HOW NATIVE ARE NATIVE VISION-LANGUAGE MODELS?

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision-language models (VLMs) have typically been built from a vision encoder, a language model, and a connector between them. Native VLMs remove these boundaries by removing the encoder and connector. How vision and language work together is learned rather than prescribed. The model must now simultaneously learn both to build visual features and to let language use them. How native models construct and use visual features remains largely under-examined. Inside public native VLMs, we find that text draws on visual features mainly in the middle layers, even when they form much earlier. In controlled training, a backbone can learn visual features yet barely change its text predictions when the image is replaced, a dissociation reminiscent of ***optic aphasia***, in which patients recognize objects but cannot name them. Through an external encoder and connector, the same features support far stronger image dependence, locating the bottleneck largely in the connection. We therefore introduce the adaptive visual reader, a connector learned inside the backbone, which raises caption precision from 20.7% to 37.0%, closing about two-thirds of the gap to a matched external-encoder model at 45.1%, while leaving visual features largely unchanged. These results suggest that native VLMs should be trained and evaluated with visual encoding and visual reading as separate objectives, since strong visual features alone do not ensure that language uses them.

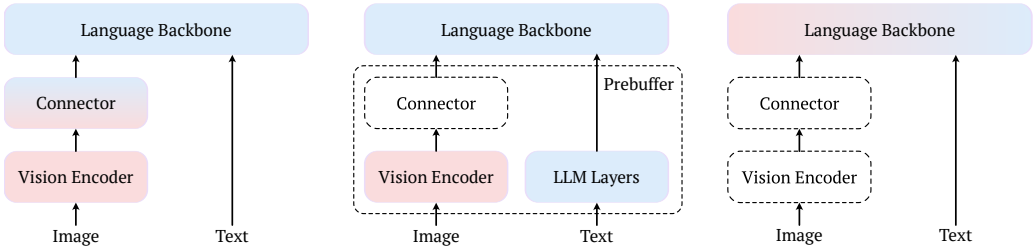
1 INTRODUCTION

Native vision-language models (VLMs) pursue a compelling goal: to learn vision and language within a single network (Chen et al., 2024; Diao et al., 2024). Image patches and words pass through the same backbone, allowing visual representations and language processing to develop within a common computation (Chen et al., 2024; Lei et al., 2025; Luo et al., 2025b; Diao et al., 2025). This opens the possibility of learning how the two modalities can work together with fewer boundaries prescribed by the architecture. Realizing that possibility requires understanding how a shared network learns to organize their interaction.

Conventional VLMs provide part of this organization by design (Liu et al., 2023b; Li et al., 2023a; Alayrac et al., 2022). A vision encoder builds image representations (Radford et al., 2021; Tschannen et al., 2025; Caron et al., 2021; Oquab et al., 2024; Siméoni et al., 2025), and a connector supplies them to a language model (Liu et al., 2023b; Li et al., 2023a; Alayrac et al., 2022). Removing these modules places more of the burden on learning (Figure 1). The backbone must both build visual features and learn how and when language should read them (Lei et al., 2025; Luo et al., 2025b; Diao et al., 2025). The second does not follow from the first. Information can be present in image states yet contribute little to the words the model predicts. Native multimodal learning therefore involves two abilities that must develop together: representing the visual world and connecting those representations to language.

In this paper, we examine how trained native VLMs (Lei et al., 2025; Diao et al., 2025; 2026a) organize the relationship between these two abilities. With layer-wise probes (Alain & Bengio, 2017; Hewitt & Manning, 2019) and attention interventions (Geva et al., 2023; Neo et al., 2025; Kaduri et al., 2025), we find that visual representations form across layers, while language draws on them mainly in the middle of the network. This pattern persists even when visual features are available much earlier. Architectural early fusion (Yang et al., 2025b; Diao et al., 2026a) therefore settles where visual features enter, not when language reads them.

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.



(a) External encoder (b) Native with a pre-buffer (c) Native without a pre-buffer

Figure 1: **Three ways to organize visual encoding and its connection to language.** (a) An external vision encoder builds image features, and a separate connector maps them into the language model. (b) Native designs with a pre-buffer use it for visual encoding and connect vision to language within the backbone. (c) Native designs without a pre-buffer learn both functions within the backbone.

This distinction becomes consequential during training. In our controlled experiments, next-token prediction alone leaves visual representations weak, consistent with the training difficulties reported in earlier work (Chen et al., 2024; Diao et al., 2024). Distilling a vision encoder into image-specific parameters gives the backbone stronger visual features, and continued distillation preserves them as the model learns to generate captions. Yet replacing the image barely changes the model's text predictions. The features are there, but language makes little use of them. This failure resembles ***optic aphasia***, a dissociation between visual recognition and naming (Riddoch & Humphreys, 1987; Beauvois, 1982). The same learned encoder, frozen and placed outside the language model, supports substantially stronger image dependence after caption training. Visual features that an external connector puts to work can sit largely unused inside a shared backbone.

To understand why, we analyze the attention updates through which image tokens reach the text (Basu et al., 2024). Visual messages carry class information, but their norms are small relative to the text states they update at the measured layers. Giving these messages their own output projection improves caption precision, while learned amplification alone brings little benefit. We call the pathway through which text states retrieve information from image tokens inside the backbone the *visual reader*; it plays the role that the connector plays in a conventional VLM. Guided by this finding, we design an **adaptive visual reader** (AVR) that reuses the backbone's attention and combines a dedicated visual output projection with text-conditioned gates. It raises caption precision from 20.71% to 37.03% and coverage from 26.20% to 36.61%, approaching a matched external-encoder model at 45.09% and 37.92%, with little change in visual probe accuracy. The improvement comes from making existing visual information more useful to language.

Native VLMs rest on the expectation that a single network can discover on its own how vision should serve language. Our findings show what this expectation requires. Vision and language do not fully organize themselves inside a shared backbone, and the training we study can build visual representations without teaching language to read them. Future native VLMs should therefore evaluate how language reads visual features, for example through image dependence and caption quality, alongside probes of the features themselves, and should give this reading pathway dedicated capacity rather than expect it to emerge from next-token prediction.

2 Related Work

Native vision-language models. Native modeling spans encoder-based models pretrained multimodally from the start and encoder-free architectures (Wu et al., 2026; Han et al., 2026; An et al., 2026). NaViL studies the relative sizes of visual and language components under joint training while retaining a vision encoder (Tian et al., 2025). Among encoder-free models, Fuyu, SOLO and SAIL use shared backbone parameters (Bavishi et al., 2023; Chen et al., 2024; Lei et al., 2025), while Mono-InternVL, EVEv2 and VoRA use image-specific parameters or low-rank updates (Luo et al., 2025b;a; Diao et al., 2025; Wang et al., 2025). NEO assigns visual encoding to a designated pre-buffer, a block before the pretrained language layers, before jointly training the full backbone (Diao et al., 2026a;c). These designs also extend to video and joint image understanding and generation (Li et al., 2025; Liu et al., 2026a; Diao et al., 2026d;b). We examine where these backbones encode images and where text reads the resulting features.

Learning visual representations. Visual encoders learn from image-text contrastive objectives in CLIP and SigLIP or caption prediction in CapPa and GenLIP (Radford et al., 2021; Zhai et al., 2023a; Tschannen et al., 2025; 2023; Fang et al., 2026). Distillation transfers features from one or more vision teachers, as in AM-RADIO, Theia and UNIC (Ranzinger et al., 2024b; Shang et al., 2024; Sariyildiz et al., 2024), with results depending on the projector, target normalization and layer supervision (Miles & Mikolajczyk, 2024; Ranzinger et al., 2024a; Sariyildiz et al., 2024). We compare parameters, initialization and losses for internal visual encoding, then test whether the learned features are retained and used during language training.

Visual information inside language models. Prior work examines where and how language models use visual information (Neo et al., 2025; Kaduri et al., 2025; Zhang et al., 2025b; Salazar et al., 2026; Basu et al., 2024), and how changing where visual features enter the backbone affects performance (Meng et al., 2024; Choraria et al., 2026). Visual information a model encodes may be absent from its answers (Fu et al., 2025; Liu et al., 2025). We study this gap in native VLMs through controlled encoder comparisons and attention analysis. Our reader learns an internal vision-to-language link, inspired by BLIP-2 and Flamingo (Li et al., 2023a; Alayrac et al., 2022).

3 Visual encoding and reading inside the backbone

Native VLMs allow text and image tokens to interact from the first layer, as illustrated in Figure 1. But does this architectural early fusion mean that visual encoding and reading develop together from the first layer? To answer this question, we examine where visual features form and when text reads them. *Visual encoding* builds visual features. *Reading* fuses visual information into text states through attention. Layer-wise probes measure how visual representations develop (Hewitt & Manning, 2019; Alain & Bengio, 2017), and attention interventions (Geva et al., 2023; Neo et al., 2025; Kaduri et al., 2025) identify the layers where the text reads the image.

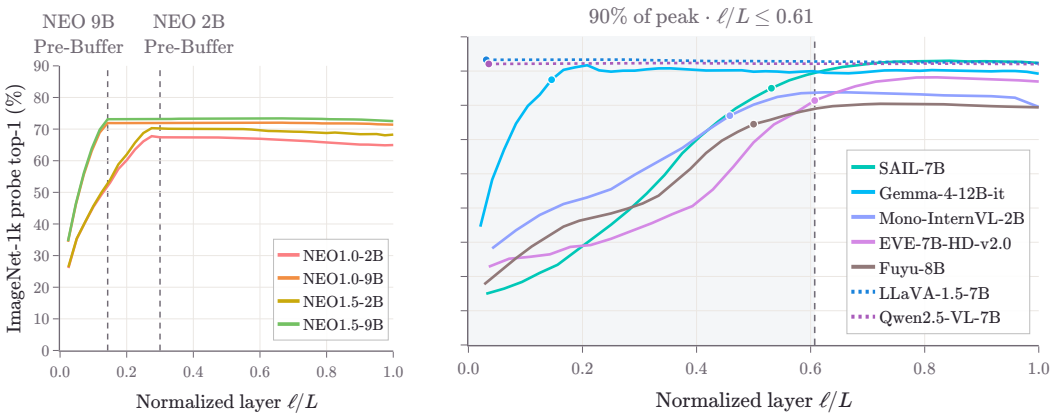
Models. We study five native checkpoints: SAIL-7B, Gemma-4-12B (Gemma Team, 2026), Mono-InternVL-2B, EVEv2-7B and Fuyu-8B (Lei et al., 2025; Luo et al., 2025b; Diao et al., 2025; Bavishi et al., 2023). For comparison, we include four NEO checkpoints from the NEO1.0 and NEO1.5 series at 2B and 9B, which assign visual encoding to a *pre-buffer*, a block placed before the pretrained language layers (Diao et al., 2026a;c). Two modular references, LLaVA-1.5-7B and Qwen2.5-VL-7B, receive features from an external vision encoder (Liu et al., 2023a; Bai et al., 2025). Appendix A gives the protocol details.

3.1 The backbone takes on the role of a vision encoder

A probe is an auxiliary model trained to predict a target from a frozen model's hidden states. Its performance measures how much information about that target can be read from those states (Alain & Bengio, 2017; Hewitt & Manning, 2019). We use probes to track the development of visual features across layers. At each backbone layer, we take the hidden states at image-token positions and train an attentive probe on ImageNet (Deng et al., 2009). Figure 2 plots top-1 accuracy against normalized depth ℓ/L, where L is the number of backbone layers.

Explicit visual encoding in NEO. NEO assigns visual encoding to a pre-buffer before the pretrained language layers (Figure 1(b)). Its depth is chosen in advance based on model size and computational cost, rather than learned during training (Diao et al., 2026a;c). NEO first trains the pre-buffer while keeping the pretrained LLM weights frozen, followed by joint training of the full model during mid-training and supervised fine-tuning. Figure 2(a) shows that probe accuracy rises steeply inside the pre-buffer and changes little afterward. Most visual encoding therefore occurs within the block assigned to this role. Although NEO removes the separate connector, it retains an explicitly designated encoding block inside the backbone, giving it an encoder-like division of computation.

Implicit visual encoding in other native models. The other five native models have no predefined encoder block (Figure 1(c)). Image and text tokens pass through the same layers, with either shared or modality-specific parameters. Figure 2(b) shows how visual features develop within these layers, with some models reaching probe accuracies comparable to the external-encoder references. Most



(a) NEO series

(b) Native models and external-encoder references

Figure 2: **Visual features form at different depths.** ImageNet-1K attentive-probe accuracy against normalized depth ℓ/L. (a) NEO models build their features almost entirely inside the pre-buffer, whose end is marked by vertical lines. (b) Native models without a pre-buffer (solid) compared with VLMs that use an external vision encoder (dashed). In (b), dots mark where each native model first reaches 90% of peak accuracy; all five do so by middle layers, the shaded region's right edge.

build visual features gradually, reaching 90% of peak probe accuracy between $\ell/L = 0.46$ and 0.61. Gemma-4-12B reaches this point much earlier, at 0.15 (Figure 2(b)). Across all five models, the visual features are thus largely formed by the middle of the backbone, even without a block explicitly assigned to visual encoding.

