ARCHITECTURE-DEPENDENT FUSION PATHWAYS IN MLLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal Large Language Models (MLLMs) achieve strong performance across vision-language tasks, yet the internal mechanisms by which visual and textual information are fused across layers remain insufficiently understood. We investigate representative MLLMs from two architectural paradigms: concatenation architectures and native multimodal architectures. We conduct three progressively connected analyses: alignment decoupling identifies which modality changes, attention routing and entropy characterize how cross-modal information is distributed, and intrinsic dimensionality examines how fusion reshapes feature spaces. Separately, we perform causal intervention experiments as a validation of the resulting interpretation. As a supplementary analysis, we use visual CKA to examine the Platonic Representation Hypothesis. Together, these analyses reveal two distinct fusion pathways: concatenation models follow a text-first, vision-later pathway, whereas native models exhibit earlier visual-textual co-adaptation and feature-space reorganization. This work provides a mechanistic perspective for understanding multimodal fusion and supports architecture-aware diagnostics of multimodal representations.

1 INTRODUCTION

MLLMs have rapidly advanced a wide range of vision-language applications, including human-computer interaction (Achiam et al., 2023), medical image analysis (Tu et al., 2024), and autonomous driving (Sima et al., 2024). Despite strong empirical progress, most studies still evaluate MLLMs primarily through downstream task accuracy and multimodal reasoning benchmarks (Zhu et al., 2023; Wang et al., 2024). Final performance, however, does not reveal how visual and textual tokens interact inside the model. The question is therefore not only whether an MLLM solves a task, but also how it fuses modalities across depth.

A central factor behind this problem is architectural diversity. Existing MLLMs can be broadly categorized into two paradigms according to how visual and textual information are integrated (Yin et al., 2024): *concatenation architectures* and *native multimodal architectures*. We refer to Section 2.1 for the architectural definitions and representative examples. Although these paradigms differ in how they introduce and process visual information, their layer-wise fusion dynamics have not been systematically characterized.

Prior work has revealed modality gaps in multimodal representation learning (Liang et al., 2022), shown that attention mechanisms shape multimodal representations (Hendricks et al., 2021), and demonstrated that neural representations evolve substantially across layers (Phang et al., 2021; Lee et al., 2025). Motivated by these findings, we first conduct a layer-wise cross-modal representation analysis using Centered Kernel Alignment (CKA) (Kornblith et al., 2019). CKA measures the similarity between two feature matrices and is widely used for comparison in multimodal analysis because it is robust to neuron permutation and common linear transformations.

As illustrated in Figure 1, concatenation architectures generally exhibit relatively high similarity in early layers, although the magnitude depends on the employed vision encoder. These models then experience a pronounced decline in intermediate layers, suggesting a temporary divergence between modalities during feature transformation. In contrast, native multimodal architectures begin changing earlier and maintain cross-modal interaction through the early and intermediate stages.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

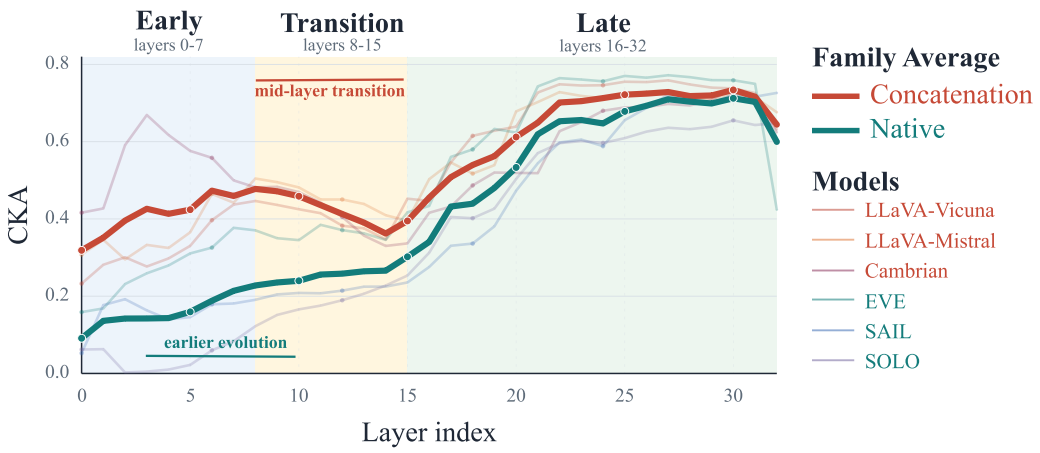


Figure 1: Cross-modal CKA similarity across layers. **Family average** represents the average values of different types of models. Shaded regions mark the early, transition, and late layer ranges. The model classification is described in EXPERIMENTAL SETUP 3.1. Model-name colors distinguish concatenation from native architectures.

Beyond the transition range, similarity generally recovers in middle-to-deep layers, with model-specific declines near the final layers. This phenomenon suggests that different architectures follow distinct fusion pathways before eventually forming more similar high-level representations.

Cross-modal CKA alone is insufficient to explain the mechanism behind this pattern. First, it is an aggregate similarity score and cannot determine which modality is primarily changing during fusion. Second, it does not reveal how information is routed between visual and textual tokens. Third, it cannot distinguish whether representation changes correspond to feature compression, expansion, or structural reorganization. Finally, representation-level analysis alone does not establish whether the identified transition layers are functionally important.

We therefore organize the study around three progressively connected analyses. First, alignment decoupling separately tracks the layer-wise evolution of visual and textual representations. Second, attention routing and entropy reveal how information flows between modalities and whether cross-modal attention is concentrated or diffuse. Third, intrinsic dimensionality measures whether fusion corresponds to feature-space compression, expansion, or reorganization. Separately, causal intervention tests whether the identified routing and layer-transition patterns are functionally relevant. Together, the three analyses reveal two fusion pathways: concatenation models follow a text-first, vision-later pathway, whereas native models show earlier visual-textual co-adaptation.

Our main contributions are summarized as follows:

- **Three progressive experiments.** We conduct three related experiments that move from which modality changes, to how cross-modal information is routed, to how feature spaces reorganize, revealing two distinct fusion pathways.
- **Causal validation.** We separately conduct intervention experiments to test whether the mechanisms identified above are functionally relevant.
- **Supplementary Platonic analysis.** We conduct a visual CKA experiment to test whether deep visual representations across MLLMs become more aligned, providing evidence relevant to the Platonic Representation Hypothesis (Huh et al., 2024).

