# Expert Collapse and Compositional Failure in Simple Multimodal MoE

**Anthony Ticinovic** ATICINOV@OUTLOOK.COM
**Caren Han** CAREN.HAN@UNIMELB.EDU.AU
*The University of Melbourne, Victoria, Australia*

## Abstract

Mixture-of-Experts (MoE) is a technique that uses multiple MLPs at each transformer layer, rather than using a single MLP. MoE architectures are hypothesised to create specialised experts, but this is often inferred rather than quantitatively measured. Further, this specialisation is in conflict with standard load-balancing losses that promote more uniform load distribution, though can encourage expert redundancy. We construct a novel two-unimodal-expert (vision/text) MoE testbed and use a three-stage protocol to first force specialisation with uni-modal data (Stage 2), then test its stability during fine-tuning (Stage 3). Our findings demonstrate that while Stage 2 successfully creates specialised experts, this specialisation persists only at the object-level. In Stage 3, the standard multimodal loss actively overwrites this structure, causing the latent space to default to only clustering by *modality*, rather than by *concept*. We identify the mechanism as *layer-level expert collapse*. Furthermore, a case study on compositional binding reveals that even when specialisation is present, it captures monolithic objects (e.g., 'car') but fails to bind attributes like colour, perhaps highlighting a source of bias in multimodal representation.

**Keywords:** Multimodal AI, Mixture of Experts, Interpretability, Bias, Modality Interference

## 1. Introduction

Mixture-of-Experts (MoE) architectures have become a cornerstone of state-of-the-art generative models, enabling significant increases in parameter count while maintaining a constant computational budget (Shi et al., 2024). This paradigm has been successfully applied to Vision-Language Models (VLMs) like Astrea (Yang et al., 2025) and MoE-LLaVA (Lin et al., 2024).

A primary challenge stems from knowledge interference, where models struggle to fairly evaluate all modalities. This leads to "parametric memory interference", where the model is biased to relying on parametric statistical patterns from its language training rather than adhering to specific visual evidence (Cai et al., 2025). Since these architectures rely on pretrained LLM backbones, this defaulting is typically dominated by the text domain. In this work 'bias' is defined as representational modality bias (e.g., the model ignoring visual evidence in favour of textual priors), distinct from societal or demographic biases.

Furthermore, the core hypothesis of the MoE paradigm is that experts develop meaningful specialisations. Work on advancing expert specialisation identifies a paradox where standard auxiliary load-balancing losses inadvertently promote redundancy (Guo et al., 2025). This loss forces a more homogeneous data stream upon experts, leading to overlap rather than the focused functions that define true specialisation.

This paper addresses these gaps by using a novel VLM testbed, using a staged training protocol inspired from Sha et al. (2024). Though unlike the Uni-MoE method (Sha et al., 2024), we train modality *pure* experts to investigate whether forced specialisation persists under standard training objectives. We also use full finetuning rather than LoRA techniques. Our central hypothesis is that by explicitly training experts to be modality-specific, we can create a system that is more interpretable, and naturally disentangles vision and text parameter interference. However, we find that standard end-to-end training actively overwrites this modal disentangling, leading to expert collapse and compositional noise in the latent space.

## 2. Methodology

We employ a three-stage progressive training protocol to enforce and then stress-test modality specialisation. The VLM is constructed from a frozen CLIP Vision Encoder (ViT-L/14) (Radford et al., 2021), a linear Vision-Language Connector (MLP), and a Mistral-7B-v0.3 backbone (Jiang et al., 2023). See Figure 1.

### 2.1. Architectural Blueprint

The Feed-Forward Network (FFN) sub-layer within each Transformer block of Mistral is replaced with our custom MoE layer containing two experts: `Expert 0` (Vision) and `Expert 1` (Text). A linear router manages top-1 gating.

### 2.2. Three-Stage Progressive Training

We adapt the Uni-MoE protocol (Sha et al., 2024) but introduce a divergence in Stage 2 to force strict uni-modal specialisation (Figure 2).