3.2 Text reads the image in the middle layers

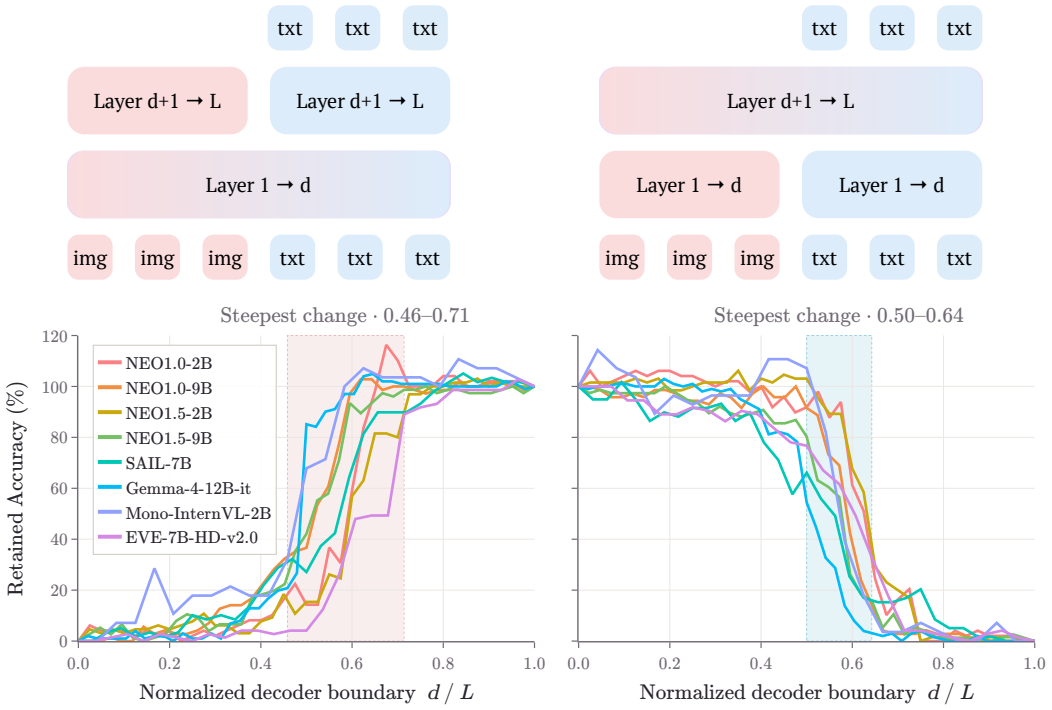
The probes show where visual features develop. We next ask where the text reads them. We use attention knockout (Geva et al., 2023; Neo et al., 2025; Kaduri et al., 2025) to block text-to-image attention over a range of layers and measure the change in Multimodal Visual Patterns (MMVP) accuracy (Tong et al., 2024). The cuts leave image-token processing unchanged and retain text-to-text attention, testing whether text predictions depend on image access in the selected layers. MMVP evaluates basic visual perception, testing whether models can distinguish visual details in images that CLIP encodes similarly. We report *retained accuracy*, which measures how much of the same model's original MMVP accuracy remains after blocking attention:

$$\text{Retained accuracy} = \frac{\text{MMVP accuracy after the cut}}{\text{MMVP accuracy without the cut}} \times 100\%.$$

We move a boundary layer d through the backbone and block attention on either side of it. The *suffix cut* (Figure 3(a)) keeps text-to-image attention in layers 1 to d and removes it above d. This tests *sufficiency*. If accuracy recovers by layer d, the model has obtained enough visual information by that layer to complete the task without further image access. The *prefix cut* (Figure 3(b)) removes text-to-image attention below d and keeps it from d to L. This tests *necessity*. If accuracy collapses once d passes a layer range, image access in later layers cannot compensate for the removed interaction. Together, the cuts locate the layers where the text reads the image. Appendix B gives the evaluation details.

Reading is concentrated in the middle layers. Figure 3 shows how accuracy changes as the boundary moves through the backbone. Under the suffix cut, accuracy rises most sharply when the retained image access extends into the middle layers. Under the prefix cut, accuracy falls most sharply when the removed image access reaches these layers. Together, these changes indicate that early access alone is insufficient, and later access cannot fully compensate for removing access in the middle layers. Across the eight intervention checkpoints, the steepest changes occur between $\ell/L = 0.46$ and 0.71, locating the main reading window in this region.

We next compare this reading window with the depth at which visual features form (Table 5). NEO and Gemma-4-12B reach 90% of peak probe accuracy early, at $\ell/L = 0.12$–0.23, but their main



(a) Suffix cut (b) Prefix cut

Figure 3: **Text tokens read the image in the middle layers.** The top diagrams illustrate (a) suffix cuts and (b) prefix cuts. Below, retained accuracy is MMVP strict-pair accuracy after a cut as a percentage of intact accuracy. (a) Keeping text-to-image attention through layer d tests *sufficiency*. (b) Removing it before layer d tests *necessity*. Both cuts show sharp changes in the middle layers. Shaded bands mark the range of steepest-transition depths across models.

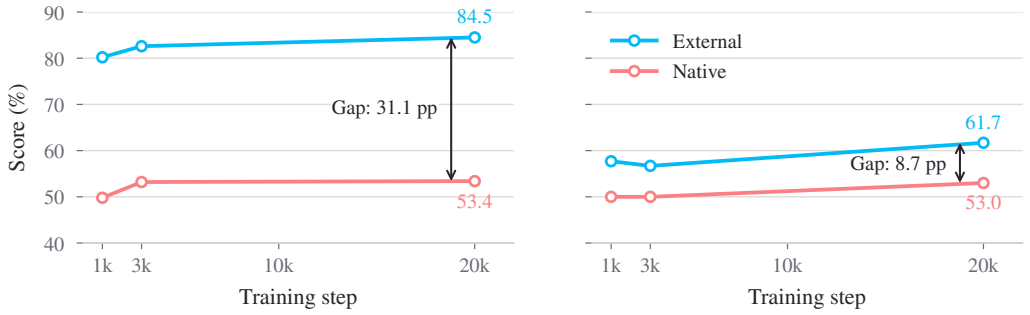
reading transitions occur at 0.48–0.63. In SAIL, Mono-InternVL and EVEv2, visual features form later, and the reading transitions overlap with or follow their formation. Thus, early visual encoding does not necessarily lead to early reading: visual features can be available well before the text makes its main use of them. In NEO, this separates encoding in the pre-buffer from reading in the subsequent language layers.

Takeaway 1. Early fusion does not mean early reading.
Text reads the image mainly in the middle layers, even when visual features form early. Architectural early fusion does not make encoding and reading develop together from the first layer.

4 Learning Visual Encoding Inside the Language Backbone

Having examined where visual encoding and reading occur in trained native VLMs, we next ask whether learning visual representations is sufficient for language to use them. In this section, we train and preserve visual representations inside the backbone. Section 5 then tests whether the language model uses these representations.

We begin with direct native training. In a Qwen3-1.7B pilot, we train a native model and an external-encoder reference with the same backbone and schedule on a mixture of FineVision (Wiedmann et al., 2025) and LLaVA-OneVision-1.5 data (An et al., 2025). The native model receives image patches, while the reference receives features from a pretrained CLIP encoder (Radford et al., 2021). The native model improves, but its gap to the reference does not shrink. After 20k steps it scores 53.4 on POPE (Li et al., 2023b), below the reference's 80.2 after 1k steps, and per-item MMVP shows the same pattern (Figure 4). EVE reports similar difficulties (Diao et al., 2024), and probing our directly trained native model gives around 13% accuracy, indicating weak visual representations



(a) POPE (b) MMVP

Figure 4: **Direct native training leaves a gap on both POPE and MMVP.** The Qwen3-1.7B pilot uses the same language backbone, data and schedule for both image pathways. Curves show scores at 1k, 3k and 20k steps. Gap labels give the percentage-point difference at 20k steps. MMVP uses the pilot's per-item metric. Appendix C also reports VMCBench and image-swap results.

under this probe. We therefore supervise the image-token states directly with feature distillation, and ask which parameters and objective learn these features best and how to keep them during language training. These experiments use a Qwen3-0.6B backbone (Yang et al., 2025a) and a SigLIP2 NaFlex teacher (Tschannen et al., 2025), with training details in Appendix D.

4.1 Learning visual features while preserving language ability

We first ask which backbone parameters can learn visual features while preserving pretrained language ability. Unfreezing all backbone parameters during visual distillation gives the strongest visual features, with 80.29% ImageNet accuracy, but severely degrades language ability. WikiText-2 negative log-likelihood (NLL) rises from 3.17 to 12.48 (Appendix K). This degradation is consistent with catastrophic forgetting (Zhai et al., 2023b). We therefore freeze the text weights and add a *visual branch*, a set of parameters used only by image tokens (Figure 7 in Appendix D). The branch spans every backbone layer, leaving the depth of visual encoding to be learned. We compare an image-specific feed-forward network (visual FFN) with image-specific attention projections (visual attention), initializing each from the matching language weights or at random. We evaluate ImageNet classification, ADE20K semantic segmentation (Zhou et al., 2018) and COCO image-text retrieval (Lin et al., 2014). Appendix D gives the evaluation metrics.

Visual attention outperforms visual FFN. Visual attention outperforms visual FFN on all evaluated metrics under both initialization settings (Table 1a), while training fewer parameters (178M compared with 266M). With language initialization, it reaches 69.52% ImageNet accuracy compared with 68.61% for visual FFN, and also performs better on segmentation and retrieval. Initializing from language weights improves every metric for both branches, providing a useful starting point for visual learning (Luo et al., 2025b). Visual attention also lets us study the connection to language by changing the image keys and values separately from the text output projection. We use it for the reader experiments in Section 5.2.

4.2 Distillation objective and data distribution

Feature supervision. Drawing on prior work on feature distillation of vision encoders (Miles & Mikolajczyk, 2024; Ranzinger et al., 2024a), we compare cosine, MSE and mixed objectives, including SmoothL1 and standardized-MSE variants (Table 1b). Cosine loss gives the best ImageNet accuracy, while cosine with a small SmoothL1 term gives the best retrieval results. Appendix E defines the losses, and Appendices F and N report the intermediate-layer supervision experiments and layer-wise probes.

Data distribution. We also vary the training data distribution, holding the 12.8M-image budget fixed (Table 1c). Alongside ImageNet-1K, we choose CC12M for its broader coverage of visual concepts (Changpinyo et al., 2021) and DCVLM-Baseline for exposure to images containing text,

Table 1: **Visual feature distillation.** Each block changes one factor and evaluates the image tokens of the final layer. All values are percentages, and bold pink values mark the best value within each block. (a) Branch type and initialization, with final-layer supervision. (b) Distillation loss for the language-initialized visual attention branch, with intermediate-layer supervision weight 0.2. (c) Data distribution for the language-initialized visual FFN under a 12.8M-image budget. Its ImageNet row is the same run as the corresponding row in (a).

	ImageNet	ADE20K	COCO retrieval R@1	
Setting	top-1	mIoU	image→text	text→image
(a) Branch and initialization				
Visual attention, language weights	**69.52**	**27.81**	**43.22**	**34.48**
Visual attention, random	67.20	25.03	41.34	32.36
Visual FFN, language weights	68.61	26.34	42.86	33.52
Visual FFN, random	66.17	23.89	38.72	31.11
(b) Distillation loss				
Cosine	**70.28**	**27.39**	44.96	35.74
0.9 cosine + 0.1 SmoothL1	70.15	27.37	**45.18**	**35.85**
0.9 cosine + 0.1 MSE	66.72	24.68	40.92	32.41
Standardized MSE	66.68	24.79	39.40	31.64
MSE	57.73	20.40	31.10	24.56
(c) Data distribution				
ImageNet-1K	**68.61**	26.34	42.86	33.52
CC12M	59.78	29.93	44.46	35.14
DCVLM-Baseline	49.25	24.53	42.16	33.98
All three mixed	66.39	**30.75**	**48.28**	**39.29**

including OCR and document data (Farina et al., 2026). We compare each source with a mixture of all three. The mixture gives the best segmentation and retrieval, while ImageNet alone gives the best ImageNet accuracy, benefiting from the match between training and evaluation distributions. Data diversity thus helps visual representations transfer beyond ImageNet classification.

