Circuit Tracing in Vision–Language Models: Understanding the Internal Mechanisms of Multimodal Thinking

Supplementary Material

9. Transcoder Training

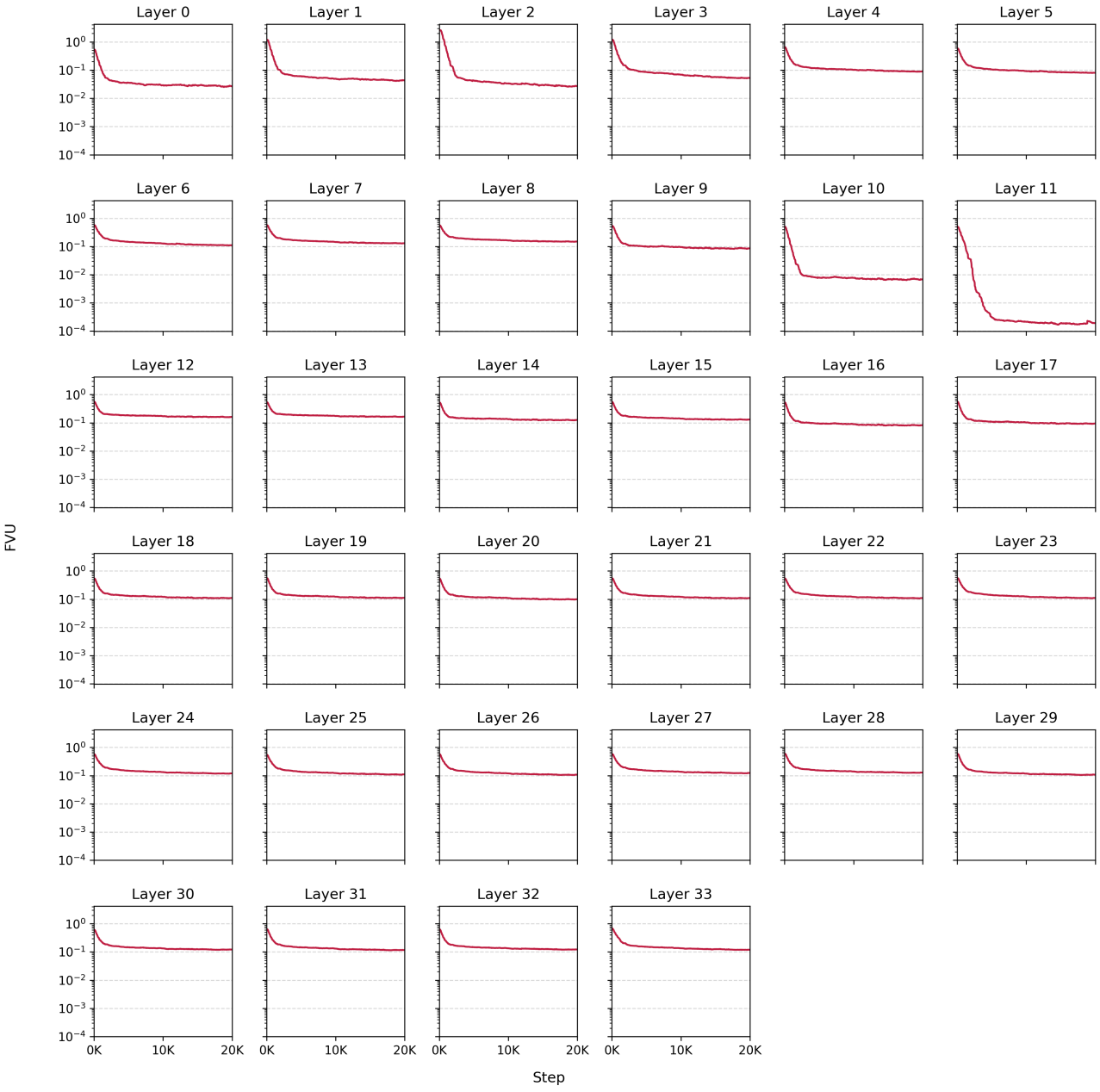

Figure 8. FVU (Fraction of Variance Unexplained) Training curve for Gemma-3-4B-IT.