2 PRELIMINARY

2.1 MLLM ARCHITECTURES

MLLMs can generally be divided into concatenation architectures and native multimodal architectures according to the design of the vision-language interface (Yin et al., 2024). Concatenation ar-

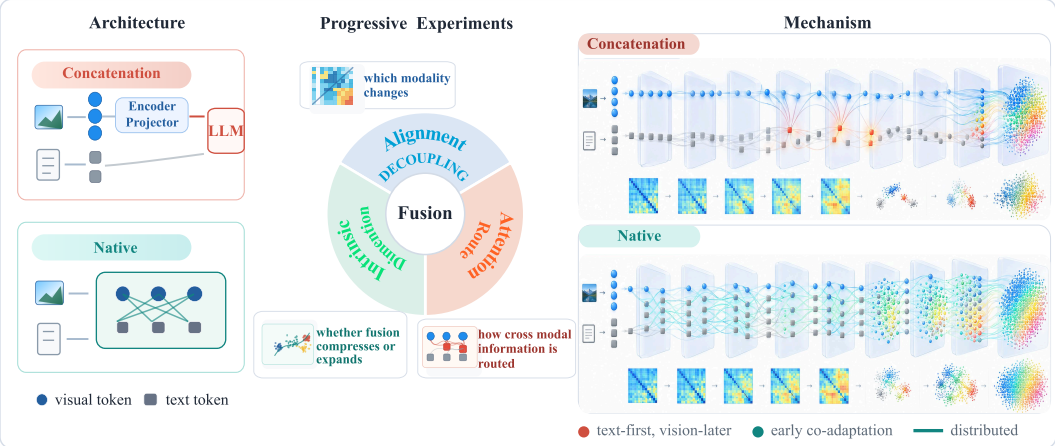


Figure 2: Schematic overview of the study design. The left side contrasts concatenation and native multimodal architectures, the center groups the three progressive experiments, and the right side summarizes the two fusion pathways suggested by these analyses.

chitectures preserve the original language model backbone while introducing lightweight projection modules to align visual features with the textual embedding space. Most of these frameworks rely on pretrained CLIP visual encoders (Radford et al., 2021), whose large-scale contrastive pretraining provides transferable semantic representations for multimodal alignment. The core assumption of such architectures is that the visual encoder already provides sufficiently high-level semantics for visual tokens (Li et al., 2023a).

Native multimodal architectures learn visual tokens and language representations within a shared end-to-end multimodal framework. Their visual front ends may be jointly optimized or replaced by encoder-free visual tokenization, so the category is defined by the multimodal learning interface rather than by a single tokenization implementation. Our work is complementary: rather than treating the empirical difference between architectures as self-evident, we examine the internal mechanisms that give rise to different fusion pathways.

2.2 MLLM Alignment and Interpretability

Recent works often adopt the heuristic that stronger alignment leads to better multimodal performance (Li et al., 2021; Almudévar et al., 2025; Singh et al., 2022). However, recent evidence suggests that stronger alignment does not always lead to better performance (Fang et al., 2026). Other studies analyze internal information flow in MLLMs but remain focused primarily on concatenation architectures (Zhang et al., 2025) or on representation alignment alone (Lu et al., 2025). Existing probing-based analyses often focus on isolated layer representations rather than the complete evolution of multimodal fusion (Phang et al., 2021). In summary, prior work has not yet provided a comprehensive explanation of how different MLLM architectures fuse modalities across depth.

3 Experiments

3.1 Experimental Setup

To explain the cross-modal CKA phenomenon, we design three progressively connected analyses from the perspectives of representation alignment, attention routing and entropy, and intrinsic feature-space structure. The first identifies *which modality changes*, the second explains *how the change is mediated*, and the third examines *whether the change corresponds to compression or expansion*. We then conduct a separate causal intervention validation to test *whether the identified mechanism is functionally relevant*.

Model Selection. We evaluate three concatenation models, LLaVA-next with Vicuna and Mistral backbones (Liu et al., 2024a) and Cambrian (Tong et al., 2024), and three native multimodal models,

SOLO (Chen et al., 2024), EVE (Diao et al., 2024), and SAIL (Lei et al., 2025). The LLaVA series uses CLIP-ViT-L (Radford et al., 2021) as its visual encoder, whereas Cambrian is a distinct concatenation model employing a hybrid visual encoder that integrates DINO (Oquab et al., 2023) and CLIP (Radford et al., 2021). Because the two architectural paradigms use fundamentally different training pipelines, a perfectly controlled comparison that fixes every factor is not feasible. Concatenation models typically combine a pretrained visual encoder, a projection module, and a language model backbone, while native models learn visual tokenization and language representations in an end-to-end multimodal framework with model-specific supervision pipelines. Our goal is to identify paradigm-level regularities under representative open models. To reduce single-model artifacts, we evaluate multiple models per paradigm and require the same interpretation to be supported across CKA, attention, intrinsic-dimensionality, and intervention evidence.

Dataset and Input Construction. We randomly sample 10,000 image-text pairs from COCO-val (Lin et al., 2014), VQA (Antol et al., 2015), and MMMU (Yue et al., 2024) to cover different levels of visual complexity, textual semantic density, and multimodal reasoning requirements. For the representation and attention analyses, we use clean caption-style image-text inputs to reduce noise from long reasoning chains or text spans that are only weakly grounded in the image. All special tokens, padding tokens, and truncation positions are recorded and masked during analysis to ensure that the measured statistics reflect only effective visual and textual tokens. We evaluate complex reasoning behavior separately through the causal perturbation experiment on MMBench (Liu et al., 2024b) using VLMEvalKit (Duan et al., 2024).

Feature and Attention Extraction. For each model and each input sample, we extract the post-norm hidden representation matrix $H^{(l)} \in \mathbb{R}^{N_l \times d}$ from every layer, where N_l denotes the number of valid tokens and d denotes the hidden dimension. In all experiments, the analyzed hidden representations are normalized layer by layer and use a common hidden size of $d = 4096$. The numbers of visual and textual tokens follow the default configuration of each model, so the analysis preserves the native tokenization and visual-token construction of different architectures. According to the token indices of each model, we separate the hidden states into visual and textual token sets, denoted as $H_V^{(l)}$ and $H_T^{(l)}$, respectively. Layer indices are zero-based (0–32), with index 0 denoting the initial CKA reference representation. We also extract post-softmax attention matrices from all layers and heads.

Statistical Robustness. To quantify sampling uncertainty, we conduct non-parametric bootstrap resampling of pairs on a 5,000-pair subset drawn from population used in the main representation analysis. For each sample-size group, we perform 1,000 independent resamplings with replacement, compute the target value for each layer, and report bootstrap mean, standard error, and relative bias. Appendix A.6 reports the detailed results for Cambrian at layers 1, 15, and 30.

3.2 Alignment Decoupling

Motivation. The global cross-modal CKA curve in Figure 1 shows that different architectures follow distinct fusion trajectories. However, cross-modal CKA alone is an aggregate measure and cannot determine which modality is primarily responsible for the observed change. A decrease in visual-textual similarity may indicate that visual tokens are being reshaped by the language manifold, that textual tokens are adapting to visual semantics, or that both modalities are changing simultaneously. This analysis answers the first mechanistic question: *during multimodal fusion, who is moving toward whom?*

Method. In addition to the cross-modal similarity $\mathrm{CKA}_{V,T}^{(l)}$, we compute intra-modality CKA relative to the initial representation (index 0). For each input sample, visual and textual token representations are first averaged within their corresponding modality to obtain sample-level pooled representations, which makes representations comparable when different models or modalities have different token numbers:

$$\mathrm{CKA}_{V,V}^{(l)} = \mathrm{CKA}\left(H_V^{(0)}, H_V^{(l)}\right), \qquad \mathrm{CKA}_{T,T}^{(l)} = \mathrm{CKA}\left(H_T^{(0)}, H_T^{(l)}\right).$$