4.3 Encoding must be maintained during language training

We next train the distilled model to generate captions using image-caption pairs from MetaCLIP (Xu et al., 2024), updating both the visual branch and the language weights, and track its visual features with probes (Appendix G). With caption CE alone, accuracy falls from 66.17% to 11.24% within the first 1k steps and ends at 13.57% (Figure 5). To preserve these features, we keep the distillation loss active during training. With continued distillation, accuracy ends at 64.83%, 98% of the starting accuracy. Continued distillation lets us study reading with the learned visual features still in place. In Section 5, we ask whether the text can use them.

Takeaway 2. Distillation builds and preserves visual features.
Distillation teaches visual branches to encode images. During caption training, CE alone degrades these features, while continued distillation largely preserves them for the language model to use.

5 Visual encoding does not ensure effective reading

5.1 The reading gap

With visual features preserved through continued distillation, we now ask whether the language model uses them. We measure image dependence as the *image gap*, the increase in caption negative log-likelihood (NLL) when the image is replaced. We also evaluate generated captions with CAPEval (Liu et al., 2026b), which checks them against a checklist of image facts. Coverage is the fraction of

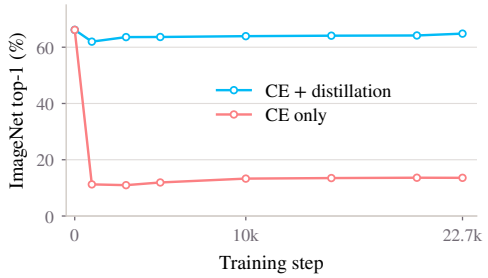


Figure 5: **Distillation preserves visual features.** ImageNet probe accuracy during caption training with CE alone or continued distillation.

	Native	Same path	
		frozen	trained
Probe ↑	64.83	66.17	52.50
Caption NLL ↓	1.895	1.848	1.856
Image gap ↑	+0.042	+0.235	+0.178

Table 2: **An external connection improves image use.** The same distilled encoder is used inside or outside the backbone. Protocols are in Appendix G.

checklist facts a caption mentions, and precision is the fraction of mentioned facts that are correct. Appendix G gives the protocols.

The same features are more useful through an external connection. The native model retains 64.83% probe accuracy, yet replacing its image changes caption NLL by only +0.042 (Table 2). Is this because the learned visual features are inadequate, or because language does not use them effectively? We freeze the distilled visual encoder and connect it to the same language model as an external encoder. After training on the same captions, its image gap reaches +0.235. The same learned encoder thus supports stronger image dependence through an external connection.

Generated captions show the same gap. For the native backbone used in the reader experiments below, caption precision is 20.71%, compared with 45.09% for its matched external reference under a shared protocol (Table 3).

We call this failure *optic aphasia*, by analogy to the neurological condition in which patients recognize an object by sight but cannot name it (Riddoch & Humphreys, 1987; Beauvois, 1982). The model retains visual features but makes little use of them in generation. Related gaps between available and used visual information have been reported in modular VLMs (Fu et al., 2025; Liu et al., 2025).

Visual messages are too weak to shape text states. To locate the reading problem, we separate the attention update into contributions from image and non-image tokens (Basu et al., 2024). We call the image contribution the *visual message*. These messages carry class information comparable to the matched external reference at the measured layers, but are much smaller relative to the text states. In a separate diagnostic, scaling the message improves text-state class accuracy. Whether amplification also improves generation is tested below. Appendices I and L give the decomposition, scaling experiment and layer-wise measurements.

5.2 An adaptive visual reader

Reading needs its own output projection. Attention uses queries and keys to select image tokens, values to carry their content, and the output projection W_O to write the result into the text state. In this backbone, image keys and values have their own parameters, but attention updates at text positions pass through the language output projection. Table 3 tests three possible changes to this connection on the same distilled backbone and training data. Learned gains test whether a larger visual message suffices, separate key/value projections whether the reader must select or represent image content differently, and a separate output projection whether the message needs its own mapping into text states. For each variant, the probe checks that visual features are kept, the image gap measures image dependence, and CAPEval measures caption quality.

Learned gains leave caption precision near the control, at 20.03% versus 20.71%. Separate keys and values reach only 21.44%, whereas a separate output projection reaches 37.69% (Table 3, first block). The K/V and output-projection variants add the same number of parameters, and their probe accuracies remain close to the control. The comparison points to the mapping into text states as a source of the reading gap. Appendix J includes a reader with both separate K/V and W_O, which brings no further improvement.

Table 3: **Connection ablations and adaptive visual reading.** Check marks denote additional reader projections, with K/V for keys and values and W_O for the output. Unmarked entries reuse backbone projections. Probe, coverage and precision are percentages, and the image gap is in nats. Bold pink values mark the best results in the reader comparison; dashes denote unavailable measurements.

Reader	New K/V	New W_O	Probe	Image gap	Coverage	Precision
Which part of the connection matters?						
No reader (control)			67.54	0.031	26.20	20.71
Learned gains only			67.54	0.032	27.06	20.03
Separate K/V only	✓		67.59	0.029	26.09	21.44
Separate W_O only		✓	67.21	0.317	33.71	37.69
Reusing attention with a separate visual output projection						
Reader without adaptive gating		✓	67.22	0.317	34.09	**37.95**
Adaptive visual reader (ours)		✓	67.13	**0.351**	**36.61**	37.03
Matched external encoder	–	–	–	–	37.92	45.09

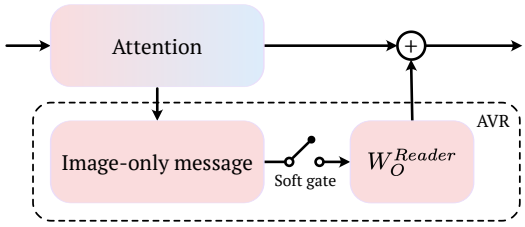


Figure 6: **Adaptive visual reader.** Reused attention weights and values form an image-only message. Text-conditioned gates scale it before a separate output projection. The projected message is added to the standard attention update; the residual connection is omitted.

Adaptive visual reader. These results motivate the *adaptive visual reader* (AVR) in Figure 6. AVR reuses the backbone's attention weights and values to form an image-only message m_i, normalized over image tokens. This removes the attenuation caused by competition with text tokens. A gate g_i, computed from the text state entering attention, controls each head's visual contribution before a separate output projection writes it into the residual stream:

$$h_i^{\text{postattn}} = h_i^{\text{in}} + A_i + W_O^{\text{Reader}}(g_i \odot m_i). \tag{1}$$

Here A_i is the standard attention output, and $\odot$ applies each head's gate to its message. The reader adds this update only at text positions and shares the backbone's attention computation and key-value cache. Appendix J gives the per-head equations and implementation details.

What does adaptive gating add? Compared with the same reader without adaptive gating, AVR raises the image gap from 0.317 to 0.351 and coverage from 34.09% to 36.61%, while precision falls slightly from 37.95% to 37.03% (Table 3). The separate output projection accounts for most of the precision gain over the native control; adaptive gating adds coverage and image dependence at a small cost in precision. Appendices G and J report fact counts and computational cost, and Appendices L–M analyze the non-adaptive reader.

Takeaway 3. Encoding alone is not enough.
Language makes little use of the retained visual features. A separate visual output projection improves caption precision, while adaptive gating adds coverage and image dependence.

6 Discussion and Conclusion

Native VLMs remove the external vision encoder and connector, but the backbone must still perform their functions. Visual features develop across layers, while text reads them mainly in the middle layers. Distillation builds and preserves these features, yet the same features yield stronger image dependence when they are read through an external connector. Reader ablations trace part of this gap to the mapping that writes visual messages into text states. Guided by this analysis, the adaptive visual reader (AVR) improves both caption precision and coverage. Building native VLMs therefore requires learning both the visual representations and the connection that lets language use them.

ETHICS STATEMENT

This work adheres to the ICLR Code of Ethics. Our experiments study existing vision-language models and datasets and do not involve human-subject recruitment or animal experimentation. We use established research datasets, including ImageNet-1K, CC12M, DCVLM-Baseline, MetaCLIP, FineVision, LLaVA-OneVision-1.5, ADE20K, COCO, MMVP, POPE and CAPEval, to study visual representations and image grounding. We are committed to respecting the usage terms of these resources and to maintaining transparency and integrity in reporting our methods and results.

REPRODUCIBILITY STATEMENT

To support reproducibility, we describe the model configurations, training schedules, hardware and evaluation protocols in the paper and appendices. Appendices A and B detail the layer-wise probes and attention interventions. Appendices D–G describe visual distillation, caption training and evaluation, and Appendix J provides the equations and implementation details of the adaptive visual reader. We use established datasets, including ImageNet-1K, ADE20K, COCO and MMVP, and specify the checkpoints and data processing procedures to help others reproduce the experiments.

AI USE STATEMENT

We used large language models to assist with code implementation and with writing and polishing the manuscript. Writing assistance included sentence rephrasing, grammar correction and improvements to clarity and readability, as well as figure formatting and LaTeX preparation. We also used `gpt-5.4-2026-03-05` as the automated judge for CAPEval, following the evaluation protocol described in Appendix G.

The authors take full responsibility for the scientific content of the paper, including the implementation, analyses and reported results, and for all text and artifacts prepared with LLM assistance.

REFERENCES

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. In *International Conference on Learning Representations (ICLR)*, 2017. URL https://arxiv.org/abs/1610.01644.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob L Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karén Simonyan. Flamingo: A Visual Language Model for Few-Shot Learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 23716–23736. Neural Information Processing Systems Foundation, Inc. (NeurIPS), 2022. doi: 10.52202/068431-1723. URL https://doi.org/10.52202/068431-1723.

Siyu An, Junru Lu, Junnan Dong, Qiufeng Wang, Yinghui Li, Weizhi Fei, Zichao Yu, Zheng Yuan, Biao Liu, Haopeng Wang, Renzhao Liang, Yixuan Yang, Yunhang Shen, Bo Ke, Keyu Chen, Linhao Luo, Difan Zou, Xiao Huang, Di Yin, Ruizhi Qiao, and Xing Sun. Toward Native Multimodal Modeling: A Roadmap. arXiv preprint arXiv:2605.25343, 2026. URL https://arxiv.org/abs/2605.25343.

Xiang An, Yin Xie, Kaicheng Yang, Wenkang Zhang, Xiuwei Zhao, Zheng Cheng, Yirui Wang, Songcen Xu, Changrui Chen, Didi Zhu, Chunsheng Wu, Huajie Tan, Chunyuan Li, Jing Yang, Jie Yu, Xiyao Wang, Bin Qin, Yumeng Wang, Zizhen Yan, Ziyong Feng, Ziwei Liu, Bo Li, and Jiankang Deng. LLaVA-OneVision-1.5: Fully Open Framework for Democratized Multimodal Training. arXiv preprint arXiv:2509.23661, 2025. URL https://arxiv.org/abs/2509.23661.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng,

Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-VL Technical Report. arXiv preprint arXiv:2502.13923, 2025. URL https://arxiv.org/abs/2502.13923.

Samyadeep Basu, Martin Grayson, Cecily Morrison, Besmira Nushi, Soheil Feizi, and Daniela Massiceti. Understanding Information Storage and Transfer in Multi-modal Large Language Models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, pp. 7400–7426. Curran Associates, Inc., 2024. doi: 10.52202/079017-0237. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/0dfe31d6e703e138d46a7d2fced38b7c-Paper-Conference.pdf.

Rohan Bavishi, Erich Elsen, Curtis Hawthorne, Maxwell Nye, Augustus Odena, Arushi Somani, and Sağnak Taşırlar. Introducing our Multimodal Models, 2023. URL https://www.adept.ai/blog/fuyu-8b.

Marie-France Beauvois. Optic aphasia: a process of interaction between vision and language. *Philosophical Transactions of the Royal Society of London. B, Biological Sciences*, 298(1089):35–47, 1982. doi: 10.1098/rstb.1982.0070. URL https://doi.org/10.1098/rstb.1982.0070.

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. Piqa: Reasoning about physical commonsense in natural language. arXiv preprint arXiv:1911.11641, 2019. URL https://arxiv.org/abs/1911.11641.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging Properties in Self-Supervised Vision Transformers. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9630–9640. IEEE, 2021. doi: 10.1109/iccv48922.2021.00951. URL https://doi.org/10.1109/iccv48922.2021.00951.

Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12M: Pushing Web-Scale Image-Text Pre-Training To Recognize Long-Tail Visual Concepts. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3557–3567. IEEE, 2021. doi: 10.1109/cvpr46437.2021.00356. URL https://doi.org/10.1109/cvpr46437.2021.00356.

Yangyi Chen, Xingyao Wang, Hao Peng, and Heng Ji. SOLO: A Single Transformer for Scalable Vision-Language Modeling. *Transactions on Machine Learning Research (TMLR)*, 2024. URL https://arxiv.org/abs/2407.06438.

