Probing Small Vision-Language Models for Global and Local Semantic Representations

Jonathan Schwab
jonathan.schwab@student.uni-tuebingen.de

Tsun Wai Wong
6670979
tsun-wai.wong@student.uni-tuebingen.de

Abstract

Building on foundational studies that mapped semantic representations in models like Kosmos-2, this project investigates how a new generation of compact vision-language models (VLMs) encode global versus local information. We probe the layer-wise representations of FastVLM-0.5B and Qwen2-VL-2B using two tasks on MS COCO: caption-image entailment for global semantics and object recognition for local semantics. Our results reveal a clear distinction: for global semantics, both models progressively build richer representations in later layers, but FastVLM-0.5B's performance exhibits a sharp decline in the final layers while Qwen2-VL-2B remains stable. In contrast, local, object-level semantics are highly and consistently decodable across all layers for both models, with strong F1 scores throughout. These findings indicate that while compact VLMs share a hierarchical build-up for global understanding, they differ in final-layer stability and robustly encode local features across their entire architecture.

1 Introduction

Vision–language models (VLMs) combine text and image understanding and have become central to multimodal AI research. Probing studies help to uncover how such models encode semantic information at different layers, complementing evaluations on end tasks. A recent study (Tao et al., 2024) examined decoder-only multimodal models such as Kosmos and LaVIT, and found that global semantics are most strongly represented in intermediate layers, while upper layers shift toward local, token-level information, which can reduce the ability to encode global meaning. In this work, we extend this line of analysis to recently released compact models such as Qwen2-VL-2B and FastVLM-0.5B. These models represent a new generation of efficient VLMs with parameter counts in the low billions, designed for practical deployment under limited resources. Whether their internal representational patterns follow the same global–local layering observed in earlier models, or show different dynamics, remains an open question. To address this, we developed a general probing framework (Schwab and Wong, 2025) that enables layerwise representation extraction, flexible pooling of token embeddings, and efficient training of lightweight classifiers. Using this framework, we construct two probing tasks on MS COCO (Lin et al., 2014): a caption–image entailment experiment probing global semantics, and an object-category experiment probing local semantics. Following established methodology (Alain and Bengio, 2018), we freeze the VLM parameters, pool hidden states into fixed-size embeddings, and train classifiers on top. We report accuracy, precision, recall, and F1 scores per layer to quantify representational quality. Our experiments provide insight into how current compact VLMs encode global and local information across layers, and allow comparison with trends previously reported for older models (Tao et al., 2024).

2 Methods

2.1 Models

We developed a probing framework for two compact vision–language models: **Qwen2-VL-2B** (Team, 2024) and **FastVLM-0.5B** (Research, 2025). 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 whether the caption correctly describes the image. Due to limited computational resources, we only use one positive and one negative caption per image in contrast to (Tao et al., 2024). We consider a set of images M with associated ground-truth captions C and a set of randomly sampled negative captions C'. For each image $m \in M$, we select one positive caption $c^+(m) \in C$ and one negative caption $c^-(m) \in C'$. This yields the dataset

$$D = \{(m, c^+(m)), (m, c^-(m)) \mid m \in M\}.$$

The probing classifier is then trained to solve the binary decision problem

$$f : D \rightarrow \{0, 1\},$$

with $f(m, c^+(m)) = 1$ and $f(m, c^-(m)) = 0$. The *category experiment* focuses on local features by probing whether the model can identify the presence of specific objects. For this task, 2 different versions of prompts were tested in the experiment. Initially, the prompt takes the form

```
This image contains the following
types of objects: {categories}
```

After reviewing the preliminary results, a prompt without the list of the object categories is implemented to prevent data leakage

```
This image contains the following
types of objects:
```

and the probe must predict the existence of each category in the image. $\mathcal{K}$ denotes the set of object categories, and for each image $m \in M$, the subset $K(m) \subseteq \mathcal{K}$ specifies the categories present in m. The probing task amounts to learning a classifier

$$f : M \rightarrow \{0, 1\}^{|\mathcal{K}|},$$

such that for an image m, the prediction $f(m)$ is a binary vector whose positive entries correspond to the categories in $K(m)$. For this experiment a set of 40 categories is considered, including common occurring objects in the dataset like *person*, *car*, *dog*, and *bottle*. For each positive category in $K(m)$, a negative category is randomly sampled from $\mathcal{K} \setminus K(m)$. Images without positive categories are excluded. Both tasks are based on the MS COCO dataset (Lin et al., 2014), which provides annotations for images, including captions for each image. We created training and evaluation splits. Due to limited computational resources, we restrict our analysis to a random subset of 20,000 images for the training set and 2,000 images for the evaluation set in both experiments.