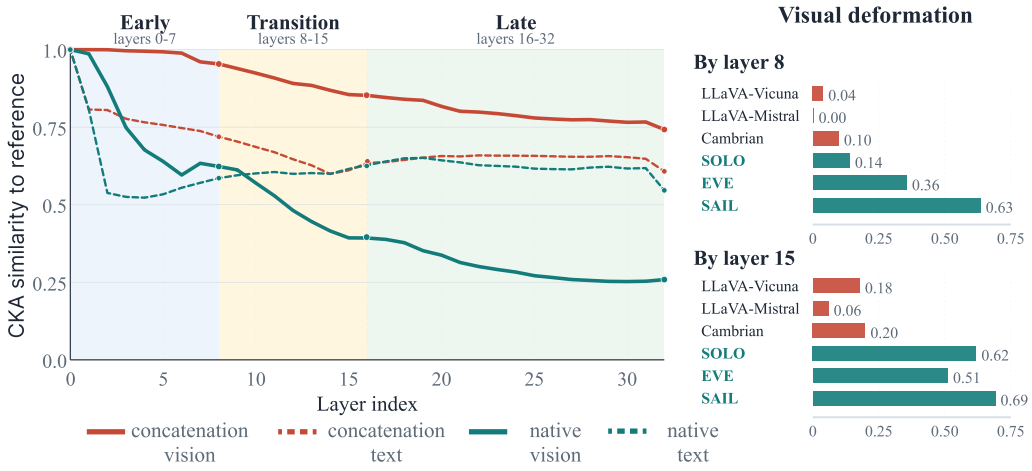


Figure 3: Intra-modality CKA relative to the initial reference representation. Left: Solid and dashed lines show family-average visual and textual CKA, respectively. Right: Per-model visual deformation ($1 - \mathrm{CKA}_{V,V}$), at displayed layer indices 8 and 15. Individual-model curves are provided in Appendix A.2.

Here, $\mathrm{CKA}^{(l)}_{V,V}$ measures how much the visual representation changes, while $\mathrm{CKA}^{(l)}_{T,T}$ measures the corresponding evolution of textual representations. A higher value indicates that the modality retains a representation close to its initial state, whereas a lower value indicates stronger layer-wise deformation.

Results. Figure 3 illustrates distinct differences between concatenation and native multimodal architectures. For concatenation models, visual curves stay close to the initial reference representation in shallow layers, meaning visual tokens largely retain the structure from the pretrained visual encoder at the input of the language model stack. By contrast, textual representations deform earlier and stabilize in middle layers. Following this textual adaptation, visual curves drop more evidently, indicating late reshaping of visual representations. The per model bars at layers 8 and 15 further show that concatenation models maintain higher visual CKA relative to native models.

Native multimodal architectures follow a different trend: their visual representations diverge from the initial reference state much earlier, while textual representations evolve with depth with model-specific curve patterns. Thus, the core finding is not perfectly synchronized cross-modality transformation across all layers, but that native models trigger earlier and more continuous visual-textual co-adaptation compared with concatenation models. The individual visual and textual trajectories in Figure 8 further show the model-specific patterns underlying these family averages.

3.3 Attention Routing and Entropy

Motivation. The previous experiment identifies which modality changes during fusion, but it does not explain how such changes are implemented inside the model. Since MLLMs are transformer-based architectures, attention is the primary operator through which tokens exchange information. Routing ratios measure how much attention mass one modality allocates to the other (Abnar & Zuidema, 2020), while entropy measures whether this attention is concentrated on a small number of tokens or dispersed across many tokens (Clark et al., 2019). This analysis answers the second mechanistic question: *through what information-flow pattern does cross-modal fusion occur?*

Method. Given visual token set V and textual token set T, we define the routing mass from query modality M to key modality N as

$$r^{(l)}_{M\to N} = \frac{1}{|M|}\sum_{i\in M}\sum_{j\in N}\bar{A}^{(l)}_{ij}, \qquad M, N \in \{V, T\},\ M \neq N.$$

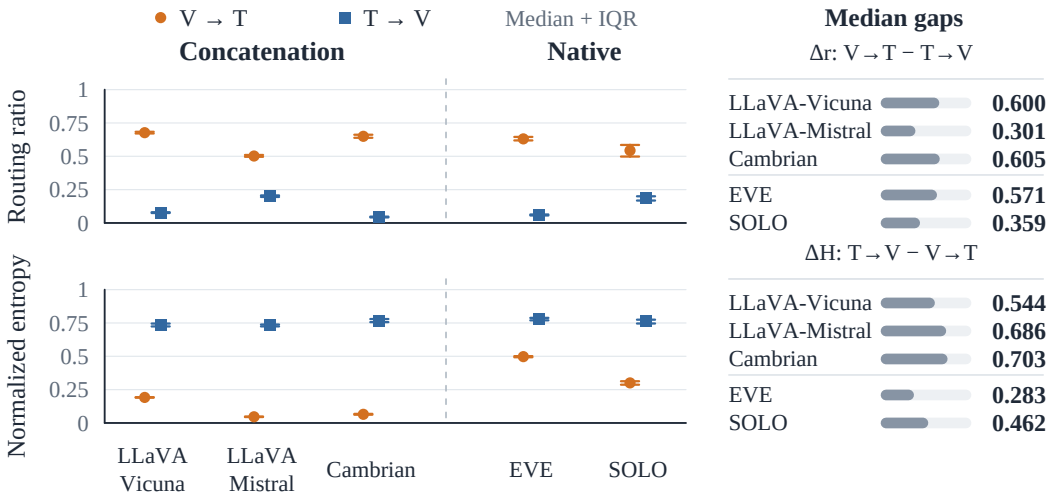


Figure 4: Cross-modal attention over layers 8–15. Top: routing ratio; bottom: normalized entropy. Orange circles denote $V \to T$ and blue squares denote $T \to V$, where arrows indicate query-to-key attention allocation. Left: medians and interquartile ranges of per-pair layer averages. Right: differences between the displayed medians, Δr ($V \to T$ minus $T \to V$) and ΔH (the reverse subtraction). SAIL is omitted because its hybrid attention path constrains visual-to-text entropy. Layer-wise curves for all six models appear in Appendix A.3.

Here, $r^{(l)}_{V\to T}$ measures how much attention mass visual query tokens allocate to textual keys, while $r^{(l)}_{T\to V}$ measures the reverse direction. We normalize over target tokens:

$$\tilde{A}^{(l,M\to N)}_{ij} = \frac{\bar{A}^{(l)}_{ij}}{\sum_{j'\in N} \bar{A}^{(l)}_{ij'} + \epsilon}, \qquad j \in N.$$

Since different modalities may contain different numbers of target tokens, we further use the normalized cross-modal entropy

$$\widehat{H}^{(l)}_{M\to N} = -\frac{1}{|M|} \sum_{i\in M} \frac{\sum_{j\in N} \tilde{A}^{(l,M\to N)}_{ij} \log\left(\tilde{A}^{(l,M\to N)}_{ij} + \epsilon\right)}{\log(|N| + \epsilon)}.$$

A lower $\widehat{H}^{(l)}_{M\to N}$ indicates more concentrated cross-modal attention, while a higher value indicates more diffuse attention over the target modality. In the implementation, entropy is normalized for each sample using its actual target-token count before aggregation. For Figure 4, we first average each pair's routing ratio and normalized entropy across layers 8–15, then report the median and interquartile range across pairs. The right-hand summaries report differences between the displayed directional medians: visual-to-text minus text-to-vision for routing (Δr), and text-to-vision minus visual-to-text for entropy (ΔH).

Results. Figure 4 shows directional routing asymmetry in all displayed models: the median visual to text routing ratio exceeds the reverse-direction median, with Δr ranging from 0.301 to 0.605. Large routing gaps occur in both architectural families; for example, Cambrian and EVE have gaps of 0.605 and 0.571, respectively. Routing magnitude alone therefore does not clearly separate the two families.