Moulik Choraria, Xinbo Wu, Akhil Bhimaraju, Nitesh Sekhar, Yue Wu, Xu Zhang, Prateek Singhal, and Lav R. Varshney. DeepInsert: Early Layer Bypass for Efficient and Performant Multimodal Understanding. In *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pp. 7059–7079. Association for Computational Linguistics, 2026. doi: 10.18653/V1/2026.EACL-LONG.332. URL https://doi.org/10.18653/v1/2026.eacl-long.332.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge. arXiv preprint arXiv:1803.05457, 2018. URL https://arxiv.org/abs/1803.05457.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 248–255. IEEE, 2009. doi: 10.1109/cvpr.2009.5206848. URL https://doi.org/10.1109/cvpr.2009.5206848.

Haiwen Diao, Yufeng Cui, Xiaotong Li, Yueze Wang, Huchuan Lu, and Xinlong Wang. Unveiling Encoder-Free Vision-Language Models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, pp. 52545–52567. Curran Associates, Inc., 2024. doi: 10.52202/079017-1665. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/5e2217482fa75556f1970be809acd3f8-Paper-Conference.pdf.

Haiwen Diao, Xiaotong Li, Yufeng Cui, Yueze Wang, Haoge Deng, Ting Pan, Wenxuan Wang, Huchuan Lu, and Xinlong Wang. EVEv2: Improved Baselines for Encoder-Free Vision-Language Models. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 21014–21025. IEEE, 2025. doi: 10.1109/ICCV51701.2025.01953. URL https://doi.org/10.1109/iccv51701.2025.01953.

Haiwen Diao, Mingxuan Li, Silei Wu, Linjun Dai, Xiaohua Wang, Hanming Deng, Lewei Lu, Dahua Lin, and Ziwei Liu. From Pixels to Words – Towards Native Vision-Language Primitives at Scale. In *International Conference on Learning Representations (ICLR)*, pp. 109909–109929, 2026a. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/b2876deb92cbd098219a10da25671577-Paper-Conference.pdf.

Haiwen Diao, Jiahao Wang, Chenjing Ding, Hanming Deng, Jiangnan Chen, Ruixi Zhang, Ruohui Wang, Wenwen Tong, Xiangyu Fan, Yubo Wang, Yue Zhu, Yuwei Niu, Zhengqi Bai, Zhiqian Lin, Zhitao Yang, Zhongang Cai, Bo Yang, Chen Feng, Chengguang Lv, Guangjia Liu, Guanlin Wang, Hanyu Zhang, Haojia Yu, Hongcan Xiao, Hongli Wang, Huan Wu, Huaping Zhong, Jian Fang, Jianan Fan, Jiaqi Li, Jiefan Lu, Jing Zuo, Jingcheng Ni, Junxiang Xu, Linjun Dai, Mutian Xu, Peishen Yan, Penghao Wu, Ruijie Mao, Ruisi Wang, Shihao Bai, Shuang Yang, Shuya Yang, Shuyan Zheng, Silei Wu, Siying Li, Tao Chu, Tianbo Zhong, Tongxi Zhou, Weichao Luo, Weichen Fan, Wenhao Jia, Wenjie Gao, Xiangli Kong, Yan Li, Yang Yong, Zimo Wen, Zixuan Qian, Wenxiu Sun, Ruihao Gong, Quan Wang, Lewei Lu, Lei Yang, Ziwei Liu, and Dahua Lin. SenseNova-U1.5: Towards Native Unified Visual Intelligence. arXiv preprint arXiv:2609.11929, 2026b. URL https://arxiv.org/abs/2609.11929.

Haiwen Diao, Jiahao Wang, Penghao Wu, Yuhao Dong, Yuwei Niu, Yue Zhu, Zhongang Cai, Weichen Fan, Linjun Dai, Silei Wu, Xuanyu Zheng, Mingxuan Li, Yuanhan Zhang, Bo Li, Hanming Deng, Huchuan Lu, Quan Wang, Lei Yang, Lewei Lu, Dahua Lin, and Ziwei Liu. From Pixels to Words – Towards Native One-Vision Models at Scale. arXiv preprint arXiv:2605.28820, 2026c. URL https://arxiv.org/abs/2605.28820.

Haiwen Diao, Penghao Wu, Hanming Deng, Jiahao Wang, Shihao Bai, Silei Wu, Weichen Fan, Wenjie Ye, Wenwen Tong, Xiangyu Fan, Yan Li, Yubo Wang, Zhijie Cao, Zhiqian Lin, Zhitao Yang, Zhongang Cai, Yuwei Niu, Yue Zhu, Bo Liu, Chengguang Lv, Haojia Yu, Haozhe Xie, Hongli Wang, Jianan Fan, Jiaqi Li, Jiefan Lu, Jingcheng Ni, Junxiang Xu, Kaihuan Liang, Lianqiang Shi, Linjun Dai, Linyan Wang, Oscar Qian, Peng Gao, Pengfei Liu, Qingping Sun, Rui Shen, Ruisi Wang, Shengnan Ma, Shuang Yang, Siyi Xie, Siying Li, Tianbo Zhong, Xiangli Kong, Xuanke Shi, Yang Gao, Yongqiang Yao, Yves Wang, Zhengqi Bai, Zhengyu Lin, Zixin Yin, Wenxiu Sun, Ruihao Gong, Quan Wang, Lewei Lu, Lei Yang, Ziwei Liu, and Dahua Lin. SenseNova-U1: Unifying Multimodal Understanding and Generation with NEO-unify Architecture. arXiv preprint arXiv:2605.12500, 2026d. URL https://arxiv.org/abs/2605.12500.

Yan Fang, Mengcheng Lan, Zilong Huang, Weixian Lei, Yunqing Zhao, Yujie Zhong, Yingchen Yu, Qi She, Yao Zhao, and Yunchao Wei. Let ViT Speak: Generative Language-Image Pre-training. arXiv preprint arXiv:2605.00809, 2026. URL https://arxiv.org/abs/2605.00809.

Matteo Farina, Vishaal Udandarao, Thao Nguyen, Selim Kuzucu, Maximilian Böther, Andreas Hochlehnert, Adhiraj Ghosh, Marianna Nezhurina, Karsten Roth, Joschka Struber, Yuhui Zhang, Sebastian Dziadzio, Elaine Sui, Soumya Jahagirdar, Dhruba Ghosh, Hasan Hammoud, Thomas De Min, Simone Caldarella, Jehanzeb Mirza, Sedrick Keh, Mehdi Cherti, Hilde Kuehne, Bernt Schiele, Serena Yeung-Levy, Muhammad Ferjad Naeem, Federico Tombari, Ana Klimovic, Elisa Ricci, Matthias Bethge, Sewoong Oh, Ameya Prabhu, Alessio Tonioni, Jenia Jitsev, Massimiliano Mancini, Ludwig Schmidt, and Nikhil Parthasarathy. DataComp-VLM: Improved Open Datasets for Vision-Language Models. arXiv preprint arXiv:2606.28551, 2026. URL https://arxiv.org/abs/2606.28551.

Stephanie Fu, Tyler Bonnen, Devin Guillory, and Trevor Darrell. Hidden in plain sight: VLMs overlook their visual representations. arXiv preprint arXiv:2506.08008, 2025. URL https://arxiv.org/abs/2506.08008.

Gemma Team, Sherif El Abd, Vaibhav Aggarwal, Robin Algayres, Alek Andreev, Olivier Bachem, Ian Ballantyne, Cormac Brick, Victor Cărbune, Michelle Casbon, Mayank Chaturvedi, Aditya Chawla, Victor Cotruta, Alice Coucke, Phil Culliton, Robert Dadashi, Lucas Dixon, Mohamed Elhawaty, Utku Evci, Clément Farabet, Johan Ferret, Filippo Galgani, Sertan Girgin, Jean-Bastien Grill, Maarten Grootendorst, Jiaxian Guo, Cassidy Hardin, Yanzhang He, Steven M. Hernandez, Omri Homburger, Léonard Hussenot, Juyeong Ji, Armand Joulin, Aishwarya Kamath, Parnian Kassraie, Olivier Lacombe, Preethi Lahoti, Gaël Liu, Gus Martins, Luciano Martins, Tatiana Matejovicova, Ramona Merhej, Nikola Momchev, Sneha Mondal, Ryan Mullins, Sindhu Raghuram Panyam,

Shreya Pathak, Sarah Perrin, André Susano Pinto, Etienne Pot, Angéline Pouget, Alexandre Ramé, Sabela Ramos, Douglas Reid, David Rim, Morgane Rivière, Karsten Roth, Louis Rouillard, Omar Sanseviero, Pier Giuseppe Sessa, Shane Settle, Danila Sinopalnikov, Sara Smoot, Piotr Stanczyk, Andreas Steiner, Lawrence Stewart, Ilya Tolstikhin, Michael Tschannen, Anton Tsitsulin, Nino Vieillard, Renjie Wu, Pingmei Xu, Haichuan Yang, Edouard Yvinec, Biao Zhang, Li Zhang, Joe Zou, Nicolas Aagnes, Abdelrahman Abdelhamed, Jakub Adamek, Shivani Agrawal, Shubham Agrawal, Ibrahim Alabdulmohsin, Jean Baptiste Alayrac, Uri Alon, Chandramouli Amarnath, Ankesh Anand, Chrysovalantis Anastasiou, Setareh Ariafar, François-Xavier Aubet, Kyriakos Axiotis, Federico Barbero, Joelle Barral, Alexei Bendebury, Urs Bergmann, Stanley Bileschi, Kat Black, Mathieu Blondel, Sebastian Borgeaud, Arthur Bražinskas, Ryan Burnell, Robert Busa-Fekete, Mu Cai, Daniele Calandriello, Glenn Cameron, Charlotte Caucheteux, Rahma Chaabouni, Garima Chadha, Jetha Chan, Blake Jianhang Chen, Jesse Chen, Lin Chen, Xu Chen, Derek Cheng, Tzu-hsiang Chien, Nikolai Chinaev, Yi Chou, Zhaohui Chu, Benjamin Coleman, Pooja Consul, Sam Conway-Rahman, Scott Crowell, Dylan Cutler, Vivek Dani, Samira Daruki, Anil Das, Daniel Deutsch, Nishanth Dikkala, Li Ding, Qiuhan Ding, Shenil Dodhia, Konstantin Donhauser, Tulsee Doshi, Anca Dragan, Alex Druinsky, Sahil Dua, Zoltan Egyed, Danielle Eisenbud, Daniel Eppens, Cindy Fan, Bahare Fatemi, Yassir Fathullah, Vlad Feinberg, Milen Ferev, Sebastian Flennerhag, Takumi Fujimoto, João Gabriel Oliveira, Isaac Galatzer-Levy, João Gante, Simon Geisler, Soham Ghosal, Antonious M. Girgis, Tamara von Glehn, Alec Go, Alhaad Gokhale, Alex Grills, Yiming Gu, Mayank Gupta, Pramod Gupta, Guru Guruganesh, Raia Hadsell, Hamza Harkous, Jitendra Harlalka, Demis Hassabis, Anja Hauth, Joe Heyward, Arian Hosseini, Chih-Yang Hsia, I-Hung Hsu, Xiaopeng Huang, Yangsibo Huang, Kevin Hui, Adrian Hutter, Te I, Fotis Iliopoulos, Advait Jain, Ganesh Jawahar, Ziwei Ji, Qilin Jin, Melvin Johnson, Kandarp Joshi, Arun Kandoor, Wang-Cheng Kang, Koray Kavukcuoglu, Mehran Kazemi, Kathleen Kenealy, Amr Khalifa, Phoebe Kirk, Ivan Korotkov, Suraj Kothawade, Vitaly Kovalev, Neel Kovelamudi, Adam Kraft, Ravin Kumar, Vivek Kumar, Harish Kuppam, Justin Lannin, Chen-Yu Lee, Seungji Lee, Dmitry Lepikhin, Alon Levkovitch, Dongdong Li, Qiujia Li, Valentin Liévin, Ethan Lin, Ziqian Lin, Casper Liu, Tianlin Liu, Tianqi Liu, Xin Liu, Ivan Lobov, Mayank Lunayach, Min Ma, Gagan Madan, Andrii Maksai, Eric Malmi, Michal Matuszak, Daniel McDuff, Gaurav Menghani, Maciej Mikuła, Daniil Mirylenka, Karolis Misiunas, Vedant Misra, Andreea Mitran, Kareem Mohamed, Maksim Mukha, Eric Noland, James O'Donnell, Brendan O'Donoghue, Kate Olszewska, Bernett Orlando, Wanqiong Pan, Rina Panigrahy, Unnati Parekh, Nicolas Perez-Nieves, Chunjong Park, Eric Paskie, Liqian Peng, Bryce Petrini, Slav Petrov, Jonas Pfeiffer, Bilal Piot, Martyna Plomecka, Siim Poder, Octavio Ponce, Arijit Pramanik, David Racz, Anish Rajan, Michelle Ramanovich, Anand Rao, Marvin Ritter, Vitor Rodrigues, Evan Rosen, Mikołaj Rybiński, Noveen Sachdeva, Michaël E. Sander, Rohit Sathyanarayana, Sagar Savla, Samuel Schmidgall, Tal Schuster, George Scrivener, Benoit Seguin, Andrew Sellergren, Aliaksei Severyn, Izhak Shafran, Dhruv Shah, Bobak Shahriari, Yuan Shangguan, Ashish Shenoy, Pradeep Shenoy, Rakesh Shivanna, Pauline Sho, Lucas Spangher, Wojciech Stokowiec, Tim Strother, Yao Su, Yinghao Sun, Mukund Sundararajan, Andrea Tacchetti, Mor Hazan Taege, Pouya Tafti, Jean Tarbouriech, Chetan Tekur, Shantanu Thakoor, Rahul Thapa, Madeleine Traverse, Lenart Treven, Tao Tu, Chien Te Tung, Çağlar Ünlü, Petar Veličković, Malini Pooni Venkat, Sagar Gubbi Venkatesh, Vidya Venkiteswaran, Francesco Visin, Alex Vitvitskyi, Kiran Vodrahalli, Weiyi Wang, Xin Wang, Tris Warkentin, Jan Wassenberg, John Wieting, Cindy Wu, Lechao Xiao, Hao Xu, Yuhui Xu, Fuzhao Xue, Arun Yadav, Jun Yan, Antoine Yang, Lin Yang, Ming-Hsuan Yang, Ziyu Ying, Jae Hyeon Yoo, Morteza Zadimoghaddam, Sajjad Zafar, Fred Zhang, Jiageng Zhang, Jianyi Zhang, Xiaofan Zhang, Chao Zhao, David Zhou, and Chen Zou. Gemma 4 Technical Report. arXiv preprint arXiv:2607.02770, 2026. URL https://arxiv.org/abs/2607.02770.

