was with three in-context examples. However, increasing to six-shot inference does not lead to further improvement, while the accuracy drops to 36.5% and the model becomes biased toward the neutral class. This indicates that additional in-context examples cannot guarantee improvement in the model's performance.

- *How does the order of class labels in the prompt affect model predictions?*
  The order of the class within the prompt influences the model's prediction, and it is clearly shown in the experiment, where each sample was evaluated on the six prompts, and we have a different accuracy for each.

- *What is the impact of varying the order of examples in few-shot inference on model performance?*

The order of the in-context examples significantly impacts the model's performance. When the contradiction class appears first, especially with Prompt 2 and random selection, the model performs best (48.7% accuracy). The importance of example ordering and content diversity in few-shot inference is further supported by the fact that random selection (diverse premises) produces better generalization than individual selection (same premise).

- *To what extent does fine-tuning improve model performance compared to zero-shot and few-shot inference?*
  Fine-tuning the model for the VE task yields a significant performance boost over both zero and few-shot settings. The model achieves an accuracy of 83.3% and the metrics per class also have a notable increase in all the classes. Compared to the OFA-X model, which achieves an accuracy of 80.9% in e-SNLI-VE, the fine-tuned Llama 3.2 Vision model surpasses it. Moreover, the model has a strong interpretability since it achieves an F1-score of 0.8916 using the BERTScore, an evaluation metric that utilizes contextual embeddings for the explanation evaluation. This indicates high semantic similarity between the reference and generated text.

7 Discussion

This section focuses on the key findings and the limitations of the project and directs future work. The experiments explored various configurations for zero-shot, few-shot, and fine-tuned inference, however, the project has some limitations. Moreover, observations and findings that come from this project can act as a strong starting point for further research, allowing for an improved understanding of multimodal inference in the future.

7.1 Discussion of key findings

This study investigates the capabilities of the Llama 3.2 Vision model on the VE task using the e-SNLI-VE dataset. The experiments yielded various

findings. First, the baseline results demonstrated modest performance, indicating the limited capabilities of the model in zero-shot inference. Three-shot inference improves the performance of the model, however, the study reveals that additional in-context examples are not always beneficial. The most significant finding is the major improvement after fine-tuning. In addition, the model's performance was reduced in the experiments with limited or absent vision, and the model was highly prone to hallucination. The experiments reveal that factors such as the order of the class labels in the prompt, the order of the in-context examples, and the examples selection strategy significantly affect the model's performance.

These findings offer several lessons for the broader MLLM research. In particular, the study underscores that while general pre-training is powerful, even advanced MLLMs such as Llama 3.2 Vision may not be suitable for complex reasoning tasks such as VE without special adaptation. The few-shot results underline that a deeper understanding of how models utilize context is needed. Additionally, the study highlights that the effectiveness of in-context learning depends on the number, quality, diversity, and ordering of examples. Also, the dramatic increase in performance after fine-tuning exposes that the model's visual and linguistic embeddings are highly adaptable.

7.2 Limitations

The findings provide helpful insights into the visual entailment capabilities of Llama 3.2 Vision, but there are some limitations that should be noted.

Firstly, every experiment was evaluated once because of time and computational constraints. The metrics are not averaged over multiple runs. Because of this, all results reported can be affected, and much more accurate estimates would result from repeated experiments.

Another limitation lies in the restricted experiments for the few-shot inference. A small number of configurations were tested, particularly for the six-shot inference, which included just one permutation. Several possible combinations are left out. Thus, a few of the combinations could yield more improved and even more stable results. However, given the issues found with

biases, sensitivity to order effects, and hallucinations, strong improvements for the right reasons are unlikely.

In addition, the fine-tuning was conducted using only the first prompt. It is not clear how performance could be differentiated with the second prompt. However, we expect that predictions will not be greatly affected by the order of the classes in the prompt, given the significant performance gain observed by the fine-tuning.

7.3 Future work

Based on this study, various potential paths can be explored for further research.

One approach would be to repeat each experiment multiple times and calculate the mean and standard deviation in order to assess the stability and robustness of the findings.

In addition, a worthwhile future work would be a further investigation with a few shots of inference. For example, exploring different sets of examples for each strategy in three shots, examining different orderings of classes for six shots, and testing a larger number of examples within a context, such as fifteen shots, could still be valuable, not primarily to focus on performance, but to gain deeper insights, such as understanding the threshold beyond which providing more examples becomes disadvantageous.

Another promising direction is to integrate **C**hain-**o**f-**T**hought (CoT) prompting into few-shot and zero-shot inference. This is an engineering technique that encourages the model to define complex tasks in a sequence of logical steps towards the final solution [34]. This technique has significantly improved many tasks, especially for reasoning [34].

A broader direction for future work includes systematic prompt engineering. This involves improving the wording and structure of the prompts. Since this study demonstrates that the design of the prompts significantly affects the predictions, optimizing the prompts could lead to better generalization

and fewer hallucinations.

Finally, another important extension would be to asses the VE task using larger multimodal models, while the Llama 3.2 Vision model with 11B parameters used for this study is a small-sized vision LLM. Evaluating larger models could reveal whether the increased number of parameters can improve the model's reasoning and robustness.

8 Conclusion

In conclusion, this thesis explored the capabilities of the Llama 3.2 Vision model regarding the VE task through the zero-shot, few-shot, and fine-tuning settings.