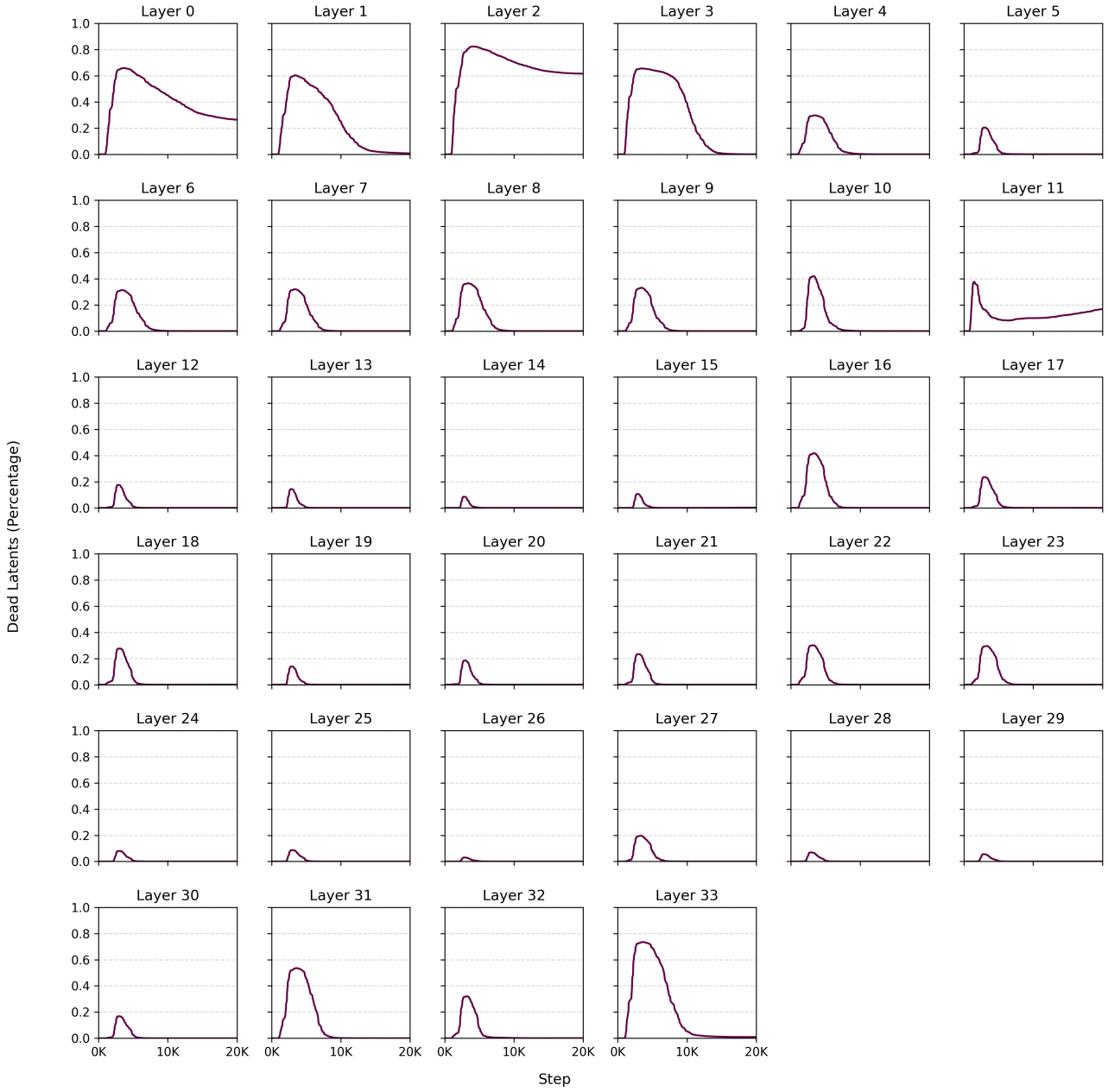

Figure 9. Percentage of dead features over training for all layers of Gemma-3-4B-IT.