Mor Geva, Jasmijn Bastings, Katja Filippova, and Amir Globerson. Dissecting Recall of Factual Associations in Auto-Regressive Language Models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 12216–12235. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.emnlp-main.751. URL https://doi.org/10.18653/v1/2023.emnlp-main.751.

Junlin Han, Shengbang Tong, David Fan, Minghao Chen, Philip Torr, Filippos Kokkinos, and Mike Lewis. Towards Physics of Multimodal Pretraining: Knowledge Flow, Modality Synergy, Early Unification, and Recipes. arXiv preprint arXiv:2608.05000, 2026. URL https://arxiv.org/abs/2608.05000.

John Hewitt and Christopher D. Manning. A Structural Probe for Finding Syntax in Word Representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4129–4138. Association for Computational Linguistics, 2019. doi: 10.18653/V1/N19-1419. URL https://doi.org/10.18653/V1/N19-1419.

Omri Kaduri, Shai Bagon, and Tali Dekel. What's in the Image? A Deep-Dive into the Vision of Vision Language Models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14549–14558. IEEE, 2025. doi: 10.1109/cvpr52734.2025.01356. URL https://doi.org/10.1109/cvpr52734.2025.01356.

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 *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 20758–20769. IEEE, 2025. doi: 10.1109/ICCV51701.2025.01930. URL https://doi.org/10.1109/iccv51701.2025.01930.

Handong Li, Yiyuan Zhang, Longteng Guo, Xiangyu Yue, and Jing Liu. Breaking the Encoder Barrier for Seamless Video-Language Understanding. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 23167–23176. IEEE, 2025. doi: 10.1109/ICCV51701.2025.02151. URL https://doi.org/10.1109/iccv51701.2025.02151.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. In *International Conference on Machine Learning (ICML)*, 2023a. URL https://arxiv.org/abs/2301.12597.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating Object Hallucination in Large Vision-Language Models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 292–305. Association for Computational Linguistics, 2023b. doi: 10.18653/v1/2023.emnlp-main.20. URL https://doi.org/10.18653/v1/2023.emnlp-main.20.

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 (ECCV)*, 2014. URL https://doi.org/10.1007/978-3-319-10602-1_48.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. arXiv preprint arXiv:2310.03744, 2023a. URL https://arxiv.org/abs/2310.03744.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual Instruction Tuning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pp. 34892–34916. Curran Associates, Inc., 2023b. doi: 10.52202/075280-1516. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/6dcf277ea32ce3288914faf369fe6de0-Paper-Conference.pdf.

Zhiheng Liu, Weiming Ren, Xiaoke Huang, Shoufa Chen, Tianhong Li, Mengzhao Chen, Yatai Ji, Sen He, Jonas Schult, Belinda Zeng, Tao Xiang, Wenhu Chen, Ping Luo, Luke Zettlemoyer, and Yuren Cong. Tuna-2: Pixel Embeddings Beat Vision Encoders for Multimodal Understanding and Generation. arXiv preprint arXiv:2604.24763, 2026a. URL https://arxiv.org/abs/2604.24763.

Zhining Liu, Ziyi Chen, Hui Liu, Chen Luo, Xianfeng Tang, Suhang Wang, Joy Zeng, Zhenwei Dai, Zhan Shi, Tianxin Wei, Benoit Dumoulin, and Hanghang Tong. Seeing but Not Believing: Probing the Disconnect Between Visual Attention and Answer Correctness in VLMs. arXiv preprint arXiv:2510.17771, 2025. URL https://arxiv.org/abs/2510.17771.

Zhipeng Liu, Haochen Wang, and Zhaoxiang Zhang. CAPEval: A Decoupled Caption Evaluation across Understanding and Generation. arXiv preprint arXiv:2608.02589, 2026b. URL https://arxiv.org/abs/2608.02589.

Gen Luo, Wenhan Dou, Wenhao Li, Zhaokai Wang, Xue Yang, Changyao Tian, Hao Li, Weiyun Wang, Wenhai Wang, Xizhou Zhu, Yu Qiao, and Jifeng Dai. Mono-InternVL-1.5: Towards Cheaper and Faster Monolithic Multimodal Large Language Models. arXiv preprint arXiv:2507.12566, 2025a. URL https://arxiv.org/abs/2507.12566.

Gen Luo, Xue Yang, Wenhan Dou, Zhaokai Wang, Jiawen Liu, Jifeng Dai, Yu Qiao, and Xizhou Zhu. Mono-InternVL: Pushing the Boundaries of Monolithic Multimodal Large Language Models with Endogenous Visual Pre-training. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24960–24971. IEEE, 2025b. doi: 10.1109/CVPR52734.2025.02324. URL https://doi.org/10.1109/cvpr52734.2025.02324.

Lingchen Meng, Jianwei Yang, Rui Tian, Xiyang Dai, Zuxuan Wu, Jianfeng Gao, and Yu-Gang Jiang. DeepStack: Deeply Stacking Visual Tokens is Surprisingly Simple and Effective for LMMs. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, pp. 23464–23487. Curran Associates, Inc., 2024. doi: 10.52202/079017-0739. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/29cd7f8331d13ede6dc6d6ef3dfacb70-Paper-Conference.pdf.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. arXiv preprint arXiv:1609.07843, 2016. URL https://arxiv.org/abs/1609.07843.

Roy Miles and Krystian Mikolajczyk. Understanding the Role of the Projector in Knowledge Distillation. In *Conference on Artificial Intelligence (AAAI)*, volume 38, pp. 4233–4241. Association for the Advancement of Artificial Intelligence (AAAI), 2024. doi: 10.1609/AAAI.V38I5.28219. URL https://doi.org/10.1609/aaai.v38i5.28219.

Clement Neo, Luke Ong, Philip Torr, Mor Geva, David Krueger, and Fazl Barez. Towards Interpreting Visual Information Processing in Vision-Language Models. In *International Conference on Learning Representations (ICLR)*, pp. 57172–57189, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/900fb3439e4968df79a6f2bfedec49cd-Paper-Conference.pdf.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning Robust Visual Features without Supervision. *Transactions on Machine Learning Research (TMLR)*, 2024. URL https://arxiv.org/abs/2304.07193.

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. In *International Conference on Machine Learning (ICML)*, 2021. URL https://arxiv.org/abs/2103.00020.

Mike Ranzinger, Jon Barker, Greg Heinrich, Pavlo Molchanov, Bryan Catanzaro, and Andrew Tao. PHI-S: Distribution Balancing for Label-Free Multi-Teacher Distillation. arXiv preprint arXiv:2410.01680, 2024a. URL https://arxiv.org/abs/2410.01680.

Mike Ranzinger, Greg Heinrich, Jan Kautz, and Pavlo Molchanov. AM-RADIO: Agglomerative Vision Foundation Model Reduce All Domains Into One. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12490–12500. IEEE, 2024b. doi: 10.1109/CVPR52733.2024.01187. URL https://doi.org/10.1109/cvpr52733.2024.01187.

M. Jane Riddoch and Glyn W. Humphreys. Visual object processing in optic aphasia: A case of semantic access agnosia. *Cognitive Neuropsychology*, 4(2):131–185, 1987. doi: 10.1080/02643298708252038. URL https://doi.org/10.1080/02643298708252038.

Israfel Salazar, Stella Frank, Dan Oneata, Desmond Elliott, and Constanza Fierro. Pathways of Visual Information Flow in Vision-Language Models. arXiv preprint arXiv:2607.03358, 2026. URL https://arxiv.org/abs/2607.03358.

Mert Bülent Sariyildiz, Philippe Weinzaepfel, Thomas Lucas, Diane Larlus, and Yannis Kalantidis. UNIC: Universal Classification Models via Multi-teacher Distillation. In *European Conference on Computer Vision (ECCV)*, 2024. doi: 10.1007/978-3-031-73235-5_20. URL https://arxiv.org/abs/2408.05088.

Jinghuan Shang, Karl Schmeckpeper, Brandon B. May, Maria Vittoria Minniti, Tarik Kelestemur, David Watkins, and Laura Herlant. Theia: Distilling Diverse Vision Foundation Models for Robot Learning. In *Conference on Robot Learning*, 2024. URL https://arxiv.org/abs/2407.20179.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. DINOv3. arXiv preprint arXiv:2508.10104, 2025. URL https://arxiv.org/abs/2508.10104.

Changyao Tian, Hao Li, Gen Luo, Xizhou Zhu, Weijie Su, Hanming Deng, Jinguo Zhu, Jie Shao, Ziran Zhu, Yunpeng Liu, Lewei Lu, Wenhai Wang, Hongsheng Li, and Jifeng Dai. NaViL: Rethinking Scaling Properties of Native Multimodal Large Language Models under Data Constraints. In *Advances in Neural Information Processing Systems (NeurIPS)*. Curran Associates, Inc., 2025. doi: 10.52202/085713-2868. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/7b99e3c648898b9e4923dea0aeb4afa1-Paper-Conference.pdf.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes Wide Shut? Exploring the Visual Shortcomings of Multimodal LLMs. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9568–9578. IEEE, 2024. doi: 10.1109/cvpr52733.2024.00914. URL https://doi.org/10.1109/cvpr52733.2024.00914.

Michael Tschannen, Manoj Kumar, Andreas Steiner, Xiaohua Zhai, Neil Houlsby, and Lucas Beyer. Image Captioners Are Scalable Vision Learners Too. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pp. 46830–46855. Curran Associates, Inc., 2023. doi: 10.52202/075280-2029. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/92369a01fbe8046a093746389b2c413e-Paper-Conference.pdf.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier Hénaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. SigLIP 2: Multilingual Vision-Language Encoders with Improved Semantic Understanding, Localization, and Dense Features. arXiv preprint arXiv:2502.14786, 2025. URL https://arxiv.org/abs/2502.14786.

Han Wang, Yongjie Ye, Bingru Li, Yuxiang Nie, Jinghui Lu, Jingqun Tang, Yanjie Wang, and Can Huang. Vision as LoRA. arXiv preprint arXiv:2503.20680, 2025. URL https://arxiv.org/abs/2503.20680.

Luis Wiedmann, Orr Zohar, Amir Mahla, Xiaohan Wang, Rui Li, Thibaud Frere, Leandro von Werra, Aritra Roy Gosthipaty, and Andrés Marafioti. FineVision: Open Data Is All You Need. arXiv preprint arXiv:2510.17269, 2025. URL https://arxiv.org/abs/2510.17269.

Haoyuan Wu, Aoqi Wu, Hai Wang, Jiajia Wu, Jinxiang Ou, and Bei Yu. Scaling Native Multimodal Pre-Training From Scratch. arXiv preprint arXiv:2607.22043, 2026. URL https://arxiv.org/abs/2607.22043.

Hu Xu, Saining Xie, Xiaoqing Ellen Tan, Po-Yao Huang, Russell Howes, Vasu Sharma, Shang-Wen Li, Gargi Ghosh, Luke Zettlemoyer, and Christoph Feichtenhofer. Demystifying CLIP Data. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://arxiv.org/abs/2309.16671.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang,

Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 Technical Report. arXiv preprint arXiv:2505.09388, 2025a. URL https://arxiv.org/abs/2505.09388.