Attention concentration provides a more differentiated comparison. The concatenation models have median visual-to-text normalized entropies of 0.046–0.192, compared with 0.497 for EVE and 0.301 for SOLO. Their directional entropy gaps are also larger: $\Delta H = 0.544$–0.703 for the concatenation models, versus 0.283 for EVE and 0.462 for SOLO. These observations are consistent with more concentrated visual-query attention to textual keys in the concatenation models, and more diffuse attention in this direction in EVE and SOLO.

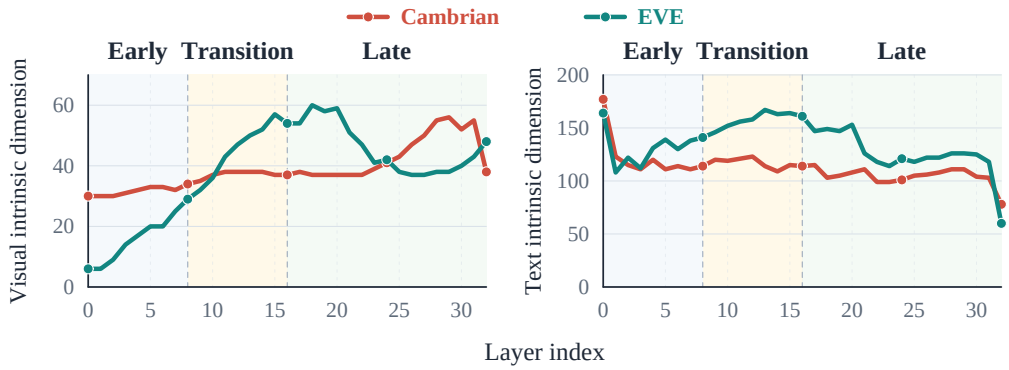


Figure 5: Intrinsic dimensionality at $\tau = 0.7$ for representative concatenation (Cambrian) and native (EVE) models. The left and right panels report visual and textual intrinsic dimensionality on separate scales. Results for additional models and threshold sensitivity are reported in Appendix A.4.

3.4 Intrinsic Dimensionality Transition

Motivation. Both CKA and attention statistics are relative measures. CKA describes similarity between representations, and attention describes how tokens interact, but neither directly indicates whether the feature space is being compressed, expanded, or reorganized during fusion. For example, a decrease in CKA may correspond either to information loss, feature specialization, or expansion into a richer semantic subspace. To address this limitation, we analyze the intrinsic dimensionality of visual and textual representations. This provides an absolute measure of the effective complexity of the feature space and answers the third mechanistic question: *what kind of structural transformation occurs during multimodal fusion?*

Method. Because the hidden representations are normalized layer by layer and compared at a fixed hidden size, the intrinsic-dimensionality curves are not caused by differences in hidden size or activation scale across layers. The token numbers are kept as each model's default setting, while the same normalization and variance-threshold protocol is applied to all models and modalities. For each modality $M \in \{V, T\}$ and each layer l, we perform SVD on the centered feature matrix. Let σ_i denote the i-th singular value and let τ denote the variance preservation threshold. The intrinsic dimensionality is defined as

$$\mathrm{ID}^{(l)}_{M,\tau} = \min\left\{ k : \frac{\sum_{i=1}^{k} \sigma_i^2}{\sum_i \sigma_i^2} \geq \tau \right\}.$$

We evaluate multiple thresholds $\tau \in \{0.5, 0.6, 0.7, 0.8\}$ to ensure that the observed trends are not caused by a particular variance cutoff. A higher intrinsic dimensionality indicates that the token representations occupy a richer and more diverse effective subspace, while a lower value suggests a more compressed or redundant representation structure.

Results. Figure 5 shows that intrinsic dimensionality provides a structural view of the fusion process. For the concatenation model Cambrian, the visual intrinsic dimension remains relatively low and stable through the early and transition ranges. This plateau suggests that the visual tokens initially preserve the compact structure inherited from the pretrained visual encoder. Around the middle-to-deep transition, the visual intrinsic dimension increases, indicating that visual representations expand into a richer effective subspace. The textual intrinsic dimension decreases from its input-side value and then changes more gradually, suggesting early reorganization of the language stream followed by stabilization.

For the native model EVE, the visual intrinsic dimension begins to increase from earlier layers and does not exhibit the long shallow-layer plateau observed in concatenation models. The textual intrinsic dimension changes together with the visual side and can peak around the transition before decreasing in later layers. Therefore, native models are better characterized as performing early

feature-space expansion and reorganization, rather than strictly monotonic expansion across all layers.

3.5 Causal Intervention Validation

The preceding analyses identify recurring internal patterns, but intervention experiments are needed to test whether those patterns are functionally relevant. We therefore conduct two complementary causal validations. The first masks sparse textual anchors selected from attention-concentrated text tokens and evaluates object hallucination behavior on POPE (Li et al., 2023b). The second injects Gaussian noise into early-layer or late-layer hidden states and evaluates general multimodal question answering on MMBench using VLMEvalKit (Liu et al., 2024b; Duan et al., 2024). Early/Late denote layers 0–7/16–32 in Figure 1; POPE scores are F1 (%), and all drops are in percentage points.

Table 1: Textual-target masking on POPE (F1, %).

Model	Original	Masked	Drop
LLaVA-Vicuna	87.5	44.6	42.9⇓
LLaVA-Mistral	86.4	39.9	46.5⇓
Cambrian	86.4	36.3	50.1⇓
EVE	85.0	53.8	31.2⇓
SOLO	78.6	43.8	34.8⇓

SAIL is omitted due to its hybrid attention mechanism.

Table 2: Layer perturbation on MMBench General VQA.

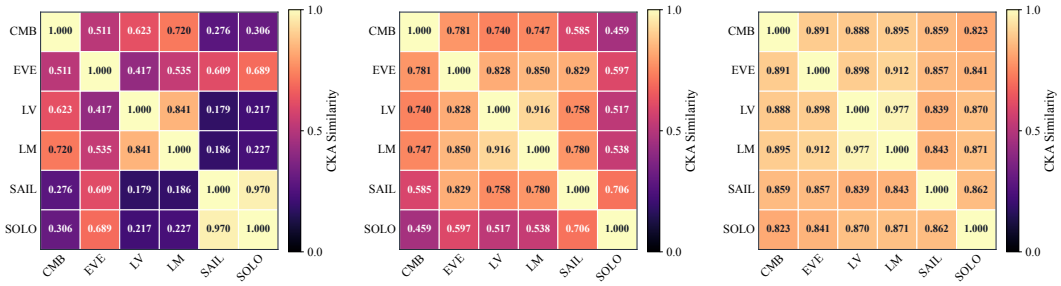
Model	Orig.	Early	E.drop	Late	L.drop
LLaVA-Vicuna	66.5	53.1	13.4⇓	46.7	19.8⇓
LLaVA-Mistral	66.5	53.4	13.1⇓	47.0	19.5⇓
Cambrian	75.9	67.0	8.9⇓	58.0	17.9⇓
SAIL	70.1	49.5	20.6⇓	58.6	11.5⇓
EVE	52.3	29.9	22.4⇓	38.6	13.7⇓
SOLO	67.7	43.6	24.1⇓	56.8	10.9⇓

Table 1 shows that concatenation models suffer larger POPE drops when concentrated textual targets are masked. This supports the attention-routing interpretation that their visual stream relies on a small set of textual targets during the transition stage. Table 2 shows the complementary MMBench pattern: native models are more vulnerable to early-layer noise, whereas concatenation models are more vulnerable to late-layer noise. This is consistent with the proposed pathways: native architectures begin visual-textual co-adaptation early, while concatenation architectures defer substantial visual reshaping to later transition layers.

