Towards Understanding Multimodal Fine-Tuning: A Case Study into Spatial Features

Extended Abstract

Large vision–language models (VLMs) achieve strong performance by fine-tuning pretrained language backbones to accept projected image tokens alongside text. Yet we lack a mechanistic account of how backbone representations adapt and when vision-specific capabilities arise. We introduce a stage-wise model diffing framework for multimodal training, extending a method previously applied only to language models into the multimodal setting. By comparing sparse autoencoder (SAE) dictionaries across regimes [1], we provide the first mechanistic account of how visual grounding reshapes backbone features. Concretely, we warm-start LLaMA-Scope SAEs trained on LLaMA-3.1-8B [2] and adapt them to multimodal activations from LLaVA-More using 50k VQAv2 pairs, yielding a feature-level view of how a language model develops the ability to "see."

We finetune SAEs at each transformer block under four regimes: full sequence, image-only, text-only, and random initialization, allowing us to separate the effects of vision and text spans. Text-only SAEs converge rapidly to near-zero reconstruction error and remain aligned with the base LLM dictionary, while image-only and full-sequence runs plateau at higher error due to the mismatch between projected image tokens and the pretrained basis. Warm-starting from pretrained SAEs clearly outperforms random initialization, stabilizing alignment and showing that finetuning is more effective than training from scratch, thereby providing a reliable foundation for stage-wise diffing.

To detect features reshaped by multimodal fine-tuning, we introduce two signals: (i) modality preference, quantified by the variance gap between multimodal and text-only activations, and (ii) geometric reorientation, measured by cosine shifts in decoder directions relative to the base SAE. A two-stage filter yields a sparse, well-defined set of vision-preferring features that reorient during training. This adapted subset represents fewer than 5% of all features, is concentrated in mid-to-deep layers, and departs significantly from the original text-only dictionary.

We then ask whether these adapted features encode spatial grounding. A controlled dataset shift from general VQA to spatial queries (e.g., left/right/above/behind) shows that a subset is selectively recruited under spatial prompts and remains active under neutral instructions, confirming that they capture spatial meaning rather than superficial lexical cues. Automated interpretation and manual inspection confirm consistent meanings such as object placement, relative position, and orientation.

Crucially, we test their causal role. Using attribution patching, a scalable gradient-based proxy for activation interventions, we find that a small set of mid to deep attention heads consistently drive these spatially selective features. High-scoring heads localize to semantically relevant regions, while low-scoring heads fail to capture structure. Ablating either the identified features or the implicated heads leads to clear drops in visual spatial reasoning accuracy, showing that these pathways are not only correlated with but necessary for spatial grounding.

Taken together, our findings show that multimodal adaptation is structured and interpretable. A sparse set of vision-preferring features reorient to encode spatial relationships, and a correspondingly small set of specialized heads act as their causal channel. While our experiments focus on LLaVA-More with a LLaMA-3.1-8B backbone, the methodology is general and provides a foundation for auditing and refining multimodal training in safety-critical or domain-specific applications.

References

[1] T. Bricken et al. *Stage-Wise Model Diffing*. Transformer Circuits, 2024. `https://transformer-circuits.pub/2024/model-diffing/index.html`

[2] Z. He et al. *LLaMA-Scope: Extracting millions of features from LLaMA-3.1-8B with sparse autoencoders*. arXiv:2410.20526, 2024.

Submitted to WiML Workshop @ 39th Conference on Neural Information Processing Systems (NeurIPS 2025).