Rui Yang, Lin Song, Yicheng Xiao, Runhui Huang, Yixiao Ge, Ying Shan, and Hengshuang Zhao. HaploVL: A Single-Transformer Baseline for Multi-Modal Understanding. In *International Conference on Machine Learning (ICML)*, volume 267 of *Proceedings of Machine Learning Research*, pp. 71073–71091, 2025b. URL https://arxiv.org/abs/2503.14694.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. Hellaswag: Can a machine really finish your sentence? arXiv preprint arXiv:1905.07830, 2019. URL https://arxiv.org/abs/1905.07830.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid Loss for Language Image Pre-Training. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11941–11952. IEEE, 2023a. doi: 10.1109/iccv51070.2023.01100. URL https://doi.org/10.1109/iccv51070.2023.01100.

Yuexiang Zhai, Shengbang Tong, Xiao Li, Mu Cai, Qing Qu, Yong Jae Lee, and Yi Ma. Investigating the Catastrophic Forgetting in Multimodal Large Language Models. arXiv preprint arXiv:2309.10313, 2023b. URL https://arxiv.org/abs/2309.10313.

Yuhui Zhang, Yuchang Su, Yiming Liu, Xiaohan Wang, James Burgess, Elaine Sui, Chenyu Wang, Josiah Aklilu, Alejandro Lozano, Anjiang Wei, Ludwig Schmidt, and Serena Yeung-Levy. Automated generation of challenging multiple-choice questions for vision language model evaluation. arXiv preprint arXiv:2501.03225, 2025a. URL https://arxiv.org/abs/2501.03225.

Zhi Zhang, Srishti Yadav, Fengze Han, and Ekaterina Shutova. Cross-modal Information Flow in Multimodal Large Language Models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19781–19791. IEEE, 2025b. doi: 10.1109/cvpr52734.2025.01842. URL https://doi.org/10.1109/cvpr52734.2025.01842.

Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic Understanding of Scenes Through the ADE20K Dataset. *International Journal of Computer Vision (IJCV)*, 127(3):302–321, 2018. doi: 10.1007/s11263-018-1140-0. URL https://doi.org/10.1007/s11263-018-1140-0.

A LAYER-WISE PROBES

This appendix describes the attentive probes behind Figure 2, which evaluate visual representations at every backbone layer through ImageNet classification. Appendix B gives the attention cuts of Figure 3, and Appendix I describes the separate nearest-centroid diagnostic used in Section 5.

A.1 MODELS AND LAYER INDEXING

Table 4 lists five native checkpoints, four NEO comparisons with a predefined front block, and two external-encoder references used in the layer-wise comparison. The native and NEO checkpoints process raw image patches within their backbones. LLaVA-1.5 and Qwen2.5-VL first process the image with an external vision encoder and pass its features to the language decoder. We extract states from the language backbone in both groups. The attention-cut study uses four native checkpoints and all four NEO comparisons, excluding Fuyu.

Table 4: **Models and probe dimensions.** L counts the backbone blocks included in the layer-wise analysis, including NEO's pre-buffer. D is the width of the extracted states, and $H = D/64$ is the number of heads in the attentive probe, not the VLM's attention head count. Rows below the rule are external-encoder references.

Checkpoint	L	D	H
`Paranioar/NEO1_0-2B-SFT`	40	2048	32
`Paranioar/NEO1_0-9B-SFT`	42	4096	64
`Paranioar/NEO1_5-2B-SFT`	40	2048	32
`Paranioar/NEO1_5-9B-SFT`	42	4096	64
`ByteDance-Seed/SAIL-7B`	32	4096	64
`google/gemma-4-12B-it`	48	3840	60
`OpenGVLab/Mono-InternVL-2B`	24	2048	32
`BAAI/EVE-7B-HD-v2.0`	28	3584	56
`adept/fuyu-8b`	36	4096	64
`llava-hf/llava-1.5-7b-hf`	32	4096	64
`Qwen/Qwen2.5-VL-7B-Instruct`	28	3584	56

Layer ℓ denotes the output of backbone block ℓ, with $\ell = 1, \ldots, L$. We plot normalized depth ℓ/L, so the first point is the output of the first block rather than the input embedding. In NEO, the pre-buffer occupies the first 12 blocks of the 2B checkpoints and the first 6 blocks of the 9B checkpoints. These blocks are included in L, placing the pre-buffer boundaries at $12/40$ and $6/42$, respectively.

A.2 IMAGE PREPARATION AND FEATURE EXTRACTION

The probes use the ImageNet-1K training split of 1,281,167 images and the validation split of 50,000 images (Deng et al., 2009). Training images receive a bicubic random resized crop to 224×224, with crop scale in $[0.08, 1.0]$ and aspect ratio in $[3/4, 4/3]$, followed by a horizontal flip with probability 0.5. Validation images are resized to a shorter side of 256 pixels and center-cropped to 224×224. Each checkpoint's image processor is applied after these transformations, so the final token sequence follows that model's input pipeline.

Each input places the image in the checkpoint's multimodal template with its attention mask; no question prompt is added. At each block, we select the valid image-token rows, excluding text and padding positions from the probe. For an image with T valid image tokens, let $X_\ell \in \mathbb{R}^{T \times D}$ denote these post-block states. We apply the checkpoint's final decoder normalization to obtain

$$\widetilde{X}_\ell = \mathrm{Norm}_{\mathrm{final}}(X_\ell). \tag{2}$$

The same frozen normalization is used at every depth. The VLM parameters remain fixed throughout feature extraction and probe training.

A.3 ATTENTIVE POOLING AND PROBE TRAINING

Each layer has its own attentive pooling classifier. We first apply non-affine batch normalization to valid image-token rows, with numerical constant $\epsilon = 10^{-6}$. A learned query then pools the normalized states using bias-free key and value projections,

$$\bar{X}_\ell = \mathrm{BN}(\widetilde{X}_\ell), \qquad K_\ell = \bar{X}_\ell W_\ell^K, \qquad V_\ell = \bar{X}_\ell W_\ell^V, \tag{3}$$

where $W_\ell^K, W_\ell^V \in \mathbb{R}^{D\times D}$. The query $q_\ell \in \mathbb{R}^D$ is initialized from $\mathcal{N}(0, 0.02^2)$. We split the query, keys and values into $H = D/64$ heads. For head h, the pooled vector is

$$\alpha_{\ell,h} = \mathrm{softmax}\left(\frac{q_{\ell,h}K_{\ell,h}^\top}{\sqrt{64}}\right), \qquad z_{\ell,h} = \alpha_{\ell,h}V_{\ell,h}. \tag{4}$$

The softmax ranges over valid image tokens. We concatenate the head outputs into $z_\ell \in \mathbb{R}^D$ and apply a bias-free linear classifier $W_\ell^C \in \mathbb{R}^{1000\times D}$ to predict the ImageNet class. The query, key and value projections, and classifier are trained separately for each layer. Probe width follows the backbone width listed in Table 4.

We train for one epoch using AdamW, learning rate 10^{-3}, weight decay 0.01, and a global batch size of 4,096 across eight GPUs. The run contains 313 optimizer steps, with cosine learning-rate decay to zero and no warmup. For an image with label y, each layer contributes a cross-entropy loss,

$$\mathcal{L}_\ell = -\log\left[\mathrm{softmax}(W_\ell^C z_\ell)\right]_y, \qquad \mathcal{L}_{\mathrm{probe}} = \sum_{\ell=1}^{L}\mathcal{L}_\ell. \tag{5}$$

The probes have disjoint parameters, so summing the losses gives each probe its own layer's gradient without averaging it over the number of layers. We report top-1 accuracy on the full validation split after this training budget.

Depth summaries. Let a_ℓ be a model's validation accuracy at layer ℓ, and let $a_{\max} = \max_\ell a_\ell$. The dots in Figure 2(b) mark

$$\ell_{90} = \min\{\ell \mid a_\ell \geq 0.9a_{\max}\}. \tag{6}$$

We report this first crossing as ℓ_{90}/L, relative to each model's peak rather than an absolute 90% accuracy. It marks when the probe reaches near-peak accuracy.

B ATTENTION-CUT DEFINITIONS

Model coverage. We include Fuyu-8B in the ImageNet probe study but exclude it from the attention-cut analysis because its intact MMVP accuracy is too low to support an informative comparison. This floor effect limits the measurable degradation under attention removal and makes retained-accuracy ratios sensitive to small changes in correct-pair counts.

Let S_{ij}^ℓ be the attention logit from query i to key j, and B_{ij}^ℓ the model's original attention mask. We add an intervention mask $C_{ij}^{\ell,d}$:

$$a_{ij}^\ell = \mathrm{Softmax}_j(S_{ij}^\ell + B_{ij}^\ell + C_{ij}^{\ell,d}). \tag{7}$$

For text queries and image keys, early-only access (the suffix cut) sets $C_{ij}^{\ell,d} = -\infty$ when $\ell > d$, while late-only access (the prefix cut) sets it to $-\infty$ when $\ell < d$. All other entries of C are zero. Thus, the boundary layer is included in both access regimes, and the original causal or modality-specific mask is preserved.

MMVP strict-pair accuracy requires both questions in each image pair to be answered correctly. Retained accuracy divides each intervention result by the intact model's accuracy. For Figure 3, we compute centered finite-difference slopes at sampled boundaries and select the largest positive slope for early-only access and the largest-magnitude negative slope for late-only access. The shaded bands span these transition locations across architectures.

Interpreting the cuts. Each cut removes a cumulative range of text-to-image attention connections. We compare the first 90% probe crossing with the steepest accuracy changes under these cuts to describe the progression from feature formation to cross-modal interaction. The cuts apply to all text queries, covering both direct image-to-answer attention and transfer through prompt tokens. These paths can substitute for one another under attention removal (Salazar et al., 2026).

Chance level. Each MMVP question has two options, so random guessing gives a strict pair accuracy of 25%. With all text-to-image attention removed, every model scores 0%, consistent with giving the same answer to both questions of each pair. Mono-InternVL-2B answers only 28 of 150 pairs correctly when intact, below chance, so small changes in correct-pair counts affect its retained ratios. Excluding it changes the early-only and late-only retained-accuracy medians of Appendix B.1 to 13% and 91%, respectively.

B.1 Encoding and reading depths per model

Table 5 lists, for each checkpoint, the depth at which the probe first reaches 90% of its peak accuracy and the depths of the steepest accuracy changes under the two cuts, computed as described above. In every model, the steepest changes lie between $\ell/L = 0.46$ and 0.71, while the probe crossing ranges from 0.12 to 0.61. Keeping text-to-image attention only in the first 40% of layers retains a median of 15% of intact accuracy, while removing it from those same layers retains 96%. These medians use, for each model, the sampled boundary nearest to $d/L = 0.4$, which lies between 0.39 and 0.42.

Table 5: **Encoding and reading depths.** Encoding is the first normalized depth at which the probe reaches 90% of its peak accuracy. Suffix and prefix give the depth of the steepest change in retained MMVP accuracy under each cut. Intact pairs counts the MMVP pairs answered correctly without intervention, out of 150. Fuyu-8B has no attention cuts.

			Encoding	Reading		
Checkpoint	L	**Pre-buffer end**	probe 90%	suffix cut	prefix cut	**Intact pairs**
NEO1.0-2B	40	0.30	0.23	0.60	0.60	49
NEO1.0-9B	42	0.14	0.12	0.57	0.60	71
NEO1.5-2B	40	0.30	0.23	0.60	0.63	65
NEO1.5-9B	42	0.14	0.12	0.57	0.60	76
SAIL-7B	32	–	0.53	0.59	0.59	59
Gemma-4-12B	48	–	0.15	0.48	0.50	101
Mono-InternVL-2B	24	–	0.46	0.46	0.54	28
EVEv2-7B	28	–	0.61	0.71	0.64	73
Fuyu-8B	36	–	0.50	–	–	–

C Direct native training pilot

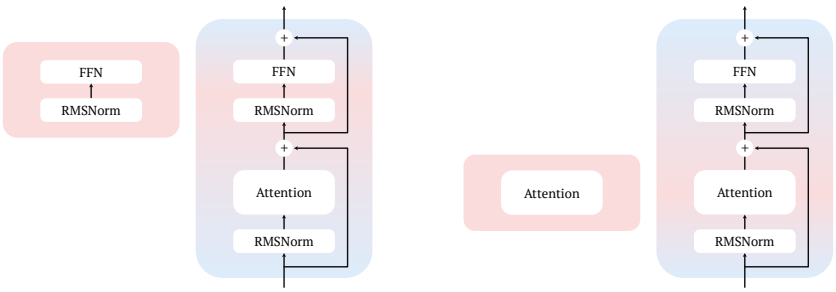
This pilot supports Section 4. The native model and external-encoder model using CLIP (Radford et al., 2021) share the Qwen3-1.7B backbone, training data and task-supervised schedule; only their image pathways differ. Both are trained on the same mixture of data from FineVision (Wiedmann et al., 2025) and LLaVA-OneVision-1.5 (An et al., 2025). Table 6 reports POPE (Li et al., 2023b), per-item MMVP scores (Tong et al., 2024), VMCBench (Zhang et al., 2025a) at 20k steps, and the score difference between intact and swapped images, which measures how much the answers depend on the image. The CLIP model at 1k steps is ahead of the native model at 20k steps on POPE, MMVP and image dependence. The pilot uses Qwen3-1.7B, and the distillation and reader experiments use Qwen3-0.6B.