1. **Stage 1: Cross-Modality Alignment.** We train only the Vision-Language Connector to align CLIP embeddings with the LLM latent space (Li et al., 2023).
2. **Stage 2: Modality-Specific Experts (Hard Routing).** We bypass the router and use deterministic hard-routing. Vision tokens are forced to `Expert 0`, and text tokens to `Expert 1`. This ensures experts diverge and specialise on uni-modal data.
3. **Stage 3: Unified Fine-Tuning (Soft Routing).** We unfreeze the router, attention layers, and experts. The model is trained on mixed-modality data using a Straight-Through Gumbel-Softmax estimator (Jang et al., 2017) for dynamic routing. Crucially, we do *not* use auxiliary load-balancing loss, as Sha et al. (2024) indicate it promotes expert redundancy rather than specialisation in this architecture.

### 2.3. Concept-Modality Analysis

To quantify alignment, we construct a Concept-Modality Comparison Matrix. We sample $N = 50$ images and captions for 8 concrete concepts (e.g., car, apple) from MS-COCO (Lin et al., 2014). We extract hidden states from specific layers and compute cosine similarity between the mean-pooled visual representation $\bar{\mathbf{h}}_{V,c_i}$ and the textual representation $\bar{\mathbf{h}}_{T,c_i}$.

