Probing Negation Understanding in Vision-Language Models With Object Attention

Anonymous ACL submission

Abstract

Negation is a fundamental aspect of language understanding and it is important vision-language models (VLMs) are able to process it effectively. Humans have shown to take longer to process negations, indicating the extra complexity of sentences containing negation. We investigate how VLMs process negation by analysing attention patterns of different models. We draw comparisons between where humans and models focus when processing affirmative vs negated statements. We compare object level model attention to human eye-fixations in both binary (2-object) and multary (4-object) settings, evaluating with both synthetic and realistic datasets. We find PaliGemma2 most closely mimics human patterns—in binary settings the negated alternative has most attention whereas in multary settings the mentioned object has more attention despite the negation. We further show a breakdown of object attention through layers, finding that PaliGemma2 uses later layers for processing negated prompts and that LLaVA-1.5 displays a bias towards left-side objects. Our evaluation of models on downstream tasks, show that while InternVL3.5 achieves the highest accuracy, strong performance doesn't necessarily correlate with human-like attention patterns.

1 Introduction

Despite their strong performance across various multi-modal tasks, vision-language models (VLMs) still struggle with negation. Investigating the ability of models to process negation is an emerging field with benchmarks to test negation accuracy being proposed (e.g. Bui et al. (2025); Zhang et al. (2025)) and explorations into how models such as CLIP handle negated captions (Alhamoud et al., 2025; Quantmeyer et al., 2024). Comparison with human performance on these benchmarks indicates that model performance is generally lower (Zhang et al., 2025), in some cases very substantially (Kamath et al., 2026).

At the same time, negation has been argued to be more difficult for humans to process (Wason, 1961a; Clark and Chase, 1972), and more recently, (Greco et al., 2020), with negated phrases taking a longer time to process than their affirmative counterparts. The cause of these longer reaction times is still under debate (Nieuwland and Kuperberg, 2008; Kaup et al., 2006). Given a negated sentence $\neg S$, e.g. *the door is not open*, the delay in processing such a sentence could be accounted for by needing to apply a symbolic negation operator to S, or by needing to simulate the alternative affirmative equivalent to $\neg S$, i.e. *the door is closed*.

One approach to understanding how human process negation is via the *visual world paradigm*, which presents verbal and visual inputs simultaneously. Orenes et al. (2014) use this paradigm to investigate whether affirmative alternatives are simulated in both binary and multary contexts, and find that while humans do indeed simulate the affirmative in binary contexts, this is not the case in multary contexts.

Since many of the errors that VLMs and generative models make in processing negation is to overgenerate or overpredict the affirmative S (Alhamoud et al., 2025; Zhang et al., 2025), it is important to understand whether VLMs exhibit the same behaviour as humans in processing negation, and where they differ.

We draw a parallel between attention to image tokens in VLMs and gaze fixation in humans. Using visual stimuli similar to those in Orenes et al. (2014), we analyse the attention values of generated output back to the input image for three VLMs: LLaVA-1.5 (LLaVA), InternVL3.5 (InternVL) and PaliGemma2. We create a synthetic dataset for evaluating models in Binary and Multary settings, similar to the human experiment setup in (Orenes et al., 2014). We also compare with attention values for realistic images more similar to the models' pre-training data by using images from the

What'sUp (Kamath et al., 2023) dataset.

Our main contributions are:

- We create two datasets for evaluating VLM negation processing in Binary (two object) and Multary (four object) settings. Additionally, we generate bounding box annotations for the What'sUp dataset to facilitate object-specific attention analysis.
- We evaluate object-level attention for 3 different VLMs, comparing how attention patterns for affirmative and negated prompts with human eye-fixation experiments. We find PaliGemma2 most closely aligns with human behaviour.
- We demonstrate that human-like attention processing does not necessarily correlate with strong downstream performance.

2 Background

Negation in Vision-Language Models VLMs have shown state-of-the-art performance in a wide range of tasks, however, recent studies show that they struggle with handling negation (Zhang et al., 2025; Alhamoud et al., 2025; Vatsa et al., 2025; Park et al., 2025; Singh et al., 2025). Many VLMs suffer from affirmation bias because they are pre-trained on large datasets of images labelled by humans who tend to focus on positive concepts in the image e.g. ("a photo of a dog") so models learn to associate the presence of words with the presence of objects (Ettinger, 2020). Negated phrases tend to appear in similar contexts to their affirmative counterparts, meaning that learned vectors of affirmative ('a dog') and negated ('not a dog') phrases can be extremely similar. This may be due to VLMs treating captions as a 'bag-of-words' and so telling the difference between 'the cow is eating the grass' and 'the cow is eating no grass' can be difficult (Yuksekgonul et al., 2022). Overall, **negation processing in VLMs is weak** and further understanding of these processes is important for the development of more reliable models.

Negation Processing in Humans Exactly how humans process negation is an open question, but it is agreed that negation is also harder for humans to process than affirmative statements (Greco, 2020; Orenes, 2021), evidenced by the fact that it takes longer for humans to process (Kaup et al., 2006; Clark and Chase, 1972; Greco et al., 2020; Wason, 1961b). One theory of negation processing is that negation is handled using a two-step process: first the affirmative version of the statement is simulated and then negated in the second step (Kaup et al., 2006). Other theories argue that negation is processed incrementally, more similar to how affirmative sentences are processed (Anderson et al., 2010; Tian et al., 2010).