9.1. Dead Latents

As shown in Figure 9, a feature is marked as dead if it fails to activate sufficiently over a given interval, and it is considered active once it activates sufficiently again. The percentage of dead latents provides insight into how features utilize the available expansion size. A high percentage does not necessarily imply high polysemanticity; it may instead reflect insufficient expansion capacity to support fully monosemantic features, causing the model to default to a denser cluster of max-k features that occupies less space than the nominal expansion factor. We hypothesize that the lower layers may either represent condensed, naturally superpositioned embeddings originating from the vision encoders, or may require a much larger expansion factor to

be properly disentangled. In either case, we currently lack a mature method for accurately extracting low-level representations—such as patterns, colors, or other subtle features—that we believe are present in these layers, due to the difficulty of identifying consistent patterns when examining activation example images in aggregate. We plan to follow up on this in future works. Layer 11, being a global-attention layer, stands out from the local layers around it and creates a clear shift in the model's training curve. Because it suddenly incorporates full-context information, its behavior differs enough to change the curvature. Its transcoder is harder to interpret for the same reason—it mixes information from across the entire sequence rather than nearby tokens. Even so, this complexity is contained within the layer, so it doesn't disrupt the rest of the circuit-tracing process.

10. Feature Discovery

Feature discovery on multi-modal inputs is significantly more expensive than on text-only LLMs. For text activations, a single token with its surrounding context is often sufficient to explain a feature's behavior. As a result, a paragraph of a few hundred tokens can provide hundreds of useful activation examples.

In contrast, image inputs in VLMs produce far more tokens. For Gemma-3-4B, each image yields 256 tokens. Although a sufficiently dense image could, in principle, provide unique and informative signals across all 256 embeddings, most images do not contain enough complexity for all tokens to meaningfully activate distinct features. In typical or unfavorable cases, the full set of 256 image tokens may provide no more explanatory coverage than a single text token. Moreover, the vision encoder introduces additional computational overhead.

These factors are the primary reason we have not yet analyzed all VLM features. We hope to optimize our pipeline and explore alternative methods for obtaining feature activations more efficiently.

11. Circuits

11.1. Hallucination - Fingers

We analyzed the common hallucination in which the model predicts five fingers instead of six. Remarkably, for this model, the logit for the token 6 is no higher than for other unrelated digits (e.g., 7 or 1). Although we have previously identified features involved in visual counting tasks (e.g., features that activate on groups of three), we were unable to find any feature corresponding to the concept of six or six visual objects in the attribution graph.

Instead, we found a direct circuit between the image embeddings, features associated with the number 5, and the final output for the '5' token. Notably, contrary to the typical assumption that the model would rely on a semantic "hand"

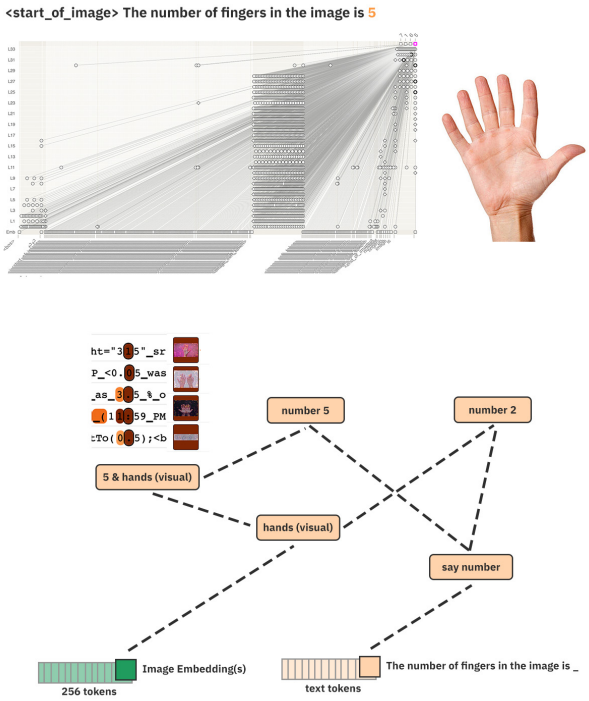

Figure 10. Circuit tracing analysis of prompt "The number of fingers in the image is " with the result being 5, despite image containing six fingers.

concept, the circuit appears to be driven by a visual feature representing the visual concept of a hand together with the number five. In other words, in this task, the visual concept of hand activates the concept of five, rather than the model performing a robust object-counting procedure.