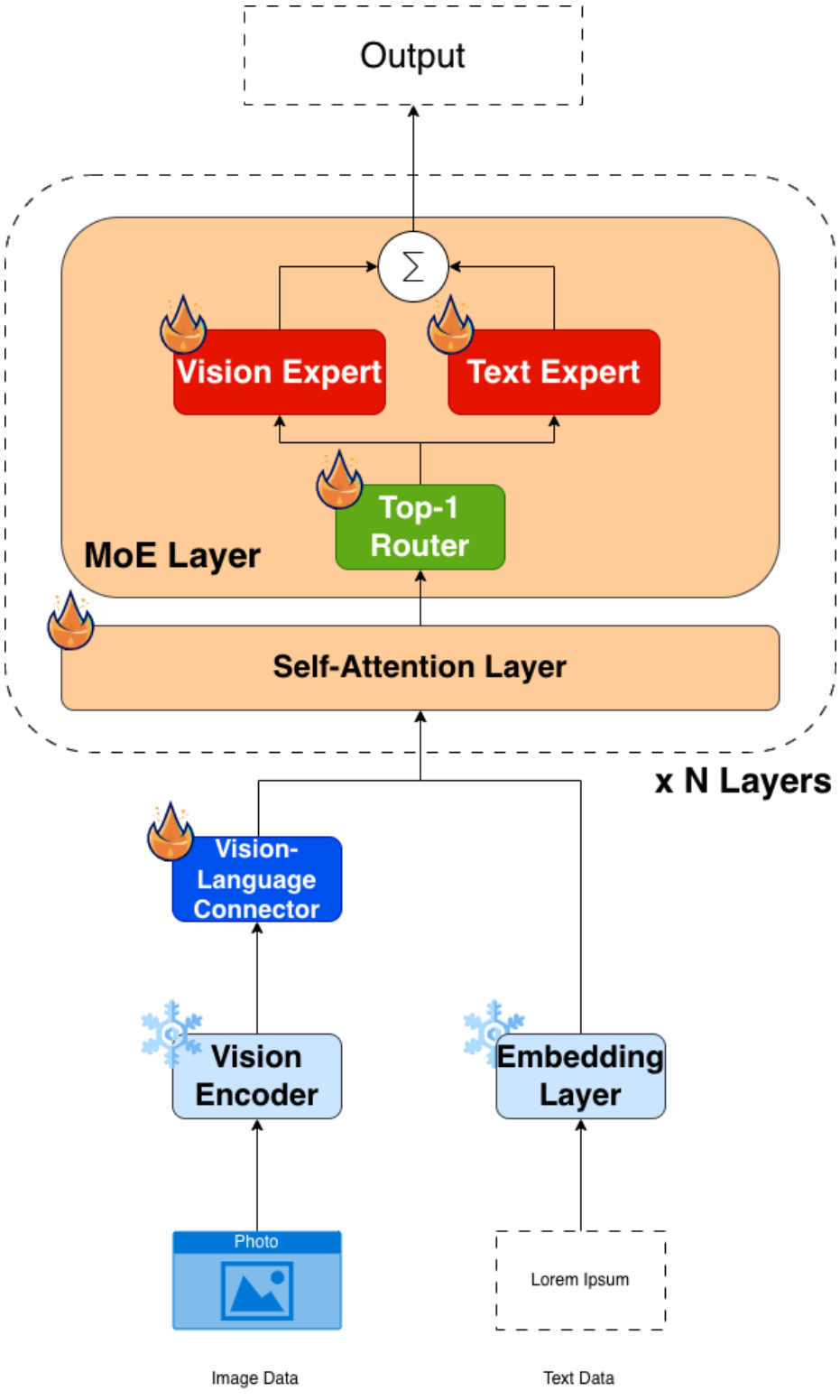


Figure 1: High-level architecture of the VLM model. Trainable parameters are denoted with a flame icon. (adapted from Sha et al., 2024)

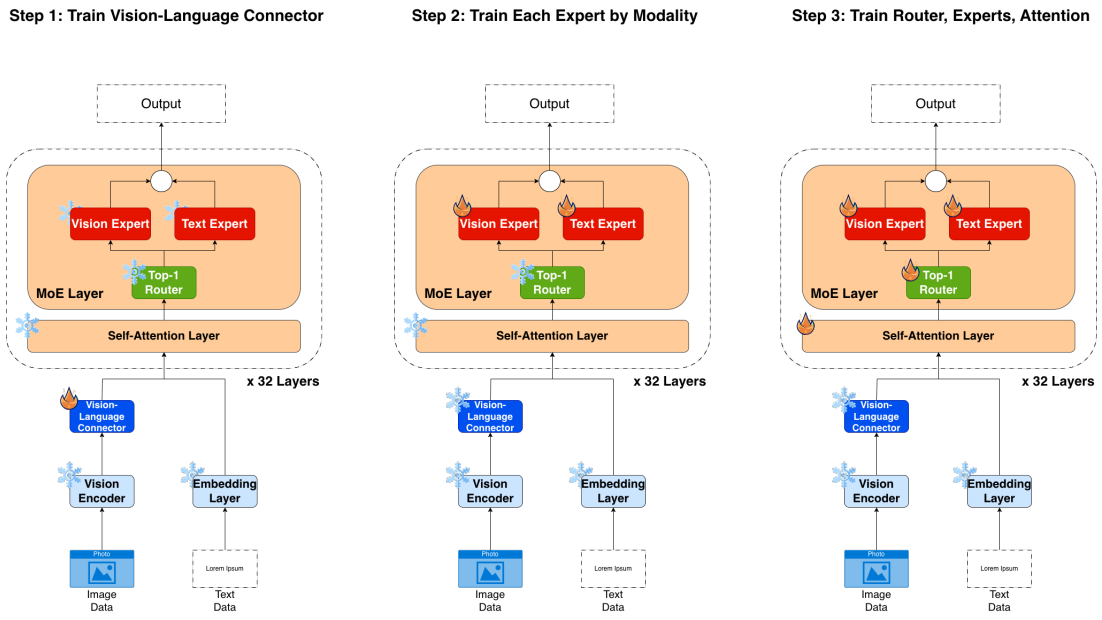


Figure 2: The three-stage training protocol. Stage 2 forces disentanglement via hard routing. Stage 3 uses learned dynamic routing.