The focus of our paper is on data collected by Orenes et al. (2014). In that paper, an eye-tracking experiment is performed to investigate how negation is processed in humans. They use 2D images of coloured shapes with 4 shapes in an image. Each participant is presented with one of these images and 2 spoken sentences. In the binary experiment, the sentence mentions two possible colours eg. "The figure could be red or blue". The second sentence is then either affirmative: "The figure was blue" or negative: "The figure was not blue". The multary experiment considers four colours: "The figure could be red, green, blue or yellow". In the binary case, the human's gaze is initially on the mentioned colour and shifts after a short time delay to the unmentioned colour as the negation is processed. In the multary case however, the gaze stays focused on the mentioned colour. These results are argued to provide evidence that in humans, in the binary case, negation is processed by simulating the affirmative alternative (i.e., "the figure was red"). This is inferred from the focus on the non-blue object. On the other hand, in the multary case, it is argued that negation may be represented symbolically. This is inferred from the focus on the blue object, with the assumption that some kind of mental negation operator is applied to a mental representation of the blue object. Since VLMs suffer from an affirmation bias, **it is important to understand what kinds of negation processing may be occurring.**

Attention Visualisation Attention visualisation has been used as an interpretability tool for VLMs to see which part of an image models are "looking" at when processing a specific word or query (Ben Melech Stan et al., 2024; Aflalo et al., 2022). Yu and Ananiadou (2024) use attention maps in addition to log probability to help explain hallucinations in VLMs and Bi et al. (2025) visualise attention weights and concentration scores to probe how models initially trained on linguistic data process visual content. Attention visualisation has also been used to diagnose failures in negation interpre-

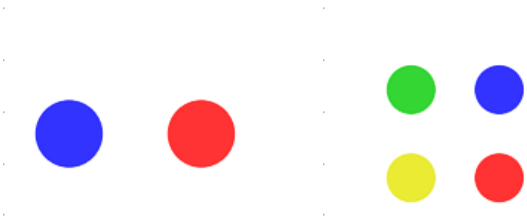
tation. Investigations into how CLIP processes negation find that there is a lack of separation between affirmative and negated captions (Alhamoud et al., 2025) and that CLIP processes negation using specific heads at early and late layers (Quantmeyer et al., 2024). While the exact role of attention in text and image processing is still open to debate (Xing et al., 2025), evidence suggests that the information for an object is localised to the token positions corresponding to its original location in the image, that attention flow from these tokens in late to middle layers carries important object information to the final token of an input, and that successful spatial reasoning correlates with the alignment of attention to tokens corresponding to object locations in images (Neo et al., 2025; Chen et al., 2025). These findings indicate that **attention is a suitable proxy for human gaze fixations** in the visual world paradigm.

3 Datasets

In this section, we provide the details for three datasets that we use for investigating models' abilities to process negation. We choose datasets to evaluate differing levels of realism, and create a "simple" abstract dataset for initial exploration in addition to selecting the What'sUp dataset (Kamath et al., 2023), which includes real-world images of objects in controlled settings.

Binary We generate images containing a plain white background and two 2D objects. We consider the colours blue, red, green and yellow and the shapes circle, square, triangle and diamond. Evaluating the models with these constraints ensures that this information will be in-domain for the vision-language models and aligns our experiments with the human eye-tracking experiments in Orenes et al. (2014). Objects within an image are of different colours but always the same shape and size. Between images, objects may be one of three sizes: small, medium or large. We choose values 0.12, 0.18, and 0.25 as the size values which act as shape diameter multipliers. For instance, a large circle will have a diameter of exactly 25% of the image width. Positions of the objects, i.e. the coordinates of the centres of the shapes are fixed. Object sizes and colours are randomly assigned and we generate 1000 images in total. This ensures there are at least 100 images containing each possible colour-shape combination allowing sufficient evaluation of the models in all cases.

Multary Additionally, we generate images with four 2D objects on a plain white background. We consider the same colours and shapes as in the Binary dataset however we choose different size parameters 0.10, 0.14 and 0.18 for small, medium and large respectively. This maintains a consistent proportion of white background in the images between the Binary and Multary datasets. Again we generate 1000 images with fixed shape positions. Each image contains one object of each colour. Figure 1 shows an example from the Binary and Multary datasets.



(a) (b)

Figure 1: Example images from the (a) Binary dataset and (b) Multary dataset.

What'sUp (Kamath et al., 2023) We choose the What'sUp dataset as a realistic set of images to test whether the attention patterns for synthetic environments generalise to more complex real-world images. Given VLMs are primarily pre-trained on real-world datasets, attention mechanisms may be more optimised for realistic images. What'sUp contains images of two objects in four positions, for example 'book left of a table', 'book on a table', 'book under a table' and 'book right of a table'. We exclude the relations 'on' and 'under' to avoid instances where objects occlude each other. Furthermore, this makes our attention analysis directly comparable with the horizontal constraint of the Binary dataset. After we have applied the above filtering, our dataset uses 206 What'sUp images (out of 408), examples of which are shown in Figure 2.

(a) book left of table (b) lamp right of chair

Figure 2: What'sUp example images

To analyse attention to objects within an image, we use bounding boxes to map image tokens to corresponding objects. The Binary and Multary datasets contain bounding box annotations from our dataset generation. However, since What'sUp dataset does not include bounding boxes, we set up a pipeline using Grounding DINO (Liu et al., 2023). Specifically, we extract the entities of two target objects from each image using the objects within the ground truth label. Grounding DINO returns a list of candidate detections from which we selected those with the highest confidence and performed a qualitative evaluation finding these bounding boxes consistently aligned with object locations within images. Therefore, we select the candidate with the highest confidence score for each label to serve as our final bounding box annotations.

4 Method

We analyse attention to shapes in visual stimuli similar to those in Figure 1, presented together with statements that are either affirmative (e.g. "the shape is blue") or negated (e.g. "the shape is *not* blue). We use the following terminology:

- **Mentioned** means that the shape was mentioned in an affirmative sentence, e.g. a blue shape given the sentence "The figure is blue".
- **Alternative** means that the shape was not mentioned in an affirmative sentence, e.g. a red shape given "The figure is blue".
- **Mentioned Negated** means that the shape was mentioned in a negated sentence, e.g. a blue shape given "The figure is *not* blue".
- **Alternative Negated** means that the shape was not mentioned in a negated sentence, e.g. a red shape given "The figure is *not* blue".

Note how we use blue to indicate a *mentioned* shape, red to indicate an *alternative* shape, and underlining to indicate that the sentence is negated. This formatting is for clarity of presentation and was not present in the original human experiments.

Binary Inspired the binary experiment in Orenes et al. (2014), we compare attention patterns to the two objects in the image in affirmative ("The figure is blue") and negated ("The figure is not blue") contexts. If following a similar pattern to humans, we expect the following pattern of attention:

Mentioned > **Alternative**

Alternative Negated > **Mentioned Negated**

To measure this, we compare the attention patterns for 3 models: LLaVA-1.5-7b (Liu et al., 2024), InternVL-3.5-4B (Wang et al., 2025) and PaliGemma2-3B-mix-224 (Steiner et al., 2024). We selected these models based on their differences in model architectures and range of vision encoders and LLM backbones: LLaVA-1.5 (CLIP-ViT/Vicuna via MLP), InternVL3.5 (InternViT/Qwen3), and PaliGemma2 (SigLIP/Gemma-2). LLaVA processes images into 576 vision tokens, InternVL dynamically varies the number of image tokens based on input image size however, the number of visual tokens for InternVL is always 256 for images in our experiments, matching the fixed number of tokens for PaliGemma2. Table 1 shows a comparison of layer and head configurations.