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. Moreover, our framework stores all extracted representations in a binary format, facilitating efficient reuse and analysis without redundant computation.

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, macro-averaged F1, precision, recall, and confusion matrix statistics.

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. 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

To evaluate the performance of the probing classifiers across different layers and models, we generated a series of plots that visualize key metrics. For the caption experiment, which focuses on global semantic features, the accuracy per layer across both models is shown in Figure 1. Since the positive and negative examples are balanced, a random baseline would achieve 50% accuracy. We can see that both probes show a similar evaluation in accuracy over the most layers, with Qwen2-VL-2B performing slightly better than FastVLM-0.5B. The accuracy increases from the first towards the later layers. Interestingly, the accuracy of FastVLM-0.5B drops significantly in the last layers, while it remains relatively stable for Qwen2-VL-2B.

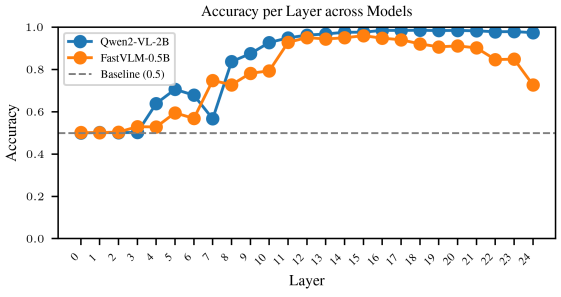

Figure 1: Accuracy per layer across the two models.

In Figure 2, we present a heatmap that summarizes the precision, recall, and F1 scores for the caption experiment across all layers and models. Here we can also observe a similar trend as for the accuracy. The scores increase from the first towards the later layers and FastVLM-0.5B shows a significant drop in the last layers.

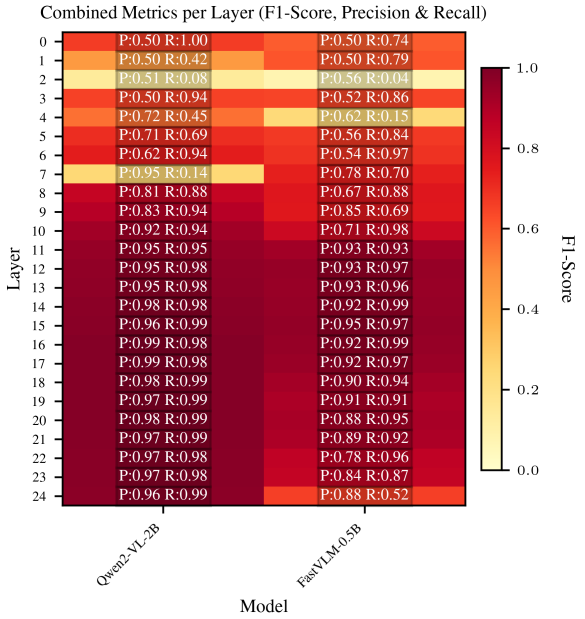

Figure 2: Precision, Recall and F1 score per layer across the two models.

For the category experiment on local semantics, the F1 score per layer across the two models is shown in Figure 3. We can see that both Qwen2-VL-2B and FastVLM-0.5B maintain a consistently high F1 scores of around 93–96% across all layers. In contrast to the global semantic results, the F1 scores here do not show a clear upward trend with increasing layer depth, but instead remains stable throughout. Moreover, no significant drop is observed in the final layers for FastVLM-0.5B, with both models performing nearly identically at each layer.

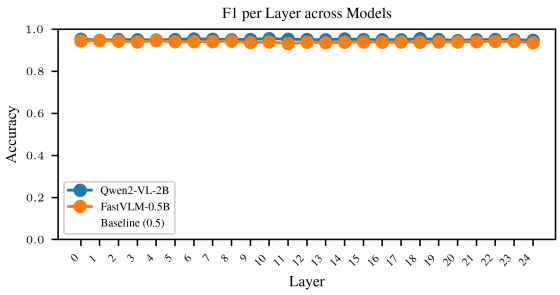

Figure 3: F1 score per layer across all two models in category experiment.