## 3. Experimental Results

### 3.1. Success of Architectural Forcing (Stage 2)

The hard-routing protocol in Stage 2 successfully created specialised experts. Figure 3 shows a clear diagonal structure in the cross-modal similarity matrix. This indicates that the vector for a text concept is more similar to the vector for its visual counterpart, despite the experts never processing mixed data.

We validated this with a one-tailed binomial test ($k = 6$ successes in $n = 8$ trials), yielding a statistically significant p-value of $\approx 8.52 \times 10^{-5}$.

We further confirmed this functional specialisation via a crossover test, observing an 11% increase in loss when expert routing was inverted.

### 3.2. The Collapse of Specialisation (Stage 3)

However, this specialisation proved unstable. After Stage 3 fine-tuning with a standard multimodal loss, the cross-modal structural alignment vanished (Figure 4).

The t-SNE analysis in Figure 5 reveals the cause, illustrating the latent space of the final model separates cleanly by *modality* rather than concept. The model prioritises separating text from images over aligning their semantic meaning.

We identified *Layer-Level Expert Collapse* (Figure 7) as the mechanism of failure. In most layers, the router defaults to sending 100% of tokens to a single expert, regardless of content. This suggests the standard loss function found a simpler local minimum that ignored the specialised experts built in Stage 2.

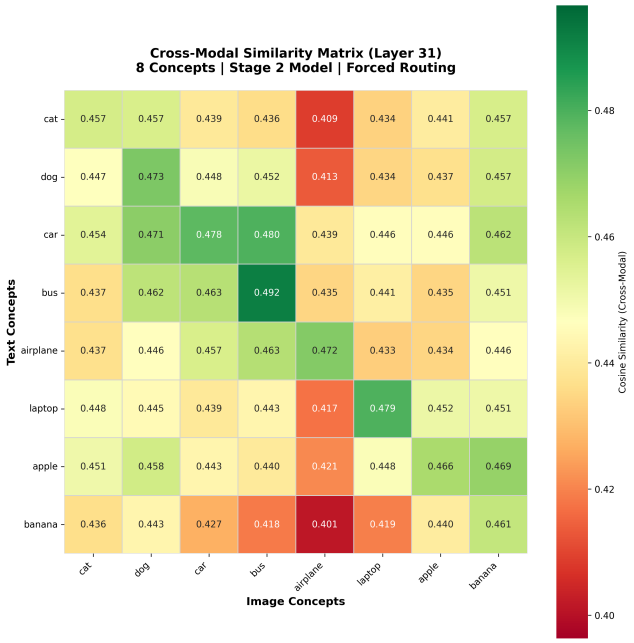


Figure 3: Stage 2 Similarity Matrix (Layer 31). The bright diagonal confirms that forced routing successfully aligns cross-modal concepts.

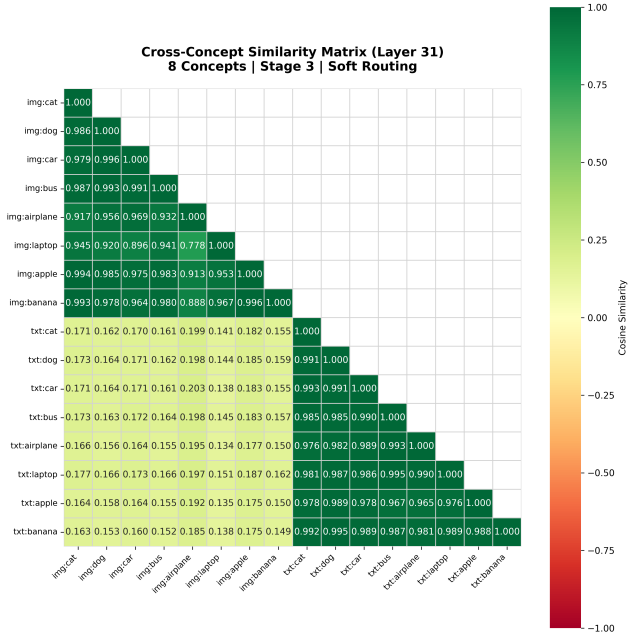


Figure 4: Stage 3 Similarity Matrix (Layer 31). The diagonal structure is lost, indicating the model has "un-learned" the concept alignment.