We also observed features related to the number 2 in upper layers, which—possibly due to the input—become activated by the concept of a hand and by features representing the function "say a number." We hypothesize that this may arise because hands occur in pairs across animals, causing the model's number-related circuits to be influenced by hand priors rather than meaningful counting information from the encoder. Overall, this suggests that the error may stem from the encoder failing to provide a sufficiently strong representation of a six-fingered hand—or from the model implicitly choosing to ignore such evidence.

11.2. Caption - Sea Otter

We conducted circuit tracing on an image of a seaotter, which required two-step circuit analysis due to seaotter being two tokens. We observe for the token "sea" the logic was in the 90+% range, and we analyzed the circuit prompted seaotter.

Interestingly, we also found that this image strongly activated a feature that represented sea lions - likely due to their

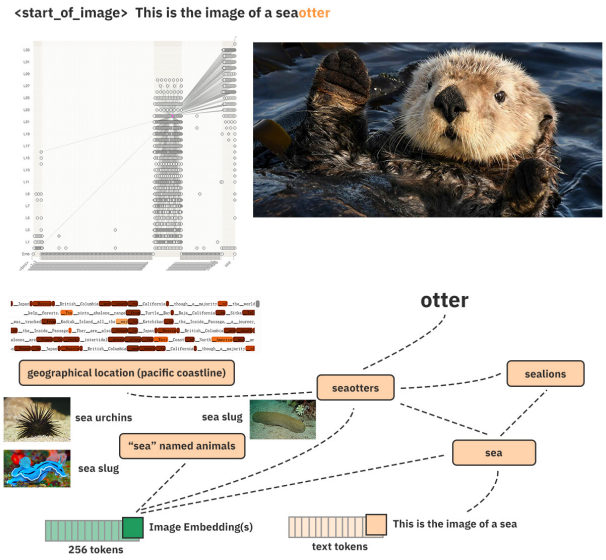

Figure 11. Circuit tracing analysis of a seaotter, with interesting feature attributions.

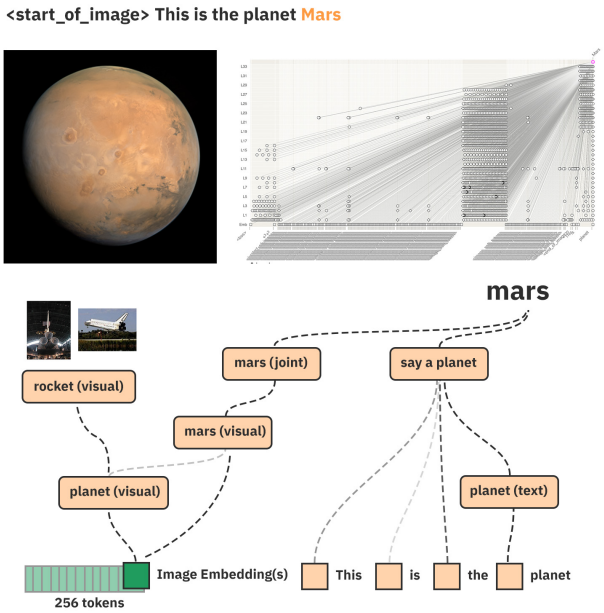

Figure 12. Analysis of the mars circuit, revealing a transition from purely visual features to a semantically grounded feature that controls the final output.

visual similarities (the feature was also present in the first circuit without the sea token input), this shows that there may exist a purely visual latent space in the model.

We also found another feature that included visual images of animals that contained the word "sea", such as sea urchins, sea slug, and sea cucumbers. There also exists a knowledge feature of a geographical range, which just happens to be the pacific coastline from California to Japan - where sea otters primarily live in.