D Controlled visual pretraining

Visual feature distillation uses a Qwen3-0.6B backbone and a SigLIP2 NaFlex teacher (Tschannen et al., 2025). The standard run uses four H200 GPUs, global batch size 1,024, 12,520 optimizer steps,

Table 6: **Direct native training in the Qwen3-1.7B pilot.** POPE, MMVP and VMCBench are scores on a 0–100 scale. MMVP is per-item, unlike the strict-pair metric in Section 3. Intact minus swapped retains the original score units. Dashes indicate unrecorded values.

Step	Pathway	POPE	MMVP	VMCBench	Intact − swapped
1k	External CLIP	80.2	57.7	–	+0.080
1k	Native	49.8	50.0	–	+0.018
3k	External CLIP	82.6	56.7	–	–
3k	Native	53.2	50.0	–	–
20k	External CLIP	84.5	61.7	59.4	+0.197
20k	Native	53.4	53.0	46.4	+0.059



(a) Visual FFN branch (b) Visual attention branch

Figure 7: **Two image-specific branches within a pretrained language block.** Pink regions mark image-specific computation, and blue regions mark the language path. Blended regions indicate shared computation. Panels (a) and (b) add a visual feed-forward network (FFN) and visual attention projections, respectively, as compared in Table 1(a). Plus signs denote residual additions.

and seed 42. This corresponds to approximately 12.8M image presentations. The image-specific branch copies the corresponding language weights or uses random initialization, while the text path retains its pretrained weights. The default objective is $0.9\mathcal{L}_{\cos} + 0.1\mathcal{L}_{\mathrm{SmoothL1}}$. Initialization comparisons use final-layer supervision. The visual attention loss comparison fixes intermediate supervision weight at 0.2. Appendix F varies that weight separately.

ImageNet classification uses top-1 accuracy. ADE20K semantic segmentation uses mean intersection over union (mIoU), and COCO image-text retrieval uses recall at 1 (R@1). The visual FFN shares the language attention projections. With visual attention, image keys and values come from the visual branch, while the attention output at text positions passes through the language output projection.

E Distillation objectives

Let $s = z_s\Pi \in \mathbb{R}^{D_t}$ be a projected student feature and $t \in \mathbb{R}^{D_t}$ the corresponding teacher feature. For each matched image patch, the objectives are

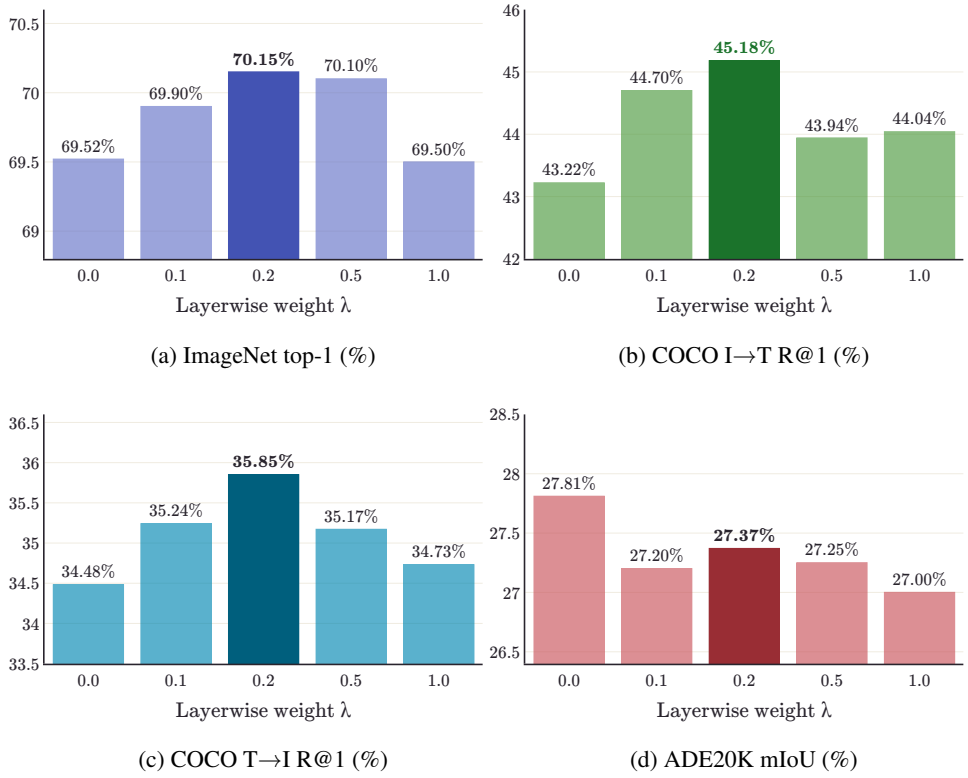
$$\mathcal{L}_{\cos}(s,t) = 1 - \frac{\langle s,t\rangle}{\|s\|_2\|t\|_2}, \tag{8}$$

$$\mathcal{L}_{\mathrm{MSE}}(s,t) = \frac{1}{D_t}\sum_k (s_k - t_k)^2, \tag{9}$$

$$\mathcal{L}_{\mathrm{SmoothL1}}(s,t) = \frac{1}{D_t}\sum_k \rho_\beta(s_k - t_k), \tag{10}$$

where $\beta = 0.001$ and

$$\rho_\beta(u) = \begin{cases} u^2/(2\beta), & |u| < \beta, \\ |u| - \beta/2, & \text{otherwise.} \end{cases} \tag{11}$$



(a) ImageNet top-1 (%)

(b) COCO I→T R@1 (%)

(c) COCO T→I R@1 (%)

(d) ADE20K mIoU (%)

Figure 8: **Intermediate-layer supervision weight sweep.** A moderate weight of $\lambda = 0.2$ gives the best ImageNet and COCO retrieval values in this sweep, while final-only supervision remains best on ADE20K. The differences between $\lambda = 0.1$, 0.2, and 0.5 are small.

Losses are averaged over valid image patches. For online-standardized MSE, channel statistics are tracked during training and feature errors are measured in standardized coordinates. The loss comparison holds the architecture and intermediate-layer supervision fixed. Cosine alignment matches student and teacher feature directions.

F Intermediate-layer supervision

We hold the visual attention architecture and combined cosine and SmoothL1 objective fixed, and vary the weight of the mean intermediate-layer loss. Figure 8 shows the trade-off across classification, segmentation, and retrieval. The final-layer loss has unit weight throughout.

G Caption training and grounding metrics

Caption training. Caption training for the feature-retention and reading experiments uses a pool of 7,046,382 image-caption pairs drawn from MetaCLIP (Xu et al., 2024). This pool is separate from the FineVision and LLaVA-OneVision-1.5 mixture used in the direct native training pilot (Appendix C). The full baseline schedule runs for 22,667 optimizer steps, corresponding to approximately 1.3M image presentations, using AdamW, a cosine schedule, learning rate 2×10^{-5}, and zero weight decay. Caption cross-entropy is applied to text targets, with image positions masked from the loss. Continued-distillation runs retain image-feature supervision during caption training.

Probes on trained models. Probes on caption-trained models, including the reader variants, feed each image followed by the fixed text `A photo of`, with no class name, answer or caption, and without generation. We extract the image-token states at each layer and train a separate attentive classifier with the protocol of Appendix A.3; the text-position states are not classified. Probes on the

distilled models before caption training use image patches and their coordinates alone, without any text.

External references. The SigLIP2 reference feeds a trainable SigLIP2 encoder through a connector into the first layer of the same language model. The matched external references instead run the distilled visual encoder of the corresponding native model as a separate encoder. Its output enters the first layer of the language model through a connector, and the language model then processes the image tokens with its own weights. For the visual FFN, Table 2 reports one run with this encoder frozen and one with it trained. The visual attention reference in Tables 11 and 3 is initialized from the distilled model of the readers. All external models are trained on the same captions and schedule as the native models. The frozen-encoder probe in Tables 2 and 7 reuses the initial encoder-only result, 66.166% rounded to 66.17%, because this encoder is frozen. It measures the encoder output before the connector, not image-token states at the final language-model layer after the external connection.

Evaluation settings. The image gap uses a fixed set of 12 image-caption pairs from the training data source, which may overlap with the training data. Tables 2 and 7 use one evaluation of this set. The layer-cut evaluation of Appendix M weights NLL by the number of supervised tokens and uses a fixed cyclic replacement, so its values differ for the same checkpoint. For the visual attention control, for example, it gives an image gap of 0.036 against 0.045. All variants in Table 3, including the external reference, are evaluated with the same generation settings and judge.

CAPEval. We follow the CAPEval checklist protocol (Liu et al., 2026b). For each caption, the judge labels each reference checklist item as correctly described, incorrectly described, or not mentioned. If the counts are N_{yes}, N_{no}, and N_{nm}, then

$$\text{Coverage} = 100\frac{N_{\text{yes}} + N_{\text{no}}}{N_{\text{yes}} + N_{\text{no}} + N_{\text{nm}}}, \qquad \text{Precision} = 100\frac{N_{\text{yes}}}{N_{\text{yes}} + N_{\text{no}}}. \tag{12}$$

We use the official prompt, parser, and aggregation, but replace the original Qwen2.5-72B judge with `gpt-5.4-2026-03-05`. Coverage measures how many checklist facts the caption attempts to describe. Precision measures correctness among those mentioned facts.

Image replacement. The NLL diagnostic keeps the reference caption fixed and replaces its image with another image from the evaluation set. For correct image I, replacement I', and caption Y, we report

$$\Delta_{\text{shuffle}} = \text{NLL}(Y \mid I') - \text{NLL}(Y \mid I). \tag{13}$$

The baseline NLL comparison uses a fixed set of 12 caption examples. For the reader comparison in Table 3, we compute the same quantity per token,

$$\Delta_t = \log p(y_t \mid I, y_{<t}) - \log p(y_t \mid I', y_{<t}). \tag{14}$$

A positive value means that the correct image makes the reference token more likely. This intervention replaces the image, preserving the patch order within each image. Table 3 reports the token-weighted mean of Δ_t over 5,000 reference captions as the image gap.

Fact counts. The reader family is evaluated on 14,965 checklist items. The matched external reference describes 2,559 checklist facts correctly and 3,116 incorrectly, compared with 1,936 and 3,166 for the visual reader.

H Controls for the reading comparison

Table 7 collects the teacher-forced measurements behind Section 5.1. All native and same-path rows use the visual-FFN family, and all rows use the same 12 images. The caption-CE-only native model starts from a different distilled model, with two visual FFNs trained on ImageNet, CC12M and DCVLM-Baseline with intermediate supervision (73.44% ImageNet accuracy). It therefore differs from the main native model in more than the absence of distillation during caption training. In the frozen-language-model run, caption CE trains only the visual branch.

Table 7: **Controls for the reading comparison.** Caption NLL and image gap on the fixed 12-image set. Probe is ImageNet top-1 accuracy (%) at the last layer after caption training, except the frozen-encoder row, which reuses the encoder-only probe (Appendix G). Dashes mark values we did not measure.

Model	Probe	Caption NLL	Image gap
Native, continued distillation	64.83	1.895	+0.042
Native, caption CE only, stronger distilled model	–	1.891	+0.084
Native, frozen language model	–	2.665	+0.043
Same distilled encoder, external, frozen	66.17	1.848	+0.235
Same distilled encoder, external, trained	52.50	1.856	+0.178
SigLIP2, external, trained	86.08	1.571	+0.741

I NEAREST-CENTROID DIAGNOSTIC AND MESSAGE SCALING

The diagnostic in Figure 9 uses the native visual-FFN model of Table 2, trained with continued distillation and caption cross-entropy, and the SigLIP2 reference. Appendix L applies the same classifier to the visual attention models of Table 11. We select a deterministic, class-balanced subset of ImageNet validation containing 100 classes and 20 images per class. For each class, 15 images fit the centroid and five are held out for evaluation, giving 1,500 fitting images and 500 test images. This 100-way diagnostic is separate from the 1,000-way attentive probes above.