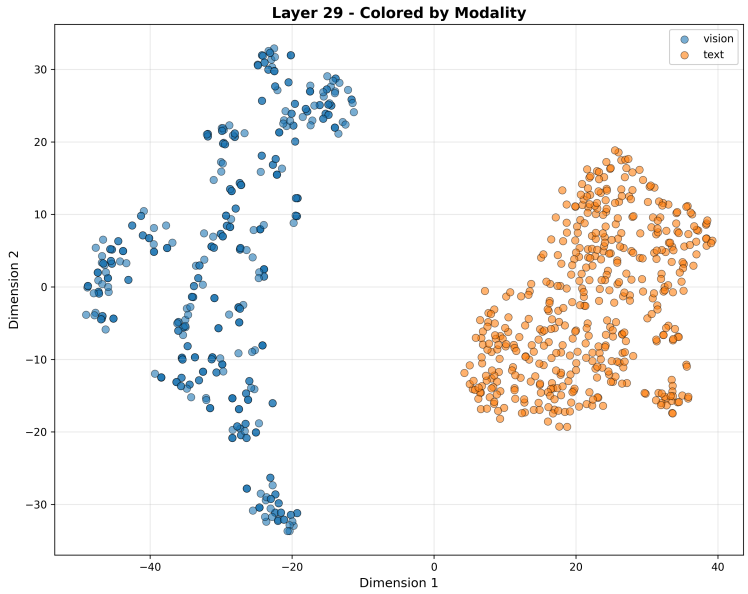


Figure 5: t-SNE of Stage 3 Hidden States (Layer 29). The model clusters tokens by modality (Text vs. Vision) rather than semantic concept.

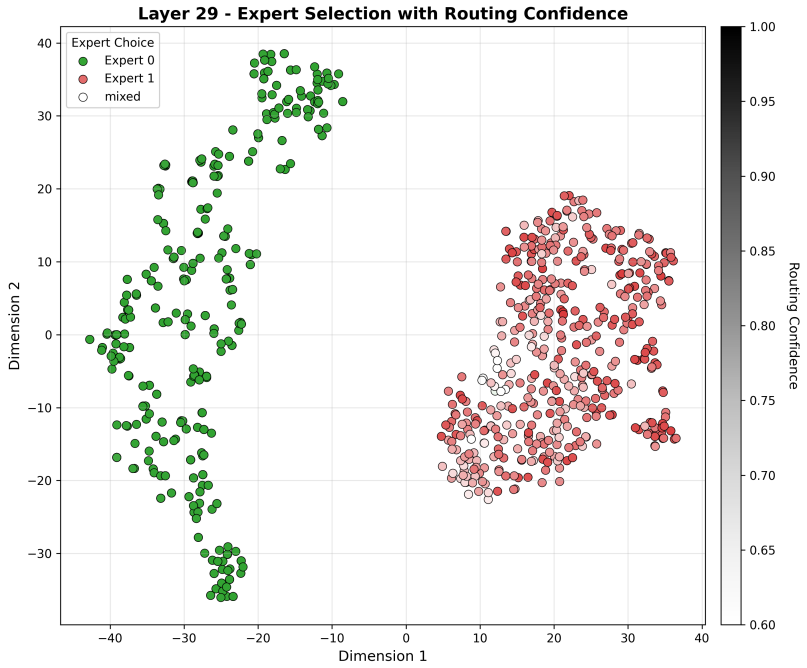


Figure 6: t-SNE of Stage 3 Hidden States (Layer 29). The router learns the modality separation perfectly at this layer, which is the exception.