Model	Layers	Attn. Heads	Visual Tokens
LLaVA-1.5-7B	32	32	576
InternVL3.5-4B	36	32	256
PaliGemma2-3B	26	8	256

Table 1: Configurations of VLMs

We analyse the cross-modal attention by capturing the weights assigned to image tokens during the generation of the output, which we limit to a single-token response. Keeping consistent with the experimental design of Orenes et al. (2014), we use prompts that specify the colour of one of the image objects in either an affirmative or negated format e.g. "The figure is {colour}" or "The figure is not {colour}", where {colour} is replaced by a colour within the image.

To analyse how different models process affirmative and negated inputs, we calculate normalised attention scores for each object within an image. For layer-wise attention we consider attention averaged over all heads.

We identify the image tokens for each model based on the relevant marker for each model (e.g. $< image >$ for LLaVA), and re-construct the 1D image token sequence to their original 2D coordinates. This results in an $n \times n$ grid of image attention weights for each transformer layer. To determine which image tokens correspond to specific objects, we project the bounding box data onto this grid. To map a coordinate from the image space to

the heatmap space, we use a scaling factor derived from the ratio of their dimensions. Let the original image dimensions be (W_{img}, H_{img}) and the attention map dimensions be (W_{attn}, H_{attn}). The scaling factors are then defined as:

$$S_x = \frac{W_{attn}}{W_{img}}, \quad S_y = \frac{H_{attn}}{H_{img}}$$

For any point (x, y) in the original image, the transformed coordinates (x', y') in the heatmap are:

$$x' = \lfloor x \cdot S_x \rfloor, \quad y' = \lfloor y \cdot S_y \rfloor \tag{1}$$

We calculate normalised attention by summing attention within the bounding box and dividing by the number of tokens within the bounding box:

$$Score = \frac{\sum_{j=y'_1}^{y'_2} \sum_{i=x'_1}^{x'_2} \text{attn}(i,j)}{(x'_2 - x'_1 + 1) \times (y'_2 - y'_1 + 1)} \tag{2}$$

We report both layer-wise attention scores averaged across heads and mean attention for all layers and heads.

Multary We repeat the Binary experiment with images containing 4 objects and compare with the multary experiment from Orenes et al. (2014). In humans presented with a negated sentence such as "The figure is not blue", rather than fixate on the alternative three shapes, fixations remain on the shape with the mentioned colour despite the negation. We therefore expect to see a reversed attention pattern for negated statements–i.e., we expect attention to the mentioned object to be higher than to the alternative. That is, we expect to see:

Mentioned > **Alternative**

Mentioned Negated > **Alternative Negated**

We calculate attention to the four objects in the image using the same method as in the Binary experiment. Mentioned attention is attention to the colour stated in the prompt, alternative attention is the mean of the attention to the other three objects.

What's Up Finally, we repeat the Binary experiment using more realistic images. Since the models are pre-trained mainly on real-world data, they should have more familiarity with the objects in these images compared with the synthetic 2D stimuli. We use the same calculations as in the Binary experiment except prompts now take the form "The object is a {noun}" where {noun} is replaced with one of the two objects in the image. Since this dataset has a similar structure to the Binary dataset, the behaviour we expect to see is that of the Binary dataset.

4.1 Downstream Tasks

To test the relationship between attention patterns and downstream performance we evaluate model performance on all three datasets. We ask the models a binary question where the answer is either 'Yes' or 'No'. For each image, we consider four question categories: affirmative-true (A.T.): an affirmative question with the correct answer Yes; affirmative-false (A.F.): an affirmative question with the correct answer No; negated-true (N.T.): a negated question with the correct answer Yes; and negated-false (N.F.): a negated question with the correct answer No. For example, the Binary example in Figure 1a may have questions 'Is the left shape blue?', 'Is the left shape red?', 'Is the left shape not red?', and 'Is the left shape not blue?'. We repeat each binary question 3 times with different prompt variations and report the average. We use Gemini 3 Flash (Google DeepMind, 2025) to generate prompt variations based on prompt template 1 (P1). We use the following prompt: "{P1} can you generate differently worded variations of this prompt template" and choose two alternatives from the produced list. Questions for the Binary dataset use the following templates:

1. Is this statement correct? The object on the {position} is {colour}. Answer yes or no.
2. Is the object on the {position} {colour}? Answer yes or no.
3. Does this statement match the image? The object on the {position} is {colour}. Answer yes or no.

The position can either be 'left' or 'right' which is chosen at random for each image and colour is chosen based on colour of objects within the image. For each prompt template we report mean and standard deviation over 3 seeds for each model. For the Multary dataset, we use the same prompt templates, however positions can be either 'top left', 'top right', 'bottom left' or 'bottom right', for example: 'Is this statement correct? The object at the top right is green'. Questions in the 'What'sUp' experiments have the same positions as the Binary dataset, however instead of colour we ask the model

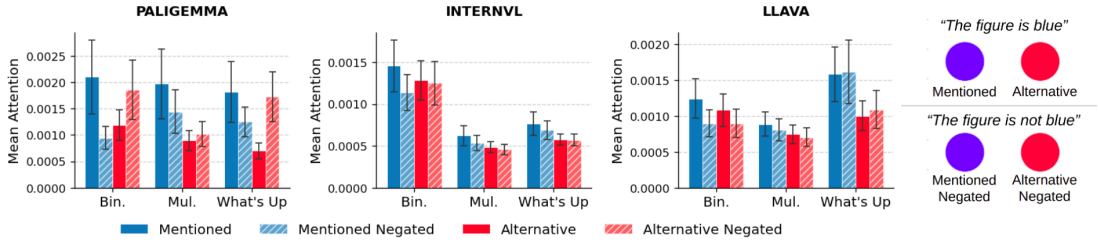


Figure 3: Mean attention to objects for models across datasets. For a human-like response we expect Mentioned > Alternative throughout, Alternative Negated > Mentioned Negated for the **Binary** and **What'sUp** datasets, and Mentioned Negated > Alternative Negated for the **Multary** dataset. Error bars represent the 95% confidence interval.

for the identity of the object. For example: "Is the object on the left a chair?" or "Is the object on the left not a chair?". A full list of prompt templates used for Multary and What'sUp datasets is provided in Appendix A.3.