11.3. Caption - Mars

We analyzed the mars prompt in Figure 12, which reveals a clean and interpretable circuit showing how raw visual representations from the encoder are gradually transformed into a unified multimodal feature representing the concept of Mars. This joint feature blends both the semantic textual notion of the word "mars" and the visual appearance of the planet, demonstrating how the model builds a shared representation that supports cross-modal grounding. The circuit further illustrates how early visual features—such as a generic "planet" detector—feed into progressively more specific features until they converge on a single joint concept that dominates the logit contribution for the final output token.

Notably, the circuit also highlights a broader associative structure present in mid-lower layers. Features representing planets reliably activate features associated with rockets and space shuttles, even though these objects do not appear in the input image. This suggests that the model has learned a latent web of visual associations that mirrors human conceptual priors: planets evoke spacecraft, just as certain animals evoke particular habitats or behaviors. These associations arise purely from visual features rather than explicit textual grounding, indicating that VLMs develop an internal "association of ideas"—where visually related objects coactivate—even before higher-level semantic features take over. This phenomenon provides insight into how VLMs integrate visual context, prior knowledge, and semantic structure when generating grounded descriptions.

11.4. Reasoning - Simple Addition

As shown in Figure 13, we analyzed a simple addition problem with an image of an incomplete equation. We found that at lower layers, the image embeddings activated features related to numerals, such as a feature representing a number between 0-5, showing that visual-semantic convergence may also occur at lower layers for low-level representations.

We also note that the model contains a very diverse feature representation of numbers. For example, we identified a feature that primarily activates on images of charts, with the attention roughly focusing on the 3 range of the axis. This feature is potentially used for visual-numerical reasoning. We also found features activating on objects in groups of three, this shows that object count information is passed from the vision encoder, and that the VLM decoder contains features that can properly represent this. We note that this

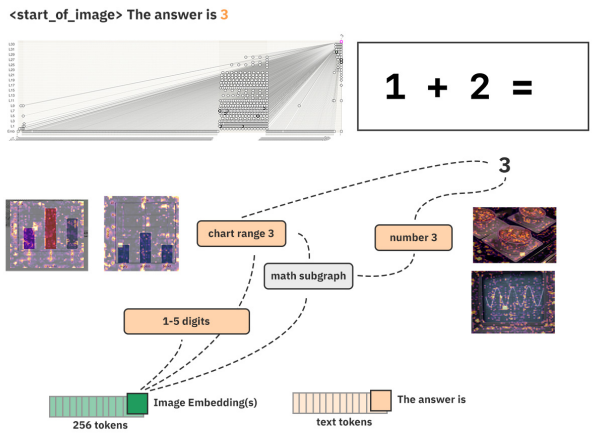

Figure 13. Circuit tracing analysis of a simple addition task.

is entirely absent in the finger hallucination example which prompts us to incline towards the vision encoder not encoding such information properly.

In the middle-upper layers, we discovered a group of semantic features that activates on mathematical operations, and we hypothesize that this is the subgraph that performs this operation in semantic space. Analyzing the purely semantic addition circuit is beyond the scope of our analysis, and we recommend referring to the circuit tracer paper by Lindsey et al. [19] for a deeper analysis.

12. Conclusion

In this appendix, we presented additional analyses generated using our proposed VLM circuit-tracing methodology, illustrating its ability to reveal structure across visual, semantic, and multi-modal representations. Through examinations of transcoder behavior, feature activation patterns, and several representative circuits, we showed how VLMs blend visual features with language-level abstractions—sometimes in unintended ways. These case studies highlight several recurring themes: the limitations of current vision encoders in providing robust object-level information; the presence of latent spaces that mix visual similarity, linguistic priors, and associative knowledge; and the emergence of compact, semantically meaningful features even within early or visually dominated layers. Our circuits serves as a useful tool to debug hallucinations, false knowledge, and internal mechanisms of VLMs for future improvements.

Overall, these findings demonstrate both the promise and challenges of interpreting VLMs. While our method successfully uncovers many of the mechanisms driving model behavior, it also reveals gaps—such as missing low-level disentangled features or overreliance on semantic priors—that motivate further refinement of both architectures and interpretability tools. We hope that the insights provided here, along with the methodology introduced in the main paper, will help pave the way for a more systematic understanding of multimodal reasoning, its failure modes, and the underlying circuits that support these abilities in modern VLMs.