In Figure 4, we present a heatmap that summarizes the precision, recall, and F1 scores for the category experiment across all layers and models. These metrics confirm the same pattern as the accuracy plot: both models achieve consistently high precision and recall (around 94% to 97%) across all

layers, leading to stable F1 scores close to 95%.

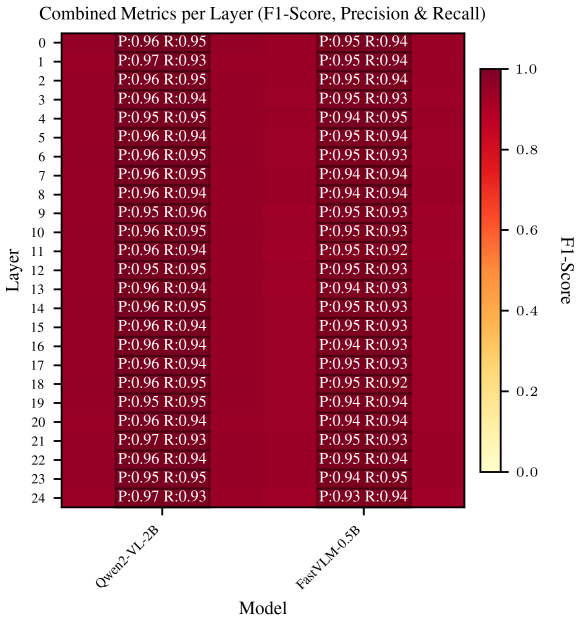

Figure 4: Precision, Recall and F1 per layer across all two models in category experiment.

4 Discussion

4.1 Reflection on the Results

In relation to our central research question of whether small VLMs follow the same global–local representational trends observed in larger models (Tao et al., 2024), our findings reveal both convergences and divergences. Regarding global semantics, we find evidence of a consistent pattern. In smaller models, as in larger ones, intermediate layers encode stronger global semantics than very early layers. This is reflected in the improved performance of middle–later layers on the caption entailment probe. Older vision models such as Kosmos-2 and LaVIT exhibit a gradual decline in global information after their peak in intermediate layers (Tao et al., 2024). FastVLM-0.5B shows a similar behavior, with a sharp drop in performance in the final layers, while Qwen2-VL-2B remains relatively stable. One possible explanation for this difference is that smaller models become increasingly specialized for immediate next-token prediction rather than maintaining a global representation. Smaller models may be forced to prioritize the next-token objective, while larger models have sufficient capacity to preserve both next-token predictive power and global contextual representations.

With respect to local semantics, our findings suggest that local semantic information is represented consistently across layers. Specifically, both the small-scale FastVLM-0.5B and the larger Qwen2-VL-2B encode object-level semantics in a highly linearly separable fashion throughout nearly all layers, with high F1 scores and only minor variation across depth. Whereas (Tao et al., 2024) reported a clear improvement in local semantic encoding from early to mid layers, our results show that small VLMs already capture local semantics with high fidelity from early layers. This could indicate that our task design was too simple. It might also suggest that smaller VLMs saturate their ability to encode simple local semantics earlier in the architecture.

4.2 Limitation of the Study

One of the limitations of our study concerns the way negative examples were constructed. We rely on random sampling to generate negative samples of captions and categories, without accounting for their semantic closeness to the correct choice. A previous study (Rösch et al., 2024) showed that using hard negatives enables models to encode subtler distinctions, leading to stronger fine-grained conceptual understanding. By analogy, when probing to uncover where such information is stored across layers, if the negatives are too easy, layer-specific variation may be underestimated. This may partly explain why our local semantic probes exhibited similar performance across layers as the distinctions of the classifiers may be flattened.

Another limitation of our study lies in the choice of pooling strategy for sequence representations. We applied mean pooling across valid tokens to obtain a single vector representation of the input. This approach implicitly treats all tokens as equally informative. Prior work (Tang and Yang, 2024) has shown that different pooling methods like taking the hidden state of the final token can yield different insights. Exploring alternative pooling strategies and comparing them systematically may provide a more nuanced view of how VLMs distribute information across layers.

Due to computational limits, our probing experiments were restricted to 20,000 training and 2,000 evaluation samples from MS COCO. In addition, coverage of semantics variation would be a potential issue. In particular, only 40 object types have been chosen for the probes in local semantics. This reduced variation means probes might overestimate

how well a model captures the information.