5 Results

5.1 High Level Results

Figure 3 shows the mean object attention for each dataset and model. Only PaliGemma2 shows the human-like pattern we outlined in section 4 across all datasets. While InternVL exhibits a similar trend on Binary and Multary, the distinction is not statistically significant given the confidence intervals, and we do not see the expected pattern on What'sUp. Interestingly, PaliGemma2 is the smallest of models tested, so it is possible that the limited number of model parameters forces a more human-like representation of more complex concepts such as negation. InternVL has more overall attention for the Binary dataset, suggesting that reduced visual complexity allows for more concentrated object attention. LLaVA however, has increased attention to the What'sUp dataset, with most attention allocated to realistic mentioned objects. In contrast, PaliGemma2 has a more uniform attention showing the least differentiation between the synthetic (Binary) and realistic (What'sUp) 2-object settings suggesting PaliGemma2 is more robust to variations in visual input. Given PaliGemma2's most human-like response to negation across all datasets, we might expect it to have the strongest downstream performance on negated questions.

5.2 Layer-Wise Attention

We now look in more detail at the attention to objects across different layers. Layer-wise attention to objects in the What'sUp dataset is shown in Figure 4. Layer-wise attention for the Binary and Multary datasets are provided in Appendices A.1 and A.2. We see that in PaliGemma2 for negated prompts, negation appears to be processed at layer 12, with a clear shift in attention from the mentioned object to the alternative. LLaVA and InternVL consistently allocate more attention to the mentioned shape than the alternative at all layers, even in binary negated contexts. This absence of a layer-wise shift toward the negated target may suggest that these models do not process negation via the simulation of an affirmative alternative as humans do. Potentially, negation is processed by applying some symbolic operator outside of the attention calculation.

5.3 Left-Right Bias

Attention patterns can be influenced by many factors. We analysed attention patterns across colour, shape, and spatial position. For most factors, with the exception of spatial position, we see similar behaviour across categories (data not shown). To analyse the influence of spatial position, we plot the layer-wise attention on the Binary dataset, categorising results by the position (left or right) of the mentioned object, shown in Figure 5. The two leftmost graphs correspond to the left object being mentioned, while the two rightmost graphs correspond to the right object being mentioned. We observe a bias in LLaVA where the attention scores for the left object are consistently higher, especially at layers 20 to 25, regardless of which object is mentioned or whether the prompt is negated. When the right object is mentioned, attention to the right is object is higher than when the left shape is mentioned showing the prompt modulates LLaVA's focus to a certain extent. However, the left shape still has more attention overall suggesting LLaVA prioritises left-hand side visual tokens when generating its output.

InternVL shows a more even distribution of at-

tention to the left and right objects. The attention patterns for affirmative left are most similar to the negated right and vice-versa. This suggests InternVL may be processing the negated prompts by converting them into their affirmative alternative, but that this is not happening in the attention weights. For example, given the 2-shape image in Figure 1, "The figure is not blue" may be interpreted as "The figure is red". Comparing with the general attention pattern illustrated in Figure 3, this suggests that the negation is not being actively computed in the attention weights, as we see in PaliGemma2, but may be calculated elsewhere in the model and causing attention to follow the pattern of the affirmative alternative.

In PaliGemma2, we see that attention patterns for affirmative left are most similar to affirmative right and similarly for the negated left and right graphs, suggesting PaliGemma2 has the most symmetrical visual processing of the models.

5.4 Downstream Tasks

Table 3 shows the accuracy of models across four different categories of questions: affirmative-true (A.T.), negated-true (N.T.), affirmative-false (A.F.) and negated-false (N.F.). Each category type is run with 3 different seeds and across 3 different prompt templates. A breakdown of accuracies and standard deviations by prompt templates is provided in Appendix A.4.

Downstream performance highlights a significant gap in performance between affirmative and negated prompts. Overall accuracies for LLaVA, InternVL and PaliGemma2 using affirmative prompts are 70.2%, 92.5% and 86.4% respectively. In comparison the overall accuracies for prompts with negation are 56.9%, 88.3% and 73.8%, highlighting the additional complexity of questions containing negation. In humans, the difficulty of negation is operationalised through processing time, humans may take longer to process negations however their overall error rate on simple attributes such as colour should be low. If VLMs had robust reasoning we would expect a high accuracy for both affirmative and negated questions. The drop in model performance on negated questions underscores the importance of understanding negation processing in VLMs and highlights the difficulty models have with handling negation.

InternVL has the highest average accuracy on both affirmative and negated prompts across datasets despite not showing clear shifts in attention patterns or specific layers for processing negative prompts. PaliGemma2 has the most human-like attention patterns however this doesn't translate into a stronger performance on downstream tasks. None of the models achieve a perfect performance on all question categories even on the Binary dataset showing even simple negations can be problematic for models, highlighting VLMs need more robust training on negated concepts. All models have multiple layers where attention to all objects in the image is very low or similar attention is given to both shapes. Perhaps assigning more attention to key image tokens would result in better comprehension of spatial reasoning questions from the models. LLaVA's left-side bias may be contributing to the fact it has the lowest accuracy on all dataset.

6 Conclusion

We investigate negation understanding in VLMs by quantifying attention allocated to objects within images. We design experimental frameworks inspired by Orenes et al. (2014) and compare our findings with human eye-gaze fixations patterns. In addition, we extend the What'sUp dataset (Kamath et al., 2023) using object segmentation to provide bounding boxes, facilitating object-level attention analysis on real-world imagery. We evaluate three models PaliGemma2, InternVL3.5 and LLaVA finding that while PaliGemma2 has overall attention to mentioned and alternative objects most similar to human eye-fixations, this doesn't translate to a better downstream performance. Instead we find InternVL3.5 has the highest accuracy on negated questions despite not showing human-like attention patterns. We theorise that this may be due to the model processing negation in a more symbolic manner, relying on components beyond attention. PaliGemma2 is the smallest of models in our comparison and the only model to display distinct layers for negation processing, suggesting human-like attention processing may emerge as a necessity for models with limited parameters. We additionally find that LLaVA has a bias to left-side shapes which may be contributing to it having the lowest accuracy on downstream tasks. All models show a drop in downstream accuracy for negated prompts, underscoring the fundamental gap in VLM negation processing. Understanding the underlying mechanisms of negation processing is crucial for developing more robust negation-aware architectures for future models.