We conducted several experiments to evaluate the performance of the model on the VE task and investigate how it is affected by factors such as the order of class labels, the number and arrangement of in-context examples, and the completeness of visual information. Our findings revealed the fact that zero-shot inference has a moderate performance. In addition, the model suffers from class bias also label-ordering sensitivity in the prompt. Three-shot inference with careful example selection led to an improvement in the performance and reduction of biases. Nevertheless, the addition of in-context examples (six-shot setting) did not lead to an increase in performance; on the contrary introduced noise.

According to the experiments, it was revealed that when they involved creating explanations and limiting vision (cropped and black images) that the model can produce believable reasoning however, it is also highly prone to hallucinations. Fine-tuning the model increases the performance significantly and demonstrates the capabilities of the model when specifically tuned to a task. Moreover, the model has obtained strong interpretability, providing human-like explanations.

Overall, the study stresses the Llama 3.2 Vision model's strengths and limitations regarding the VE task. Future directions include the integration

of advanced prompt techniques, such as Chain-of-Thought prompting, and investigating the effect of scaling in larger multimodal models.

References

[1] Aishwarya Agrawal, Jiasen Lu, Stanislaw Antol, Margaret Mitchell, C. Lawrence Zitnick, Dhruv Batra, and Devi Parikh. Vqa: Visual question answering, 2016.

[2] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. Flamingo: a visual language model for few-shot learning, 2022.

[3] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy, 2017.

[4] Anna Bavaresco and Raquel Fernández. Experiential semantic information and brain alignment: Are multimodal models better than language models?, 2025.

[5] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020.

[6] Davide Caffagni, Federico Cocchi, Luca Barsellotti, Nicholas Moratelli, Sara Sarto, Lorenzo Baraldi, Lorenzo Baraldi, Marcella Cornia, and Rita Cucchiara. The revolution of multimodal large language models: A survey, 2024.

[7] Kilian Carolan, Laura Fennelly, and Alan F. Smeaton. A review of multi-modal large language and vision models, 2024.

[8] Aman Chadha and Vinija Jain. Few-shot multimodal multitask multilingual learning, 2023.

[9] Michael Han Daniel Han and Unsloth team. Unsloth, 2023.

[10] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized llms, 2023.

[11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.

[12] Hugging Face. What is llama 3.2 vision?, 2025. Accessed: 2025-02-25.

[13] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, and et al. The llama 3 herd of models, 2024.

[14] Sophia Gu, Christopher Clark, and Aniruddha Kembhavi. I can't believe there's no images! learning visual tasks using only language supervision, 2023.

[15] Muhammad Usman Hadi, Qasem Al-Tashi, Rizwan Qureshi, Abbas Shah, Amgad Muneer, Muhammad Irfan, Anas Zafar, Muhammad Shaikh, Naveed Akhtar, Jia Wu, and Seyedali Mirjalili. Large language models: A comprehensive survey of its applications, challenges, limitations, and future prospects, 07 2023.

[16] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021.

[17] Yizhang Jin, Jian Li, Yexin Liu, Tianjun Gu, Kai Wu, Zhengkai Jiang, Muyang He, Bo Zhao, Xin Tan, Zhenye Gan, Yabiao Wang, Chengjie Wang, and Lizhuang Ma. Efficient multimodal large language models: A survey, 2024.

[18] Maxime Kayser, Oana-Maria Camburu, Leonard Salewski, Cornelius Emde, Virginie Do, Zeynep Akata, and Thomas Lukasiewicz. e-vil: A dataset and benchmark for natural language explanations in vision-language tasks, 2021.

[19] Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. Foundations and trends in multimodal machine learning: Principles, challenges, and open questions, 2023.

[20] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics.

[21] Shervin Minaee, Tomas Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, and Jianfeng Gao. Large language models: A survey, 2024.

[22] Dimitri Coelho Mollo and Raphaël Millière. The vector grounding problem, 2023.

[23] Milad Moradi, Ke Yan, David Colwell, Matthias Samwald, and Rhona Asgari. Exploring the landscape of large language models: Foundations, techniques, and challenges, 2024.

[24] Ollama. Llama 3.2 vision model, 2025. Accessed: 2025-02-25.

[25] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, page 311–318, USA, 2002. Association for Computational Linguistics.

[26] Björn Plüster, Jakob Ambsdorf, Lukas Braach, Jae Hee Lee, and Stefan Wermter. Harnessing the power of multi-task pretraining for ground-truth level natural language explanations, 2023.

[27] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer, 2023.

[28] Eva Sánchez Salido, Julio Gonzalo, and Guillermo Marco. None of the others: a general technique to distinguish reasoning from memorization in multiple-choice llm evaluation benchmarks, 2025.

[29] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, and et al. Gemini: A family of highly capable multimodal models, 2024.

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

[31] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2023.

[32] Jiaqi Wang, Hanqi Jiang, Yiheng Liu, Chong Ma, Xu Zhang, Yi Pan, Mengyuan Liu, Peiran Gu, Sichen Xia, Wenjun Li, Yutong Zhang, Zihao Wu, Zhengliang Liu, Tianyang Zhong, Bao Ge, Tuo Zhang, Ning Qiang, Xintao Hu, Xi Jiang, Xin Zhang, Wei Zhang, Dinggang Shen, Tianming Liu, and Shu Zhang. A comprehensive review of multimodal large language models: Performance and challenges across different tasks, 2024.

[33] Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework, 2022.

[34] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models, 2023.

[35] Ning Xie, Farley Lai, Derek Doran, and Asim Kadav. Visual entailment task for visually-grounded language learning, 2019.

[36] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention, 2016.

[37] Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. The dawn of lmms: Preliminary explorations with gpt-4v(ision), 2023.

[38] Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. A survey on multimodal large language models. *National Science Review*, 11(12), November 2024.

[39] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert, 2020.