3.6 Mechanistic Summary

The three progressive analyses provide an explanation for the fusion. Alignment decoupling shows that concatenation architectures follow a sequential pathway: textual representations adapt earlier, while visual representations remain relatively stable in shallow layers and are reshaped later. Within layers 8–15, attention routing shows that concatenation models allocate substantial visual-query attention to textual keys, with lower normalized entropy in this direction than in the reverse direction. Intrinsic dimensionality reveals the structural nature of this process: the later deformation of visual representations corresponds to delayed feature-space expansion. Therefore, concatenation models can be summarized as following a text-first, vision-later fusion pathway.

Native architectures exhibit a different pattern. Their visual representations change from earlier layers, and textual representations also evolve with depth. The attention summaries show model-specific directional asymmetries, including more diffuse visual-to-text attention in EVE and SOLO than in the concatenation models, with SAIL forming a structural exception documented in Appendix A.3. Their intrinsic dimensionality curves suggest earlier feature-space expansion and reorganization, and their stronger sensitivity to early perturbation supports the view that fusion begins earlier. Thus, the native-model results support earlier visual-textual co-adaptation while also revealing heterogeneous attention profiles. Separately, the causal intervention results support the functional relevance of these two pathway interpretations.



(a) Layer 1. (b) Layer 15. (c) Layer 30.

Figure 6: Cross-model visual CKA heatmaps at shallow, transition, and deep layers. Each cell measures the similarity between the visual representation spaces of two MLLMs at the same layer. CMB, LV, and LM denote Cambrian, LLaVA-Vicuna, and LLaVA-Mistral. Additional selected layers are reported in Appendix A.5.

4 Deep Visual Representation Convergence

To further examine whether different MLLMs converge toward similar high level representations, we conduct a supplementary analysis from the perspective of representation similarity. This experiment is motivated by the Platonic Representation Hypothesis (Huh et al., 2024), which suggests that representations learned by different neural networks may gradually converge toward structurally similar latent spaces. Here, we use layer-wise visual CKA similarity to test whether such a convergence trend can be observed across MLLMs with different architectures.

Method. We compute pairwise CKA similarity between the visual representation spaces of different models:

$$\mathcal{S}_{ij}^{(l)} = \mathrm{CKA}\left(Z_i^{(l)}, Z_j^{(l)}\right),$$

where $Z_i^{(l)}$ and $Z_j^{(l)}$ denote the stacked visual representations of two models at layer l. This produces a model-by-model visual CKA matrix for each layer.

Results. As shown in Figure 6, visual representations are highly architecture-dependent at layer 1. Models with similar visual processing strategies show relatively high similarity, while cross-paradigm similarities remain low. This indicates that shallow visual representations are mainly determined by each model's visual encoder, tokenizer, and architectural design. In contrast, the deep-layer heatmap at layer 30 shows substantially higher similarity across almost all model pairs. This suggests that although different MLLMs start from different visual representation spaces, their deep visual representations gradually move toward a more similar semantic structure.

These results provide empirical evidence consistent with the Platonic Representation Hypothesis in the context of MLLMs. From the perspective of representation similarity, multimodal fusion appears to reduce architecture-specific differences and encourages different models to form more aligned high-level visual representations.

5 Conclusion

The main contribution of this work is to uncover internal fusion mechanisms and design experiments to validate them. The findings can suggest several concrete directions. First, transition layers and sparse textual anchors can be used as robustness diagnostics, especially for studying adversarial vulnerability and hallucination behavior. Second, training strategies for concatenation models may benefit from earlier cross-modal feedback, anchor regularization, or layer-selective adaptation around the transition band. Finally, layer-selective fine-tuning, adapters, or LoRA modules can be guided by the observed pathway differences instead of being placed uniformly across all layers.

AI USE STATEMENT

Generative AI tools were used to assist with manuscript editing, LaTeX formatting. We did not use generative AI tools to generate synthetic datasets, develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proofs, write proofs, propose or refine hypotheses, design the core experiments, clean or reformat datasets, or interpret experimental results. All AI-assisted text, figures, tables, code edits, citations, and formatting changes were reviewed by the authors, who take full responsibility for the final content of this work.

REFERENCES

Samira Abnar and Willem Zuidema. Quantifying attention flow in transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4190–4197. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.acl-main.385.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Antonio Almudévar, José Miguel Hernández-Lobato, Sameer Khurana, Ricard Marxer, and Alfonso Ortega. Aligning multimodal representations through an information bottleneck. *arXiv preprint arXiv:2506.04870*, 2025.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pp. 2425–2433, 2015.

Yangyi Chen, Xingyao Wang, Hao Peng, and Heng Ji. Solo: A single transformer for scalable vision-language modeling. *arXiv preprint arXiv:2407.06438*, 2024.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D Manning. What does bert look at? an analysis of bert's attention. In *Proceedings of the 2019 ACL workshop BlackboxNLP: analyzing and interpreting neural networks for NLP*, pp. 276–286, 2019.

Haiwen Diao, Yufeng Cui, Xiaotong Li, Yueze Wang, Huchuan Lu, and Xinlong Wang. Unveiling encoder-free vision-language models. *Advances in Neural Information Processing Systems*, 37: 52545–52567, 2024.

Haodong Duan, Junming Yang, Yuxuan Qiao, Xinyu Fang, Lin Chen, Yuan Liu, Xiaoyi Dong, Yuhang Zang, Pan Zhang, Jiaqi Wang, et al. Vlmevalkit: An open-source toolkit for evaluating large multi-modality models. In *Proceedings of the 32nd ACM international conference on multimedia*, pp. 11198–11201, 2024.

Wanlong Fang, Tianle Zhang, and Alvin Chan. To align or not to align: Strategic multimodal representation alignment for optimal performance. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 21056–21064, 2026.

Fabian Gröger, Shuo Wen, and Maria Brbić. Revisiting the platonic representation hypothesis: An aristotelian view. *arXiv preprint arXiv:2602.14486*, 2026.

Lisa Anne Hendricks, John Mellor, Rosalia Schneider, Jean-Baptiste Alayrac, and Aida Nematzadeh. Decoupling the role of data, attention, and losses in multimodal transformers. *Transactions of the Association for Computational Linguistics*, 9:570–585, 2021.

Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. The platonic representation hypothesis. *arXiv preprint arXiv:2405.07987*, 2024.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International conference on machine learning*, pp. 3519–3529. PMLR, 2019.

Hyunji Lee, Danni Liu, Supriti Sinhamahapatra, and Jan Niehues. How do multimodal foundation models encode text and speech? an analysis of cross-lingual and cross-modal representations. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pp. 600–610, 2025.

Weixian Lei, Jiacong Wang, Haochen Wang, Xiangtai Li, Jun Hao Liew, Jiashi Feng, and Zilong Huang. The scalability of simplicity: Empirical analysis of vision-language learning with a single transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 20758–20769, 2025.

Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705, 2021.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pp. 19730–19742. PMLR, 2023a.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pp. 292–305, 2023b.

Victor Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Y Zou. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. *Advances in Neural Information Processing Systems*, 35:17612–17625, 2022.

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. In *European conference on computer vision*, pp. 740–755. Springer, 2014.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llavanext: Improved reasoning, ocr, and world knowledge, 2024a.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pp. 216–233. Springer, 2024b.