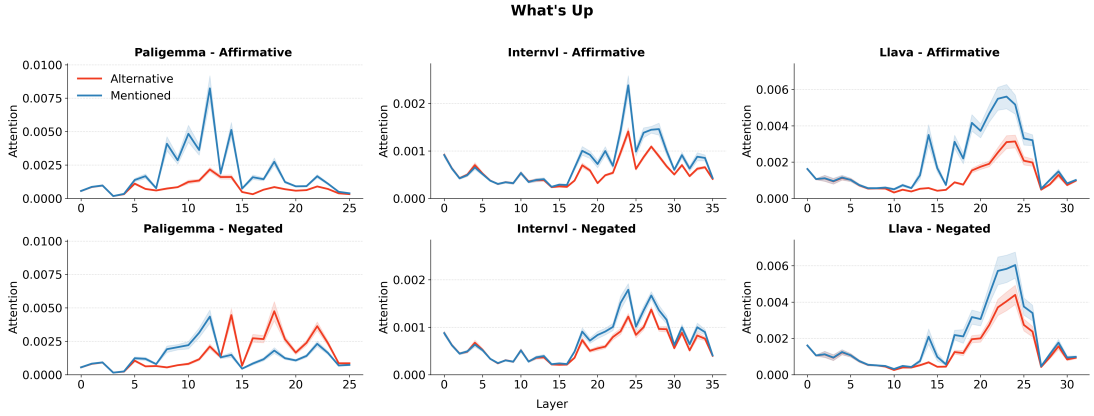


Figure 4: Layer-wise attention for What'sUp dataset

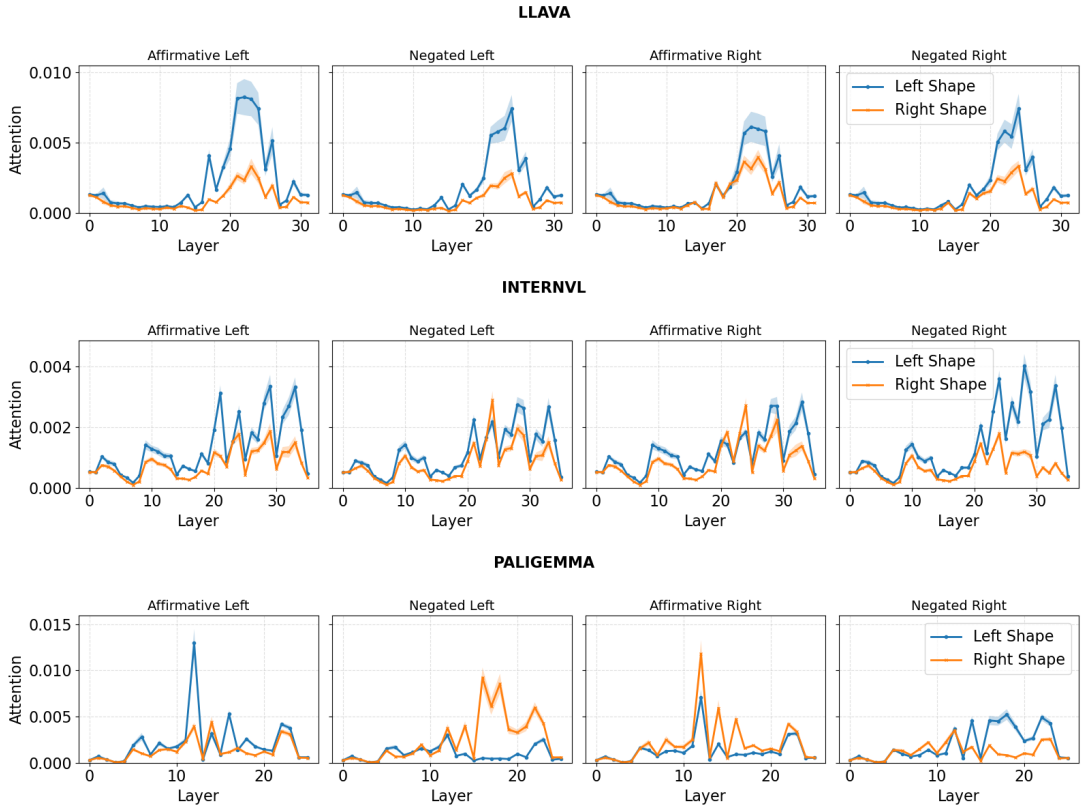


Figure 5: Layer-wise attention patterns for left-shape vs. right-shape colour mentions in both positive and negated contexts.

	Binary				Multary				What'sUp			
Model	**A.T**	**N.T**	**A.F.**	**N.F.**	**A.T**	**N.T**	**A.F.**	**N.F.**	**A.T**	**N.T.**	**A.F.**	**N.F.**
LLaVA	**100.0**	24.2	70.8	99.3	89.7	52.56	36.89	66.22	87.2	3.0	36.8	96.2
InternVL	**100.0**	**100.0**	**100.0**	**99.8**	87.56	**68.7**	**99.89**	**99.11**	79.33	**80.22**	**88.22**	82.11
PGemma	**100.0**	67.2	**100.0**	98.6	**100.0**	60.56	70.67	96.0	**91.89**	21.0	56.0	**99.33**

Table 2: Downstream Accuracy For Models On All Datasets. We report mean accuracy for four question types: affirmative-true (A.T.), negative-true (N.T.), affirmative-false (A.F.) and negative-false (N.F.)

Limitations

We acknowledge the following limitations in our work. Our analysis is limited to three models, while selected for their differences in model architectures, our experiments could be extended to a wider range of models in future work. Additionally, our experiments use simplistic images, even the more naturalistic What'sUp datasets contains only two objects under positional constraints. Exploring how models negation handling changes when faced with more complex images would be an interesting avenue for future research. We focus our analysis on attention to image tokens as a proxy for human eye-fixations. However, exploring other mechanisms for example the residual stream may reveal more insights into how the models process negation. Finally, we look only at negation in simple settings and in English. Exploring other languages is an area for future work.