4.3 Future Research Direction

A possible extension of this work would be to broaden the probing framework beyond the current focus on local versus global semantics. While this distinction of semantics captures one important perspective on how models encode information, it is only a partial lens through which to view their representational capacities. Future studies could explore a wider taxonomy of knowledge types. More fine-grained semantic representations, such as relational, compositional, and contextual understanding, as examined in recent probing approaches (Schiappa et al., 2023) can be explored. Beyond relatively more literal and structural dimensions of meaning covered in this study, another potential direction is to investigate how VLMs handle deeper or more abstract forms of knowledge. For example, in the perspective of combinational creativity (Peng et al., 2025), it would be insightful to investigate the models' ability to interpret novel combinations of familiar concepts. Such extensions would provide a richer picture of the models' capabilities in generalization.

While our investigation has yielded useful insights using Qwen2-VL-2B and FastVLM-0.5B, the architectural diversity in our models is still quite limited. To more fully understand where and how information is stored or lost across vision-language models, future work should extend this probing to a broader set of VLM architectures. For example, models under the architecture Encoder–decoder cross-modal fusion like BLIP-2 (Li et al., 2023), which integrate vision and language through deeper cross-attention layers, may preserve global and local semantic information differently. Furthermore, Mixture-of-Experts (MoE) architectures (Lin et al., 2024), which route information dynamically, may distribute semantic content in another way.

5 Conclusion

In this work, we investigated the internal semantic representations of compact vision–language models, specifically Qwen2-VL-2B and FastVLM-0.5B, by probing their layerwise encoding of global and local information. Our experiments on global semantics, measured through a caption–image entailment task, revealed a hierarchical pattern where representational quality generally increases from early to later layers. This trend of improving accuracy in later layers aligns with findings in larger models, providing strong evidence that global understanding is progressively constructed. However, we observed model-specific dynamics in the final layers, with the smaller FastVLM-0.5B exhibiting a notable performance decline while the larger Qwen2-VL-2B remained stable, suggesting that greater model capacity is advantageous for maintaining global representations throughout the network. Conversely, for local semantics, probed via an object category recognition task, our results showed that both models encode object-level information with high fidelity consistently across almost all layers. This uniform high performance may indicate that the task itself was simple enough for the necessary information to be readily decodable across the entire model depth. In conclusion, our study demonstrates that while small VLMs follow the general trend of building global semantic understanding hierarchically, they encode local semantics robustly throughout their architecture. The divergence in final-layer performance also highlights that representational patterns are not uniform across different compact model architectures and scales. These insights contribute to a more nuanced understanding of how architectural scale and design choices impact semantic representation in multimodal models, which can inform the development and application of efficient VLMs.

6 References

References

Guillaume Alain and Yoshua Bengio. 2018. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint*.

Bin Lin, Zhenyu Tang, Yang Ye, Jinfa Huang, Junwu Zhang, Yatian Pang, Peng Jin, Munan Ning, Jiebo Luo, and Li Yuan. 2024. Moe-llava: Mixture of experts for large vision-language models. *arXiv preprint*.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. *arXiv preprint arXiv:1405.0312*, abs/1405.0312.

J. Opitz and S. Burst. 2021. Macro f1 and macro f1. *arXiv preprint arXiv:1911.03347*.

Yongqian Peng, Yuxi Ma, Mengmeng Wang, Yuxuan Wang, Yizhou Wang, Chi Zhang, Yixin Zhu, and Zilong Zheng. 2025. Probing and inducing combinational creativity in vision-language models. *arXiv preprint*.

Apple Machine Learning Research. 2025. Fastvlm-0.5b. https://huggingface.co/apple/FastVLM-0.5B. Apple-amlr License, arXiv:2412.13303 (CVPR 2025).

Philipp J. Rösch, Norbert Oswald, Michaela Geierhos, and Jindřich Libovický. 2024. Enhancing conceptual understanding in multimodal contrastive learning through hard negative samples. *arXiv preprint*.

Madeline Schiappa, Raiyaan Abdullah, Shehreen Azad, Jared Claypoole, Michael Cogswell, Ajay Divakaran, and Yogesh Rawat. 2023. Probing conceptual understanding of large visual-language models. *arXiv preprint*.

Jonathan Schwab and Tsun Wai Wong. 2025. Ulm_probing_vlms. https://github.com/JDev2001/ULM_Probing_VLMs.

Yixuan Tang and Yi Yang. 2024. Pooling and attention: What are effective designs for llm-based embedding models? *arXiv preprint*.

M. Tao et al. 2024. Probing multimodal large language models for global and local semantic representations. *arXiv preprint*.