It is important to note that the collapse was not uniform. As seen in the expert load distribution (Figure 7), while most layers collapsed to a single expert, specific deeper layers (e.g., Layer 29) maintained a balanced load. We further investigate this behaviour in Figure 6, which visualises the expert routing decisions at this layer. Unlike the collapsed layers, Layer 29 exhibits a clean, bimodal routing distribution that perfectly aligns with the modality clusters. The router successfully learned to send visual tokens to Expert 0 and text tokens to Expert 1. This finding is critical as it shows that the architecture *retained* the capacity for disentangled, modality-specific processing. The global collapse is therefore not due to a lack of model capacity, but rather an optimisation bias where the standard loss function failed to incentivise this disentangled structure against the simpler local minimum of expert collapse.

The transition to LLaVA-Instruct in Stage 3 was designed as a robustness stress test. While the domain shift from captioning to instruction tuning introduces natural *target* distribution variance, it does not necessitate the observed layer-level expert collapse. If the specialisation were robust, the router could have adapted the existing experts to the new task. Instead, it largely abandoned the specialised structure.

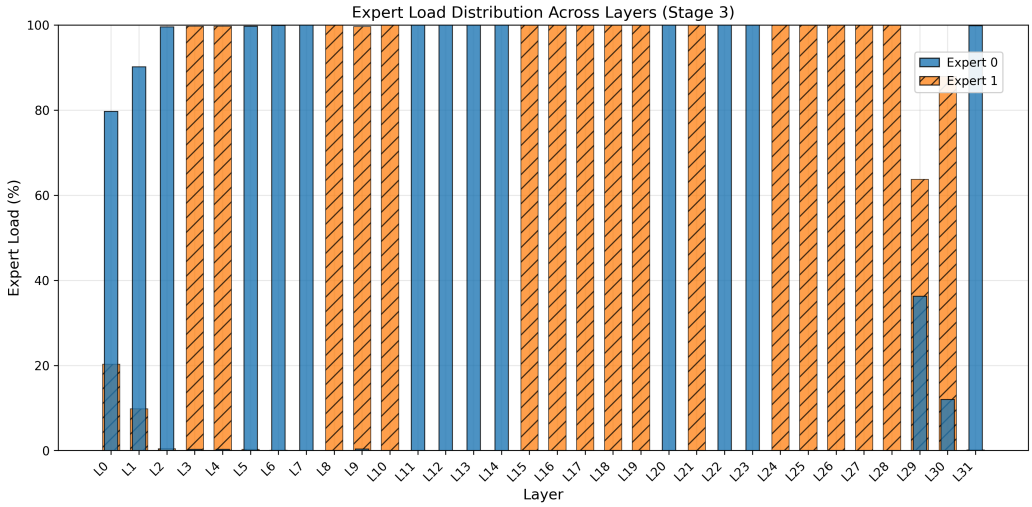


Figure 7: Expert load distribution by layer (Stage 3). Most layers collapse to a single expert (blue or orange bars), ignoring the specialised structure.

### 3.3. Compositional Bias Case Study

To test the richness of the representations cross-modal representations formed by the model, we analysed compositional concepts (e.g., ‘red_apple‘ vs ‘green_apple‘).

As shown in Figure 8, even the specialised Stage 2 model exhibits *Object Dominance*. ’Red apple’ is as similar to ’green apple’ as it is to itself. The model fails to bind the attribute ’colour’ to the object. This representational failure highlights a risk. The model relies on monolithic object priors rather than attending to specific visual attributes (like colour or size).