Acknowledgments

References

Estelle Aflalo, Meng Du, Shao-Yen Tseng, Yongfei Liu, Chenfei Wu, Nan Duan, and Vasudev Lal. 2022. Vl-interpret: An interactive visualization tool for interpreting vision-language transformers. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pages 21406–21415.

Kumail Alhamoud, Shaden Alshammari, Yonglong Tian, Guohao Li, Philip HS Torr, Yoon Kim, and Marzyeh Ghassemi. 2025. Vision-language models do not understand negation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 29612–29622.

SARAH E Anderson, Stephanie Huette, Teenie Matlock, and MJ Spivey. 2010. On the temporal dynamics of negated perceptual simulations. *Meaning, form, & body*, pages 1–20.

Gabriela Ben Melech Stan, Estelle Aflalo, Raanan Yehezkel Rohekar, Anahita Bhiwandiwalla, Shao-Yen Tseng, Matthew Lyle Olson, Yaniv Gurwicz, Chenfei Wu, Nan Duan, and Vasudev Lal. 2024. Lvlm-intrepret: An interpretability tool for large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8182–8187.

Jing Bi, Junjia Guo, Yunlong Tang, Lianggong Bruce Wen, Zhang Liu, Bingjie Wang, and Chenliang Xu. 2025. Unveiling visual perception in language models: An attention head analysis approach. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 4135–4144.

Nhat-Tan Bui, Dinh-Hieu Hoang, Quoc-Huy Trinh, Minh-Triet Tran, Truong Nguyen, and Susan Gauch. 2025. NeIn: Telling What You Don't Want. *arXiv preprint*. ArXiv:2409.06481 [cs].

Shiqi Chen, Tongyao Zhu, Ruochen Zhou, Jinghan Zhang, Siyang Gao, Juan Carlos Niebles, Mor Geva, Junxian He, Jiajun Wu, and Manling Li. 2025. Why Is Spatial Reasoning Hard for VLMs? An Attention Mechanism Perspective on Focus Areas.

Herbert H. Clark and William G. Chase. 1972. On the process of comparing sentences against pictures. *Cognitive Psychology*, 3(3):472–517.

Allyson Ettinger. 2020. What bert is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *Transactions of the Association for Computational Linguistics*, 8:34–48.

Google DeepMind. 2025. Gemini 3 flash. `https://deepmind.google/models`. Model card.

Matteo Greco. 2020. On the syntax of surprise negation sentences: A case study on expletive negation. *Natural Language & Linguistic Theory*, 38(3):775–825.

Matteo Greco, Paolo Canal, Valentina Bambini, and Andrea Moro. 2020. Modulating "surprise" with syntax: A study on negative sentences and eye-movement recording. *Journal of Psycholinguistic Research*, 49(3):415–434.

Amita Kamath, Jack Hessel, Khyathi Chandu, Jena D. Hwang, Kai-Wei Chang, and Ranjay Krishna. 2026. Scale Can't Overcome Pragmatics: The Impact of Reporting Bias on Vision-Language Reasoning. *arXiv preprint*. ArXiv:2602.23351 [cs] version: 1.

Amita Kamath, Jack Hessel, and Kai-Wei Chang. 2023. What's "up" with vision-language models? Investigating their struggle with spatial reasoning. *arXiv preprint*. ArXiv:2310.19785 [cs].

Barbara Kaup, Jana Lüdtke, and Rolf A Zwaan. 2006. Processing negated sentences with contradictory predicates: Is a door that is not open mentally closed? *Journal of Pragmatics*, 38(7):1033–1050.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 26296–26306.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. 2023. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *Preprint*, arXiv:2303.05499.

Clement Neo, Luke Ong, Philip Torr, Mor Geva, David Krueger, and Fazl Barez. 2025. Towards Interpreting Visual Information Processing in Vision-Language Models. *arXiv preprint*. ArXiv:2410.07149 [cs].

Mante S Nieuwland and Gina R Kuperberg. 2008. When the truth is not too hard to handle: An event-related potential study on the pragmatics of negation. *Psychological Science*, 19(12):1213–1218.

Isabel Orenes. 2021. "looking at" negation: Faster processing for symbolic rather than iconic representations. *Journal of Psycholinguistic Research*, 50(6):1417–1436.

Isabel Orenes, David Beltrán, and Carlos Santamaría. 2014. How negation is understood: Evidence from the visual world paradigm. *Journal of Memory and Language*, 74:36–45.

Junsung Park, Jungbeom Lee, Jongyoon Song, Sangwon Yu, Dahuin Jung, and Sungroh Yoon. 2025. Know" no" better: A data-driven approach for enhancing negation awareness in clip. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2825–2835.

Vincent Quantmeyer, Pablo Mosteiro, and Albert Gatt. 2024. How and where does clip process negation? In *Proceedings of the 3rd Workshop on Advances in Language and Vision Research (ALVR)*, pages 59–72.

Jaisidh Singh, Ishaan Shrivastava, Mayank Vatsa, Richa Singh, and Aparna Bharati. 2025. Learning the power of "no": Foundation models with negations. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 8002–8012. IEEE.

Andreas Steiner, André Susano Pinto, Michael Tschannen, Daniel Keysers, Xiao Wang, Yonatan Bitton, Alexey Gritsenko, Matthias Minderer, Anthony Sherbondy, Shangbang Long, Siyang Qin, R. Reeve Ingle, Emanuele Bugliarello, Sahar Kazemzadeh, Thomas Mesnard, Ibrahim M. Alabdulmohsin, Lucas Beyer, and Xiao-Qi Zhai. 2024. Paligemma 2: A family of versatile vlms for transfer. *ArXiv*, abs/2412.03555.

Ye Tian, Richard Breheny, and Heather J Ferguson. 2010. Why we simulate negated information: A dynamic pragmatic account.

Mayank Vatsa, Aparna Bharati, Surbhi Mittal, and Richa Singh. 2025. From no to know: Taxonomy, challenges, and opportunities for negation understanding in multimodal foundation models. *arXiv preprint arXiv:2502.09645*.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, and 1 others. 2025. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*.