Jianglin Lu, Hailing Wang, Yi Xu, Yizhou Wang, Kuo Yang, and Yun Fu. Representation potentials of foundation models for multimodal alignment: A survey. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 16680–16695, 2025.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Jason Phang, Haokun Liu, and Samuel Bowman. Fine-tuned transformers show clusters of similar representations across layers. In *Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pp. 529–538, 2021.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Chonghao Sima, Katrin Renz, Kashyap Chitta, Li Chen, Hanxue Zhang, Chengen Xie, Jens Beißwenger, Ping Luo, Andreas Geiger, and Hongyang Li. Drivelm: Driving with graph visual question answering. In *European conference on computer vision*, pp. 256–274. Springer, 2024.

Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. Flava: A foundational language and vision alignment model. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 15638–15650, 2022.

Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai C Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, et al. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *Advances in Neural Information Processing Systems*, 37:87310–87356, 2024.

Tao Tu, Shekoofeh Azizi, Danny Driess, Mike Schaekermann, Mohamed Amin, Pi-Chuan Chang, Andrew Carroll, Charles Lau, Ryutaro Tanno, Ira Ktena, et al. Towards generalist biomedical ai. *Nejm Ai*, 1(3):AIoa2300138, 2024.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. A survey on multimodal large language models. *National Science Review*, 11(12):nwae403, 2024.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9556–9567, 2024.

Zhi Zhang, Srishti Yadav, Fengze Han, and Ekaterina Shutova. Cross-modal information flow in multimodal large language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 19781–19791, 2025.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023.

A TECHNICAL APPENDICES AND SUPPLEMENTARY MATERIAL

Appendix A.1 reports the Mutual KNN analysis, which verifies that the global CKA trend is also reflected in local neighborhood structures. Appendix A.2 provides the individual-model visual and textual CKA curves underlying the alignment-decoupling summary. Appendix A.3 presents the complete attention routing and cross-modal entropy results for all evaluated models. Appendix A.4 provides additional intrinsic-dimensionality curves to further support the observed feature-space expansion patterns. Appendix A.5 reports cross-model visual CKA heatmaps at intermediate layers, showing that deep-layer representation convergence develops progressively across depth rather than appearing only at the deepest displayed layer. Appendix A.6 reports the bootstrap robustness analysis.

A.1 MUTUAL KNN

The global CKA curves in the introduction reveal the overall layer-wise similarity trend between visual and textual representations. To further verify that this trend is not only reflected by kernel-level similarity, we additionally evaluate the local neighborhood consistency of representations using a Mutual KNN analysis (Gröger et al., 2026). For each model and each layer, we compute the nearest-neighbor structure of visual and textual representations under neighborhood size $k = 10$. Given a representation $z_i^{(l)}$ at layer l, we obtain its k nearest neighbors and compare neighborhood consistency across layers and modalities. Higher Mutual KNN consistency indicates that the local semantic structure of representations is more stable or better aligned.

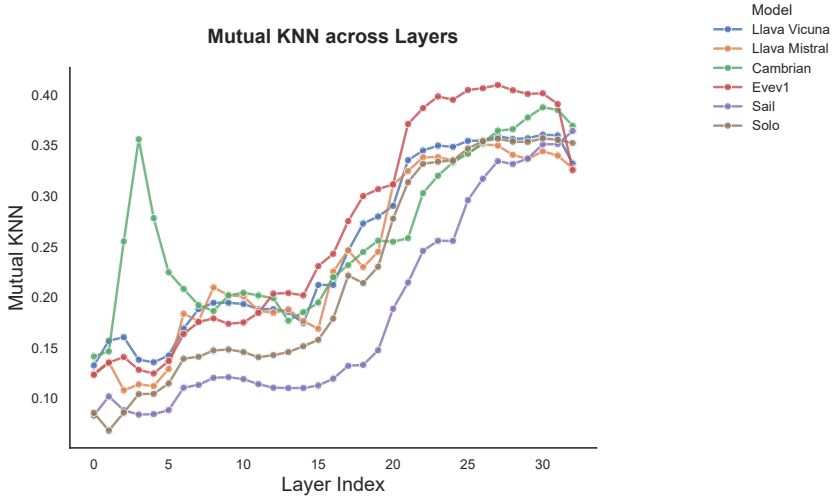
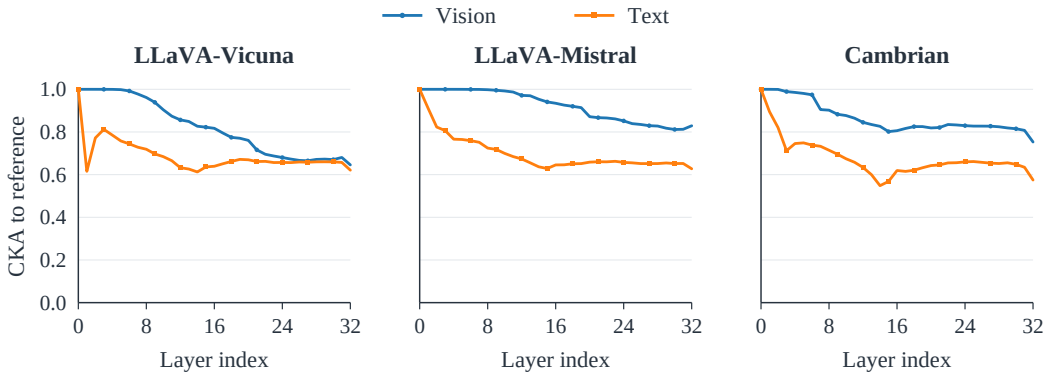


Figure 7: Supplementary Mutual KNN results. The observed trends are consistent with the global CKA analysis, suggesting that the layer-wise alignment pattern is also reflected in local neighborhood structure.

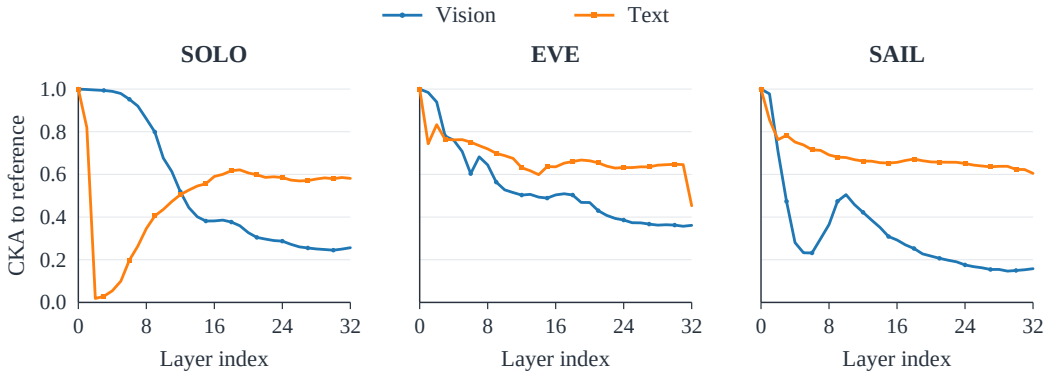
As shown in Figure 7, the Mutual KNN results generally follow the same trend as the global CKA curves. Concatenation models exhibit a more obvious transition in the middle layers, while native models show more continuous representation evolution. This supports the reliability of the global CKA observation in the main text.

A.2 PER-MODEL INTRA-MODALITY CKA

Figure 8 provides the detailed visual and textual trajectories underlying the family-level summary in Figure 3. The two panels separate concatenation and native models and retain each model's individual curves relative to the initial reference representation. The detailed panels use the same zero-based layer indices (0–32) as Figure 3, with the initial recorded representation at index 0 as the reference.