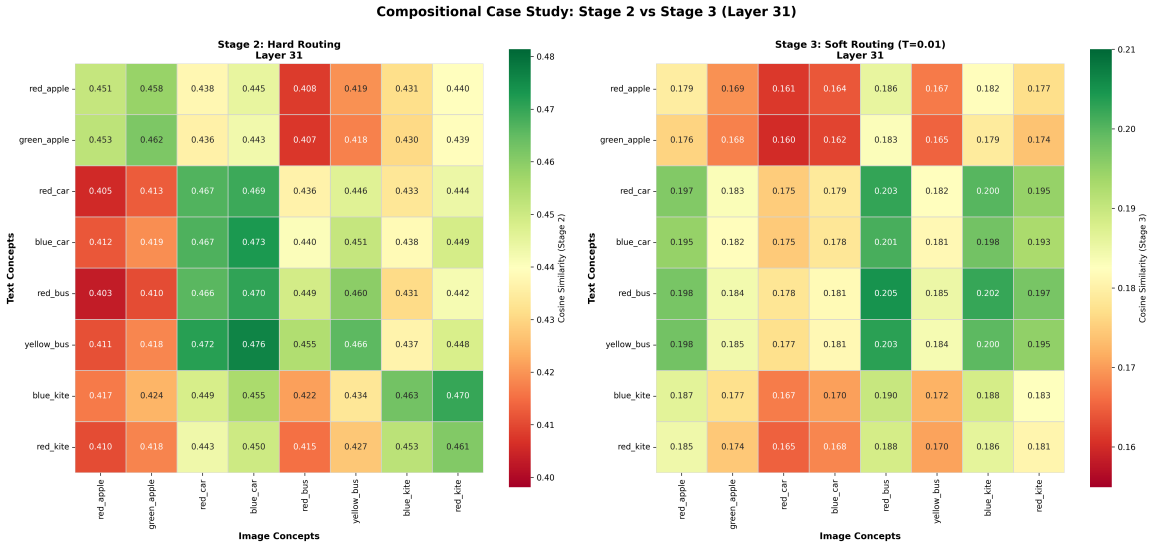


Figure 8: Compositional Matrix. Note the failure in Stage 2 (Left): 'Red Apple' and 'Green Apple' are similar but non distinct, indicating the model relies on object stereotypes rather than visual attributes. In Stage 3 (Right), no compositional structure exists.

### 3.4. Performance Impact

The loss of specialisation in Stage 3 correlated with catastrophic forgetting. On the COCO captioning benchmark, the Stage 2 model achieved a CIDEr score of 76.2, while the Stage 3 model dropped to 8.1 (Table 1), showing the VQA centred training of Stage 3 greatly diminished the model's captioning ability. Though not central to this study, further work should be done to ensure training stability between the tasks/data.

Table 1: Performance on COCO Karpathy test split (Bucciarelli et al., 2024).

| Model | Data | B-4 | M | R-L | C |
|---|---|---|---|---|---|
| *Reference: LLaVA-v1.5-7B* | | | | | |
| + Full FT | COCO | **38.2** | 23.5 | **57.3** | **111.4** |
| *Our Models (MoE)* | | | | | |
| Stage 2 | COCO | 31.9 | **33.3** | 55.4 | 76.2 |
| Stage 3 | COCO → LLaVA-Ins | 4.2 | 12.2 | 29.9 | 8.1 |

## 4. Discussion

This study illustrates the tension between architectural interpretability and standard multimodal training objectives. Our findings highlight three critical implications for the design

of robust, unbiased Vision-Language Models. That is, the fragility of architectural forcing, greedy router optimisation, and grounding failure as a source of bias.

We demonstrated that while hard-routing (Stage 2) successfully instantiates distinct semantic representations, this structure is brittle. The collapse in Stage 3 suggests the standard loss favours a path of least resistance. Separating tokens by modality, the optimisation process abandoned cross-modal alignment for a simpler local minimum, rarely using the structure derived in Stage 2.

Crucially, Layer 29 reveals the architecture *is* capable of the desired behaviour, correctly routing tokens even after fine-tuning. This isolates the failure mode. It is not a lack of capacity, but a failure of the objective function to incentivise its utilisation more globally. The standard loss permits locally greedy routing, allowing layer-level expert collapse to propagate.

The compositional case study reveals a shallow encoding of concepts in the latent space. Being unable to distinctly cluster based on compositional concepts shows evidence for only object-level semantic structure. Further, this is evidence for reliance on prioritising statistical priors over pixel-level evidence, and may be evidence of parametric memory interference (Cai et al., 2025).