P. C. Wason. 1961a. Response to affirmative and negative binary statements. *British Journal of Psychology*, 52(2):133–142. Num Pages: 10.

Peter C Wason. 1961b. Response to affirmative and negative binary statements. *British Journal of Psychology*, 52(2):133–142.

Xiaoying Xing, Chia-Wen Kuo, Li Fuxin, Yulei Niu, Fan Chen, Ming Li, Ying Wu, Longyin Wen, and Sijie Zhu. 2025. Where do large vision-language models look at when answering questions? *arXiv preprint arXiv:2503.13891*.

Zeping Yu and Sophia Ananiadou. 2024. Understanding multimodal llms: the mechanistic interpretability of llava in visual question answering. *arXiv preprint arXiv:2411.10950*.

Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. 2022. When and why vision-language models behave like bags-of-words, and what to do about it? In *The Eleventh International Conference on Learning Representations*.

Yuhui Zhang, Yuchang Su, Yiming Liu, and Serena Yeung-Levy. 2025. Negvqa: Can vision language models understand negation? In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 3707–3716.

A Appendix

A.1 Binary Results

We visualise the layer-wise attention distribution for all images in the binary dataset. Figure 6 shows the layer-wise attention for models in the positive and negative settings. All models have similar attention to both the mentioned and unmentioned colours in the first set of layers, up to 10 for LLaVA, 20 for InternVL and 5 for PaliGemma2. This suggests earlier layers may be focused more on the image as a whole and processing low-level features rather than filtering specific information based on the text prompt. InternVL, while having similar attention values throughout most layers, does display slight overall shift to having more attention on the alternative object in the negated case indicating some acknowledgment of 'not green' meaning focus on the blue shape for example. Paligemma2 shows similar patterns to in the What'sUp datset, with attention to the alternative negated object peaking at later layers than attention to the mentioned affirmative, again suggesting later layers may be responsible for high level processing required for negation.

A.2 Multary

We report average attention with 95% confidence intervals at each layer for the three models LLaVA, InternVL and Paligemma2 in Figure 7. Attention to the not mentioned colour is reported as the average attention between the three alternative shapes.

LLaVA attention patterns are similar to the Binary dataset where LLaVA exhibits a minimal

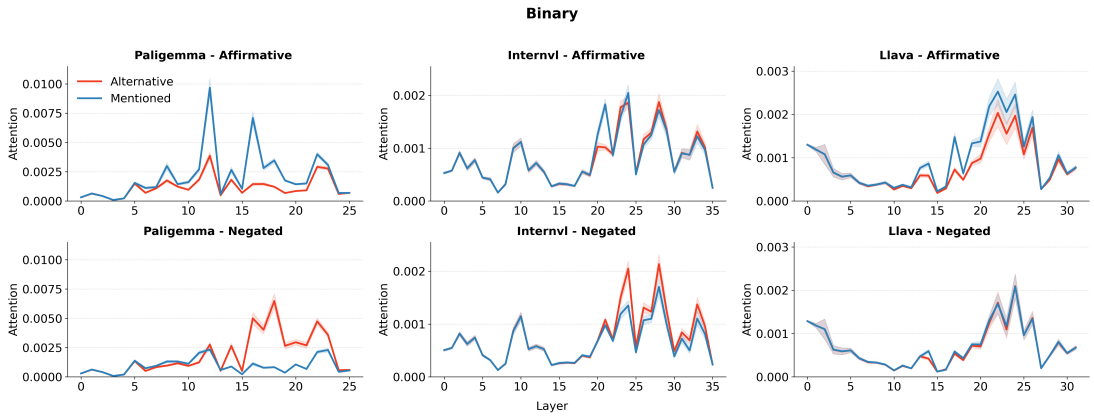


Figure 6: Layer-wise attention for 2D dataset 2 shapes

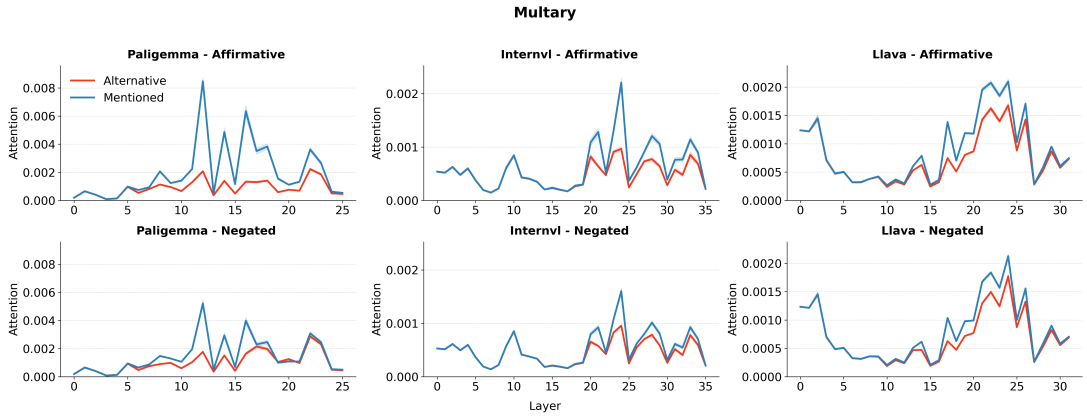


Figure 7: Layer-wise attention for models in the Multary setting

shift in attention between the positive and negative prompts. LLaVA has overall more attention to the mentioned object in both affirmative and negated settings with most attention going to both objects between layers 20 and 25. InternVL's attention to the alternative object is fairly similar for both affirmative and negated prompts however attention to the mentioned object is clearly reduced especially in layer 24. PaliGemma2 again shows an increase in alternative negated attention at later layers however attention to the mentioned object at these layers are higher. This indicates that the model acknowledges the negated object but does not fully shift its primary focus.

A.3 Downstream Tasks Prompt Templates

Questions for the Multary dataset use the following templates:

1. Is this statement correct? The object at the {position} is {colour}. Answer yes or no.

2. Is the object at the {position} {colour}? Answer yes or no.

3. Does this statement match the image? The object at the {position} is {colour}. Answer yes or no.

Where the position is chosen randomly from 'top left', 'top right', 'bottom left', 'bottom right', and the colour is selected based on the object at the chosen position. False question colours are chosen at random from the three alternative colours.