(a) Concatenation models: LLaVA-Vicuna, LLaVA-Mistral, and Cambrian.



(b) Native models: SOLO, EVE, and SAIL.

Figure 8: Per-model intra-modality CKA relative to the initial reference representation. Blue and orange curves denote visual and textual representations, respectively; both use solid lines. Layer indices are zero-based (0–32), consistent with Figure 3.

Visual representations remain closer to the reference in the shallow layers of the concatenation models, whereas the native models show stronger early visual changes. Textual trajectories vary across models, including a pronounced early decrease followed by recovery in SOLO. The individual curves therefore do not imply synchronized or monotonic changes in both modalities.

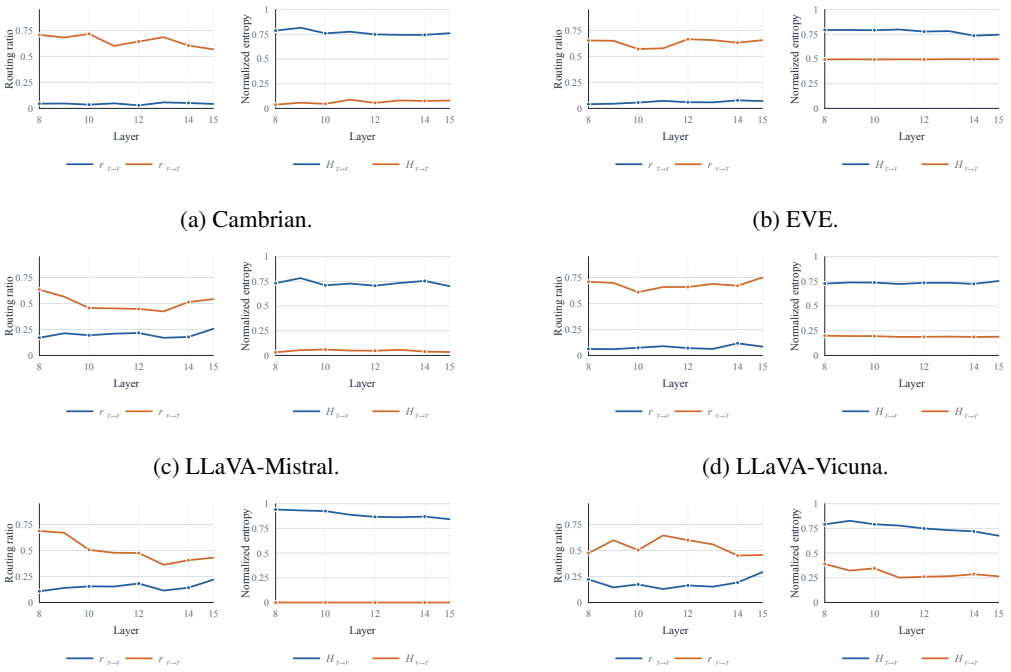
A.3 Attention Routing and Cross-modal Entropy

The main text compares attention distributions and directional median gaps across five models over layers 8–15, omitting SAIL because of its constrained attention path. Here, we provide the corresponding layer-wise routing and normalized-entropy summaries for all six models, including SAIL.

The layer-wise profiles in Figure 9 complement the distribution summaries in Figure 4. Concatenation models combine relatively high visual-to-text routing with low entropy in that direction across the analyzed range. Native models exhibit heterogeneous attention profiles, including more diffuse visual-to-text attention in EVE and SOLO and a constrained visual-to-text entropy path in SAIL. These curves characterize the sampled layer range; the onset of fusion is examined through the full-depth representation analyses.

A.4 Intrinsic Dimension

The intrinsic dimensionality analysis in the main text characterizes whether multimodal fusion corresponds to feature-space compression, expansion, or reorganization. Here, we provide the original per-model visual and textual curves for additional models, followed by a threshold-sensitivity analysis for the two representative models used in the main paper.



(a) Cambrian.

(b) EVE.

(c) LLaVA-Mistral.

(d) LLaVA-Vicuna.

(e) SAIL.

(f) SOLO.

Figure 9: Attention routing ratios and per-sample normalized cross-modal entropy for all evaluated models over layers 8–15. SAIL has zero visual-to-text entropy because its hybrid attention path constrains that direction.

As shown in Figure 10, concatenation models generally show delayed visual dimensionality expansion, while native models show earlier and more continuous feature-space growth. This is consistent with the main observation that concatenation models follow a text-first, vision-later fusion pathway, whereas native models perform earlier modality co-adaptation. Each model's plot contains multiple curves corresponding to different variance retention thresholds for the sensitivity analysis. The trends of all models remain consistent across varying thresholds, indicating that the observed patterns are not caused by a specific variance cutoff.

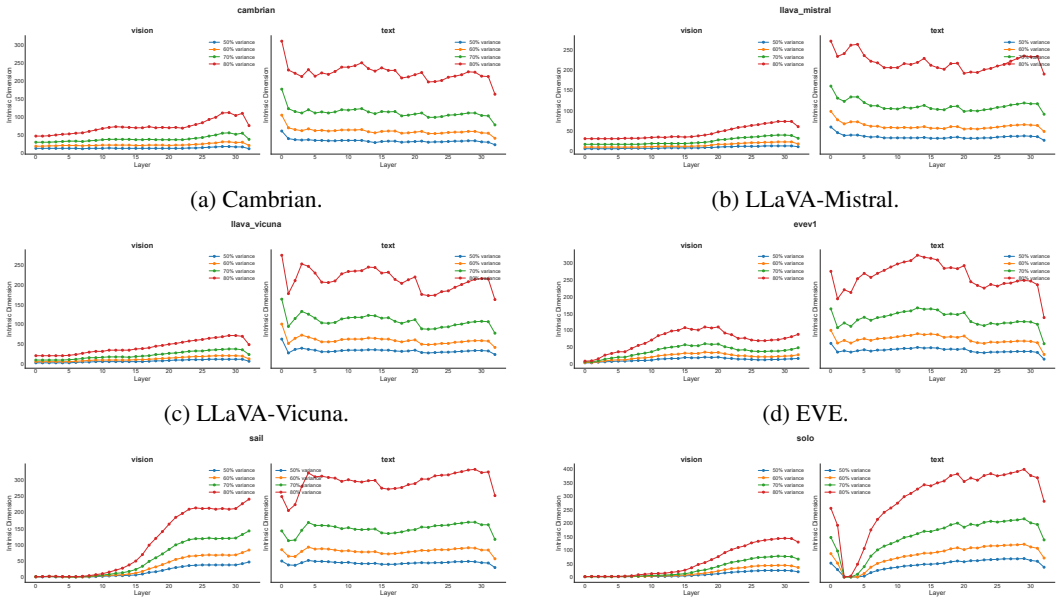
A.5 Cross-model Visual Representation Similarity

To further examine representation convergence across different MLLMs, we compute pairwise visual CKA similarity between models at shallow and deep layers. This supplementary experiment is motivated by the Platonic Representation Hypothesis (Huh et al., 2024): from the perspective of representation similarity, if different MLLMs gradually converge toward similar high-level visual semantics, their deep-layer visual representations should become more aligned.