## 5. Conclusion and Future Work

We conclude that architectural priors alone are insufficient to guarantee grounding and clean concept-level representations. While we successfully forced expert specialisation, standard training actively destroyed it, replacing concept alignment with modality-based clustering. To address limitations, future research must pursue the following:

1. **Hard-Routing in Stage 3**: Future work could investigate maintaining hard-routing during Stage 3 fine-tuning to further determine if expert collapse is driven by the routing mechanism or the multimodal objective.
2. **Specialisation-Preserving Loss Functions:** To prevent collapse, auxiliary loss terms should explicitly penalise the router for deviating from established expert roles.
3. **Compositional Auditing Frameworks:** Auditing datasets and benchmarks that specifically test attribute-object binding (e.g., separating "red car" from "blue car") may help to identify and quantify parametric interference.
4. **Verifying Findings Across Model Families:** As an $N = 1$ study on Mistral 7B, the immediate next step is to replicate this experiment on other backbones to distinguish fundamental properties from model-specific artefacts.

## References

Davide Bucciarelli, Nicholas Moratelli, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. Personalizing multimodal large language models for image captioning: An experimental analysis, 2024. URL https://arxiv.org/abs/2412.03665.

Rui Cai, Bangzheng Li, Xiaofei Wen, Muhao Chen, and Zhe Zhao. Diagnosing and mitigating modality interference in multimodal large language models, 2025. URL https://arxiv.org/abs/2505.19616.

Hongcan Guo, Haolang Lu, Guoshun Nan, Bolun Chu, Jialin Zhuang, Yuan Yang, Wenhao Che, Sicong Leng, Qimei Cui, and Xudong Jiang. Advancing expert specialization for better moe, 2025. URL https://arxiv.org/abs/2505.22323.

Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax, 2017. URL https://arxiv.org/abs/1611.01144.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Mesnard, Thibaut Gauthier, and Badr Youbi Idrissi. Mistral 7b, 2023. URL https://arxiv.org/abs/2310.06825.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models, 2023. URL https://arxiv.org/abs/2301.12597.

Bin Lin, Zhenyu Tang, Yang Ye, Jinfa Huang, Junwu Zhang, Yatian Pang, Peng Jin, Munan Ning, Jiebo Luo, and Li Yuan. Moe-llava: Mixture of experts for large vision-language models, 2024. URL https://arxiv.org/abs/2401.15947.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context, 2014. URL https://arxiv.org/abs/1405.0312.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. URL https://arxiv.org/abs/2103.00020.

Feng Sha, Run-Ze Fan, Runda Wu, Zekun Wang, Yifei Shen, Difei Gao, Lichao Sun, and Mike Zheng Shou. Uni-moe: Scaling unified multimodal llms with mixture of experts, 2024. URL https://arxiv.org/abs/2405.11273.

Bokai Shi, Yutao Ning, Zankun Zhang, Zhaowei Zhang, Wenxin Piao, Yuxiang Zhang, Zhaohui Li, Wenyu Du, Zhepeng Lv, WenChao Zhang, Guanyu Li, Biao Geng, Yixuan Li, Xin Zhang, Guanjun Liu, and Lisai Zhang. A survey on mixture-of-experts in large language models, 2024. URL https://arxiv.org/abs/2407.11181.

Xiaoda Yang, Xin C. Guo, Tianhe Ren, Zhaofeng Niu, Jiahe Shi, Cheng-Ching Tseng, Wenyu Liu, Shuo Chen, Jianlong Chang, Albert Xing-Jian Jiang, Chenglei Wu, Qi Tian, Gloria Jie-Yu Zhao, Alex Yue-Ming Yin, and Dit-Yan Yeung. Astrea: A moe-based visual understanding model with progressive alignment, 2025. URL https://arxiv.org/abs/2503.09445.