Instructions for ACL 2023 Proceedings

Anonymous ACL submission

Abstract

1 Introduction

2 Methods

2.1 Models

We developed a probing framework for three compact vision–language models: **Qwen2-VL-2B**, **Gemma-3-4B-IT**, and **FastVLM-0.5**. All models were used in frozen form, i.e., without updating their parameters, so that only lightweight probing classifiers were trained on top of their internal representations. This allows us to isolate the representational capacity of the models at different layers without confounding effects from fine-tuning.

2.2 Probing Tasks

To investigate the distinction between global and local semantic representations, we designed two probing tasks. The *caption experiment* targets global features by testing whether the model can align an image with a candidate caption. For this task, inputs are constructed using the prompt `This image contains: {caption}. Is this right?`, and the probe performs binary classification of entailment versus non-entailment. The *category experiment* focuses on local features by probing whether the model can identify the presence of specific objects. For this task, prompts take the form `This image contains the following type of object: {category}`, and the probe must predict the correctness of the statement. Both tasks are based on the MS COCO dataset, with positive and negative examples sampled to ensure balanced training and evaluation splits.

2.3 Representation Extraction

Our framework computes hidden representations for every transformer layer of a model given an image–prompt pair. From the token-level hidden states, we derive pooled embeddings that serve as inputs to the probing classifiers. We implemented a general-purpose `pool_tokens` function that supports multiple pooling strategies, including CLS token extraction, mean pooling across valid tokens, max pooling, token-index selection, and a default strategy that retrieves the last non-padding token. Unless otherwise noted, we employ mean pooling, which aggregates information across the entire input sequence while respecting attention masks. This yields a single fixed-size vector per input and per layer, enabling layerwise comparison of representational quality.

2.4 Probing Classifiers

On top of the pooled embeddings, we train lightweight classifiers that map the representations to task labels. For both experiments, we use a simple linear projection with dropout regularization, optimized with Adam and cross-entropy loss. The caption experiment is framed as binary classification, while the category experiment is treated as multi-label prediction with label and mask vectors indicating valid categories for each image. All probes are trained independently per layer, which allows us to quantify how well each layer captures global or local semantic information. Our trainer additionally computes detailed evaluation metrics, including accuracy for caption entailment and macro-averaged F1, precision, recall, and confusion matrix statistics for category recognition.

2.5 Experimental Workflow

In both experiments, the workflow follows the same high-level structure. Datasets are preprocessed and split into training and evaluation sets, with balanced numbers of positive and negative instances. The target model is then loaded, and representations are computed for all inputs. Probing classifiers are subsequently trained and evaluated for each layer. After finishing one model, GPU memory is released and the next model is processed in the same way. The probing results across layers and models are then aggregated for analysis. This setup ensures that our experiments are efficient, reproducible, and directly comparable across models of different architectures and scales.

3 Results

4 Discussion

5 Conclusion

6 References

The LaTeX and BibTeX style files provided roughly follow the American Psychological Association format. If your own bib file is named `custom.bib`, then placing the following before any appendices in your LaTeX file will generate the references section for you:

```
\bibliographystyle{acl_natbib}
\bibliography{custom}
```

You can obtain the complete ACL Anthology as a BibTeX file from https://aclweb.org/anthology/anthology.bib.gz. To include both the Anthology and your own .bib file, use the following instead of the above.

```
\bibliographystyle{acl_natbib}
\bibliography{custom}
```

Please see Section **??** for information on preparing BibTeX files.

6.1 Appendices

Use `\appendix` before any appendix section to switch the section numbering over to letters. See Appendix **??** for an example.

A Example Appendix