Qwen Team. 2024. Qwen2-vl-2b-instruct. https://huggingface.co/Qwen/Qwen2-VL-2B-Instruct. Apache-2.0 License, arXiv:2409.12191.

A Appendix

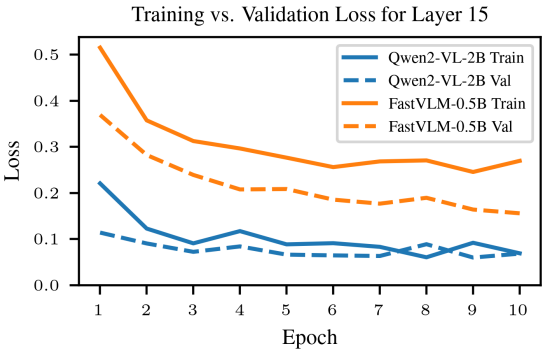

Figure 5: A comparison of training and validation loss for the layer 15 probe (layer with best accuracy scores), analyzing model performance in the global caption experiment. The Qwen2-VL-2B model (blue) demonstrates a significantly lower and more stable loss than the FastVLM-0.5B model (orange). The small gap between the training (solid) and validation (dashed) curves for the Qwen2 model suggests robust training with no signs of overfitting. Moreover, we can see that the strongest learning happens in the first 2 epochs.

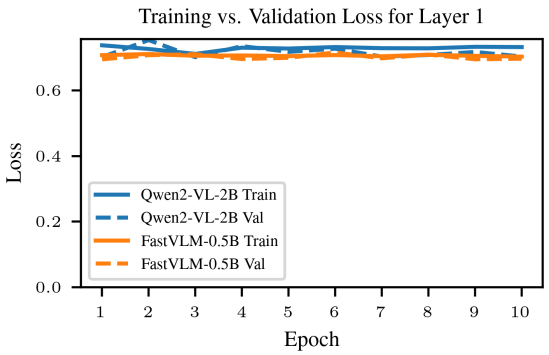

Figure 6: A comparison of training and validation loss for the layer 1 probe (layer with lowest accuracy scores), analyzing model performance in the global caption experiment. Both models show no significant learning, with losses remaining high and stable across epochs. This indicates that the representations at this early layer do not contain sufficient information for the caption entailment task, leading to poor probe performance.

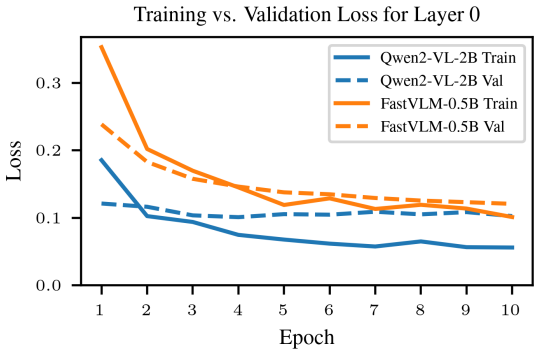

Figure 7: A comparison of training and validation loss for the layer 0 probe (layer with best F1 scores), analyzing model performance in the local category experiment. Both models show most of their improvement within the first four to five epochs, after which the curves flatten. The Qwen2-VL-2B model (blue) achieves consistently slightly lower training and validation losses than the FastVLM-0.5B model (orange), reflecting its stronger performance overall.

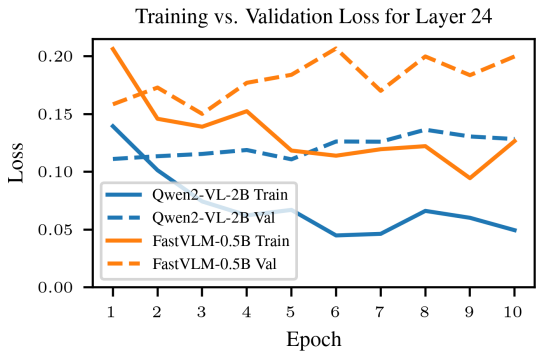

Figure 8: A comparison of training and validation loss for the layer 24 probe (layer with lowest F1 scores), analyzing model performance in the local category experiment. Both models show limited improvement after the initial epochs, with losses remaining relatively flat overall. The validation losses remain relatively low, indicating that this layer also carries useful information for the task. The FastVLM-0.5B model (orange) shows slightly higher and less stable losses compared to Qwen2-VL-2B (blue).