As shown in Figure 11, visual representations are strongly architecture-dependent in shallow layers. Models with similar visual processing strategies show higher similarity, while cross-paradigm similarities remain relatively low. In contrast, deeper-layer heatmaps show substantially higher similarity across model pairs. This result suggests that although different MLLMs start from different visual representation spaces, their deep visual representations become more aligned after multimodal fusion. Therefore, this supplementary analysis provides representation-level evidence consistent with the Platonic Representation Hypothesis in the context of MLLMs.

A.6 Bootstrap Robustness Analysis

To quantify sampling uncertainty and evaluate robustness, we perform non-parametric bootstrap resampling. We use a 5,000-pair analysis population, define four sample-size groups of 500, 1,000,



(a) Cambrian.

(b) LLaVA-Mistral.

(c) LLaVA-Vicuna.

(d) EVE.

(e) SAIL.

(f) SOLO.

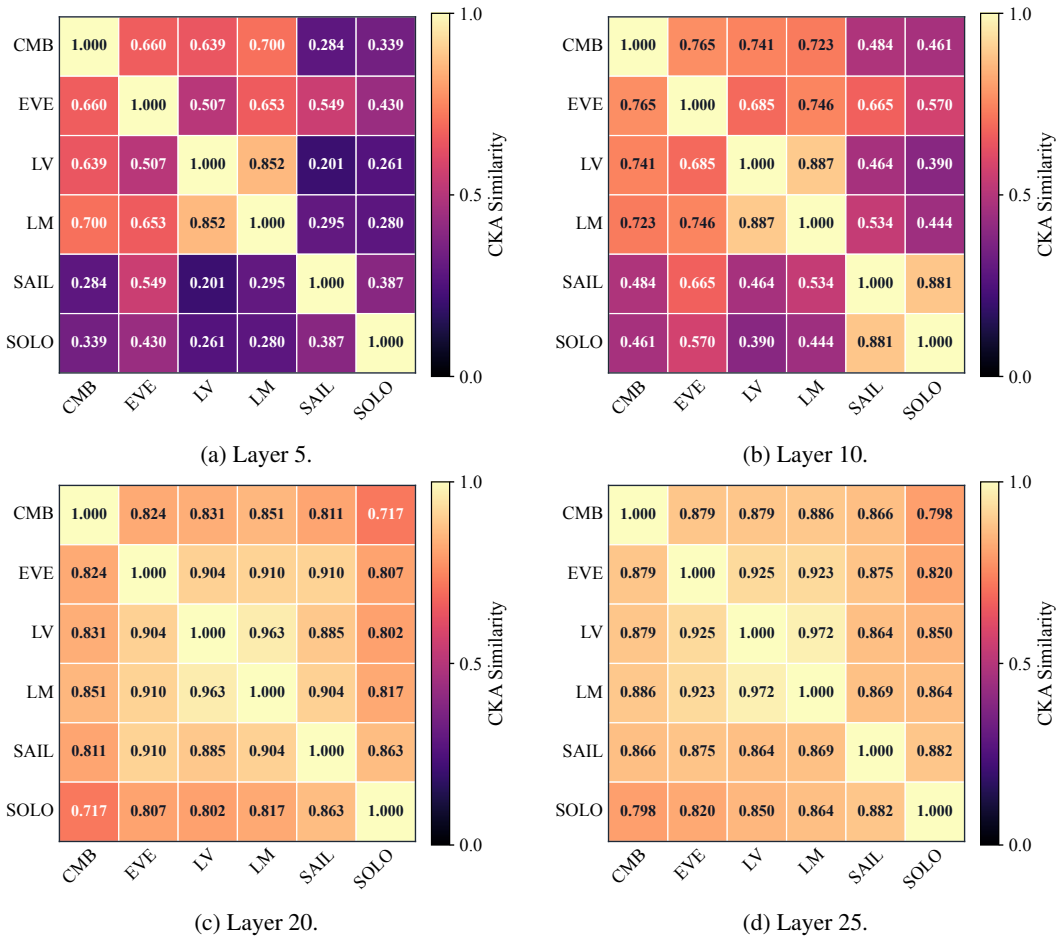
Figure 10: Supplementary intrinsic dimensionality results for additional models. Each panel uses the same layer-wise protocol as the main text and reports multiple variance preservation thresholds.

Table 3: Bootstrap robustness analysis for Cambrian in the Figure 1 experiment.

Layer / sample size	Bootstrap mean	Standard error	Relative bias (%)
Layer 1 / 500	0.4679	0.0014	8.58
Layer 1 / 1000	0.4488	0.0012	4.15
Layer 1 / 2000	0.4384	0.0009	1.73
Layer 1 / 5000	0.4309	0.0005	0.00
Layer 15 / 500	0.3844	0.0013	12.63
Layer 15 / 1000	0.3616	0.0011	5.94
Layer 15 / 2000	0.3489	0.0008	2.24
Layer 15 / 5000	0.3413	0.0004	0.00
Layer 30 / 500	0.7540	0.0011	1.88
Layer 30 / 1000	0.7475	0.0008	1.00
Layer 30 / 2000	0.7431	0.0006	0.40
Layer 30 / 5000	0.7401	0.0004	0.00

2,000, and 5,000 pairs, and conduct 1,000 independent random resamplings with replacement for each group, using image-text pairs as the resampling unit. Layers 1, 15, and 30 are selected as representative layers, and the bootstrap mean under the 5,000-sample condition is used as the reference estimate for computing relative bias. For layer l, relative bias (%) is $100(\bar{c}_{n,l} - \bar{c}_{5000,l})/\bar{c}_{5000,l}$, where $\bar{c}_{n,l}$ is the bootstrap mean for sample size n. It measures deviation from the 5,000-pair reference estimate, rather than bias relative to an unknown population CKA value. All stochastic processes use a fixed random seed. Figure 12 visualizes the relative-bias trend, while Table 3 reports the exact Cambrian CKA values.

The standard errors remain small across all sample sizes, and the relative bias decreases as the sample size increases. This indicates that the observed shallow-layer, transition-layer, and deep-layer CKA values are stable under bootstrap resampling.



(a) Layer 5. (b) Layer 10.

(c) Layer 20. (d) Layer 25.

Figure 11: Cross-model visual CKA heatmaps at selected appendix layers. A shared color scale and model order make the progressive increase in cross-model visual similarity easy to compare; lighter yellow cells indicate higher similarity.

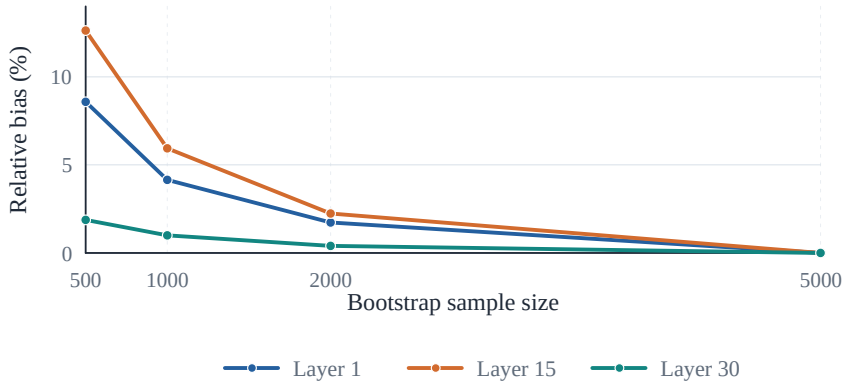


Figure 12: Bootstrap relative bias for Cambrian CKA at representative shallow, transition, and deep layers. The 5,000-sample estimate is the reference point for each layer.