Questions for the What'sUp dataset use the following templates:

1. Is this statement correct? The object on the {position} is a {noun}. Answer yes or no.

2. Is the object on the {position} a {noun}? Answer yes or no.

3. Does this statement match the image? The object on the {position} is a{noun}. Answer yes or no.

Where position can be either 'left' or 'right' chosen at random and noun is chosen based on objects in the image.

A.4 Downstream Tasks Results

The high variance between different prompt styles for models highlights the prompt sensitivity of these models. Models show the most variance on negated or positive false question types, for example PaliGemma2 accuracy on 2-Object negated questions varies from 100% to 27% and LLaVA shows a range of 0% to 84.7% on negated questions in the 4-Object dataset. In particular, LLaVA and PaliGemma2 exhibit more prompt sensitivity on negated question styles for the 2D datasets, maybe indicating that negation loweres the robustness of the models and makes following semantic cues more difficult. Generally, prompt sensitivity is overall lower for all models on the What's Up dataset than the 2D datasets indicating that the more familiar images may help stabilise prompt sensitivity.

		Binary				Multary				WhatsUp			
Model	**Pr.**	**A.T**	**N.T**	**A.F.**	**N.F.**	**A.T**	**N.T**	**A.F.**	**N.F.**	**A.T**	**N.T.**	**A.F.**	**N.F.**
LLaVA	1	$100.0^{0.0}$	$18.0^{2.16}$	$92.0^{1.63}$	$98.0^{1.41}$	$96.7^{0.94}$	$0.0^{0.0}$	$22.0^{2.94}$	$100.0^{0.0}$	$91.0^{1.63}$	$3.3^{0.47}$	$35.0^{1.63}$	$97.3^{0.47}$
	2	$100.0^{0.0}$	$54.7^{2.62}$	$41.0^{2.16}$	$100.0^{0.0}$	$100.0^{0.0}$	$73.0^{4.32}$	$13.3^{1.70}$	$53.3^{4.03}$	$82.7^{1.25}$	$5.0^{1.41}$	$35.7^{4.03}$	$92.0^{0.82}$
	3	$100.0^{0.0}$	$0.0^{0.0}$	$79.3^{1.70}$	$100.0^{0.0}$	$72.3^{2.49}$	$84.7^{1.25}$	$75.3^{0.47}$	$45.3^{2.05}$	$88.0^{1.63}$	$0.7^{0.47}$	$39.7^{1.70}$	$99.3^{0.94}$
	Avg	$\mathbf{100.0}^{0.0}$	$\mathbf{24.2}^{22.83}$	$\mathbf{70.8}^{2.76}$	$\mathbf{99.3}^{1.25}$	$\mathbf{89.7}^{12.43}$	$\mathbf{52.56}^{37.56}$	$\mathbf{36.89}^{27.49}$	$\mathbf{66.22}^{24.25}$	$\mathbf{87.2}^{3.76}$	$\mathbf{3.0}^{2.0}$	$\mathbf{36.8}^{3.39}$	$\mathbf{96.2}^{3.19}$
InternVL	1	$100.0^{0.0}$	$100.0^{0.0}$	$100.0^{0.0}$	$100.0^{0.0}$	$100.0^{0.0}$	$58.7^{4.99}$	$100.0^{0.0}$	$100.0^{0.0}$	$68.7^{0.94}$	$74.0^{0.82}$	$97.3^{0.94}$	$88.3^{2.05}$
	2	$100.0^{0.0}$	$100.0^{0.0}$	$100.0^{0.0}$	$99.3^{0.47}$	$89.0^{0.82}$	$79.0^{1.41}$	$100.0^{0.0}$	$97.7^{0.94}$	$95.3^{0.47}$	$76.3^{1.70}$	$69.3^{2.05}$	$82.3^{0.47}$
	3	$100.0^{0.0}$	$100.0^{0.0}$	$100.0^{0.0}$	$100.0^{0.0}$	$73.7^{1.25}$	$68.3^{3.77}$	$99.7^{0.47}$	$99.7^{0.47}$	$74.0^{0.82}$	$90.3^{2.05}$	$98.0^{0.0}$	$75.7^{2.05}$
	Avg	$\mathbf{100.0}^{0.0}$	$\mathbf{100.0}^{0.0}$	$\mathbf{100.0}^{0.0}$	$\mathbf{99.8}^{0.42}$	$\mathbf{87.56}^{10.83}$	$\mathbf{68.7}^{9.09}$	$\mathbf{99.89}^{0.31}$	$\mathbf{99.11}^{1.20}$	$\mathbf{79.33}^{11.55}$	$\mathbf{80.22}^{7.39}$	$\mathbf{88.22}^{13.42}$	$\mathbf{82.11}^{5.45}$
PGemma	1	$100.0^{0.0}$	$74.0^{1.41}$	$100.0^{0.0}$	$100.0^{0.0}$	$100.0^{0.0}$	$22.3^{3.40}$	$64.0^{4.32}$	$90.3^{0.47}$	96.7	23.3	55.3	98.7
	2	$100.0^{0.0}$	$100.0^{0.0}$	$100.0^{0.0}$	$95.7^{0.47}$	$100.0^{0.0}$	$69.0^{3.74}$	$83.3^{2.05}$	$99.7^{0.47}$	84.7	29.0	80.7	99.3
	3	$100.0^{0.0}$	$27.7^{3.09}$	$100.0^{0.0}$	$100.0^{0.0}$	$100.0^{0.0}$	$90.3^{1.89}$	$64.7^{5.31}$	$98.0^{0.82}$	$94.3^{0.94}$	$10.7^{1.25}$	$32.0^{4.90}$	$100.0^{0.0}$
	Avg	$\mathbf{100.0}^{0.0}$	$\mathbf{67.2}^{29.98}$	$\mathbf{100.0}^{0.0}$	$\mathbf{98.6}^{2.06}$	$\mathbf{100.0}^{0.0}$	$\mathbf{60.56}^{28.57}$	$\mathbf{70.67}^{9.87}$	$\mathbf{96.0}^{4.11}$	$\mathbf{91.89}^{5.40}$	$\mathbf{21.0}^{8.14}$	$\mathbf{56.0}^{20.12}$	$\mathbf{99.33}^{0.67}$

Table 3: Downstream Accuracy (Mean % $\pm$ SD) Across Datasets