Attention decomposition. At each layer, we separate the attention update into image and non-image contributions (Basu et al., 2024). With attention weights α_{ij} from text query i to position j, value vectors v_j and output projection W_O,

$$A_i = \underbrace{W_O \sum_{j \in \text{image}} \alpha_{ij} v_j}_{M_{\text{image}}} + \underbrace{W_O \sum_{j \notin \text{image}} \alpha_{ij} v_j}_{M_{\text{nonimage}}}, \qquad h_i^{\text{postattn}} = h_i^{\text{in}} + M_{\text{image}} + M_{\text{nonimage}}. \quad (15)$$

We call M_{image} the visual message. It enters the text state through residual addition before the feed-forward network. We measure its norm relative to the incoming text state and separately probe class information in the message and the layer's output state.

Prompt and measurement position. Each image is followed by the fixed prefix `A photo of`. We measure the final non-image prefix token at decoder layer 28. Using the model's existing attention weights, we isolate the image-source contribution M_{image} after the attention output projection, as defined in Equation 15. We also record the post-block text state h_i^{out}, after residual addition and the feed-forward update. The two representations are classified separately.

Centroid fitting. For a representation vector r_n, we first compute $\hat{r}_n = r_n / \|r_n\|_2$. For class c with fitting set $\mathcal{F}_c$, the centroid and prediction rule are

$$\mu_c = \frac{1}{15} \sum_{n \in \mathcal{F}_c} \hat{r}_n, \qquad \hat{\mu}_c = \frac{\mu_c}{\|\mu_c\|_2}, \qquad \hat{y}_n = \arg\max_c \hat{r}_n^\top \hat{\mu}_c. \quad (16)$$

We report top-1 accuracy over the 500 held-out images. The recorded confidence intervals use 10,000 bootstrap resamples at the class level.

Gain intervention. We retain the original attention weights and values, and replace the attention output at the measured text position with

$$A_i^{(g)} = A_i + (g-1) M_{\text{image}}, \qquad g \in \{1, 4, 8, 16, 32, 64\}. \quad (17)$$

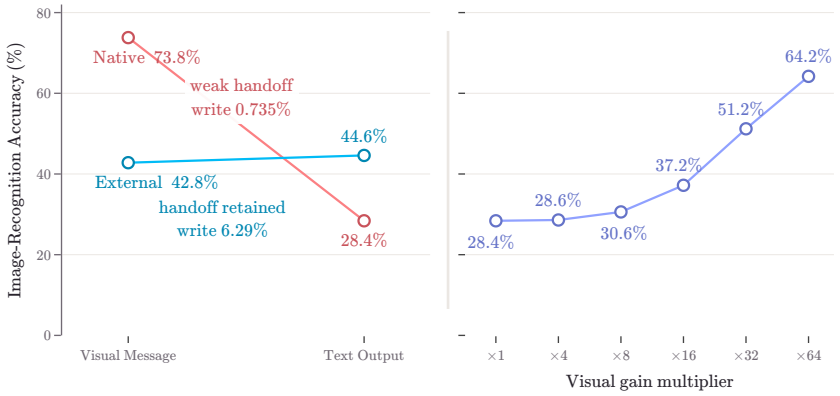
The modified output passes through the original residual addition, normalization and feed-forward update. Model weights, the incoming residual, and the non-image attention contribution remain fixed. We then apply the nearest-centroid diagnostic to the resulting post-block text states using the same fitting and held-out images. Thus, the gain curve measures how scaling the visual message changes class information accessible in the text state.

I.1 VISUAL MESSAGES IN THE VISUAL-FFN MODEL

Figure 9 shows the final-layer analysis and the gain intervention for the visual-FFN model. Across layers 15 to 28, the native message stays between 49.4% and 73.8% accurate, while the output state falls from 50.6% at layer 15 to 28.4% at layer 28. The native message is 15.5% of the text-state norm at layer 15, but at most 1.6% after layer 17. For the SigLIP2 reference, it is between 5.1% and 13.3%, and the output state stays between 32.6% and 45.0% (Table 8).

Table 8: **Visual messages across layers in the visual-FFN model.** Measured at the last prompt token, as in Table 11. Write is the norm of the visual message relative to the incoming text state. Message and text give 100-class nearest-centroid accuracy. All values are percentages.

	Native visual FFN			SigLIP2 reference		
Layer	write	message	text	write	message	text
15	15.5	55.0	50.6	5.9	30.6	35.0
16	2.2	49.4	47.4	6.1	34.4	34.2
17	3.2	52.6	48.2	7.2	31.8	32.6
18	0.7	60.2	45.4	10.4	43.4	39.4
19	0.1	67.8	43.8	6.7	32.2	36.4
20	0.6	59.4	42.6	5.1	25.0	33.6
21	1.1	69.4	40.4	13.3	47.4	41.4
22	1.6	64.4	36.2	11.4	50.2	43.2
23	0.2	72.2	31.8	8.5	53.6	42.6
24	0.9	63.4	29.8	6.6	47.0	41.8
25	0.0	61.4	28.8	10.6	48.0	42.8
26	0.1	64.6	29.2	13.3	47.4	44.4
27	0.0	58.4	28.6	9.1	55.0	45.0
28	0.7	73.8	28.4	6.3	42.8	44.6



(a) Visual message and text state (b) Gain intervention

Figure 9: **Visual messages in the visual-FFN model.** (a) Nearest-centroid accuracy of the visual message and of the output text state at the final layer, for the native model and the SigLIP2 reference. The native message has a norm of only 0.735% of the incoming text state's norm. (b) Multiplying only the visual message by g raises the accuracy of the output state from 28.4% to 64.2%.

J READER ARCHITECTURES AND FACTORIZATION

The reader experiments use the visual attention distilled backbone. Every reader variant starts from the same distilled image-specific computation and differs in how the visual message reaches the text state. The model weights are updated during training. Table 9 lists the added parameters and decoding speed. The main text calls the text-conditioned variant AVR. The non-adaptive variant is labeled

"Reader without adaptive gating" in Table 3 and "Visual reader" in the diagnostic tables below. The layer-wise writing and cut results use this non-adaptive checkpoint. The separate-attention control, with both new K/V and a new output projection, reaches 67.29% probe accuracy, 0.258 image gap, 32.44% coverage and 36.22% precision. It does not improve over the output-only control, which reaches 37.69% precision. The output-only control and the reader without adaptive gating route the same image-only message: both use the backbone's queries, keys and values, normalize attention over image tokens and write through a new output projection. The control computes these weights in a second attention pass, whereas the reader recovers them from the backbone's single attention call (below). In exact arithmetic, the two give the same image-only attention weights, so their rows in Table 3 differ in implementation and training run rather than in the routing mechanism.

Two-pass readers. The separate-attention reader adds a second attention from text queries to image keys only, with its own image normalization and its own key, value and output projections:

$$\alpha_{ij} = \frac{\exp(q_i^\top k_j/\sqrt{d_k})}{\sum_{m\in\text{image}}\exp(q_i^\top k_m/\sqrt{d_k})}, \qquad M_{\text{image}} = W_O^{(\text{sep})}\Big(\sum_{j\in\text{image}} \alpha_{ij} v_j^{(\text{sep})}\Big). \tag{18}$$

The two factorized readers keep this second pass but share one part with the backbone.

1. With shared W_{KV} and its own W_O, the reader reuses the backbone key-value projections and learns a separate matrix $W_O^{(\text{image})}$ to map the aggregated visual values into text states.
2. With its own W_{KV} and the shared W_O, the reader learns separate key-value projections and maps their aggregated values into text states through the backbone output projection W_O.

The learned-gains reader also computes the second pass, but with the backbone's own query, key, value and output projections. It scales the image-only message of each head with a learned gain, 448 gains in total. Caption precision is 20.03%, compared with 20.71% for the control. The final gain values were not recorded, so this experiment does not determine whether a larger fixed scale would improve generation.

Visual reader and adaptive visual reader. For text position i and head h, let $\alpha_{ij,h}$ be the attention weight on image token j and $v_{j,h}$ its value vector. The image-only message is

$$\bar{v}_{i,h}^{\text{image}} = \frac{\sum_{j\in\text{image}} \alpha_{ij,h} v_{j,h}}{\max\Big(\sum_{j\in\text{image}} \alpha_{ij,h}, \epsilon\Big)}. \tag{19}$$

The text-conditioned gates and visual update are

$$g_i = 2\sigma(f_{\text{gate}}(z_i)), \qquad \Delta h_i = W_O^{\text{Reader}}\,\text{Concat}_h\Big(g_{i,h}\bar{v}_{i,h}^{\text{image}}\Big). \tag{20}$$

Here z_i is the text state entering attention, f_{gate} produces one logit per head and σ is the sigmoid function. In Equation 1, m_i denotes the concatenation of these per-head image-only messages, and the gates broadcast over the value dimensions within each head.

Both readers compute the image-only message within the backbone's single attention call, with no second softmax. For each key position j and head h, the value is extended with the masked value $b_j v_{j,h}$ and the image indicator $b_j = \mathbb{I}(j \in \text{image})$. One attention call therefore returns the usual mixed message, the weighted image-value sum and the image attention mass. The extended value width is padded to a multiple of eight for the attention kernel. Queries, keys, masks and attention scores are shared with the standard path.

The image-value sum is divided by $\max(p_{i,h}, \epsilon)$ as in Equation 19, with $p_{i,h} = \sum_{j\in\text{image}} \alpha_{ij,h}$, where ϵ is the smallest positive normal number of the floating-point format used for the mass. The denominator is thus clamped rather than shifted by a constant. AVR applies $g_i = 2\sigma(f_{\text{gate}}(z_i))$ to these normalized messages before concatenating heads and applying the separate output projection W_O^{Reader}. The gate input z_i is the hidden state entering the attention module, not the projected query. The gate broadcasts over each head's value dimensions and can attenuate or amplify the message, with values between zero and two. Zero logits give unit gates. The non-adaptive reader skips this multiplication and sends the normalized messages directly to its separate output projection.

The standard message passes through the original output projection. The reader output is added only where the query is a text token and at least one head has positive image mass. Image-token rows and text rows with no positive image mass receive the standard output alone. This addition occurs inside the attention module, whose output is then added to the residual by the decoder block. Our single-pass implementation widens the value vectors and adds computation beyond the standard attention output. In our measurements, the learned-gains reader still decodes faster than the two visual readers despite its second pass. Table 9 reports parameter counts and decoding speed.

Table 9: **Reader cost.** Parameters added to the visual attention model and cached decoding speed. Parameters of the two factorized readers follow from their projection shapes. Dashes mark values we did not record.

Reader	Added parameters	Tokens per second
No reader (control)	0	78.3
Learned gains, shared W_O	448	48.2
Own K/V, shared W_O	58.72M	–
Shared K/V, own W_O	58.72M	–
Separate attention	117.47M	39.1
Visual reader	58.72M	45.3
Adaptive visual reader	59.18M	40.7

K Language ability after visual distillation

We evaluate the text decoder and final normalization from the original fully distilled model, attached to the base Qwen3-0.6B embeddings and tied language head. The patch connector, visual branch and distillation readout are excluded. Its image tokens reach 80.29% ImageNet top-1 accuracy under the evaluation of Table 1. This is the original fully distilled checkpoint at step 12,520, distinct from the later fully shared caption run discussed in Section 4.1.

WikiText-2 (Merity et al., 2016) uses the full test split with non-overlapping 1,024-token windows, scoring 298,646 tokens. We evaluate all 10,042 HellaSwag (Zellers et al., 2019) validation examples, 2,376 ARC-Easy (Clark et al., 2018) test examples and 1,838 PIQA (Bisk et al., 2019) validation examples. Multiple-choice answers are ranked by mean token log-likelihood. Table 10 reports this fixed diagnostic protocol. All 309 effective text-route tensors in the visual attention distilled model match the base model bitwise, leaving its text-only function unchanged.

Table 10: **Full visual distillation changes the pretrained language function.** Task scores are percentages under the protocol above.

Model	WikiText NLL	Perplexity	HellaSwag	ARC-Easy	PIQA
Base Qwen3-0.6B	3.1740	23.9038	44.762	49.958	65.887
Full distillation	12.4785	262627.92	24.776	23.316	50.218

Generation after fully shared training. A separate fully shared model retains 75.77% ImageNet probe accuracy after caption training, but its caption NLL is 3.68, compared with 1.89 for the visual-FFN model. Of 300 sampled captions, 229 repeat the same four-token sequence at least five times. This is a different run from the original fully distilled checkpoint evaluated above.

L Visual writing across the backbone

We compare the visual attention control, the visual reader, and the matched external reference built from the corresponding visual attention distilled model. All three are evaluated after language training at step 22,667. The external reference uses the distilled visual backbone, not the SigLIP2 teacher. We use 100 ImageNet classes with 20 images per class, split into 1,500 fitting images and 500 held-out images. Image preprocessing and the prefix `A photo of` are shared. We measure the final text query at every decoder layer and report layers 14, 21 and 28. Image mass is the head-averaged

probability assigned to image keys by the original mixed softmax. The visual message includes both the base attention's image contribution through the text output projection and, where present, the actual additional reader output captured by a hook. Its norm is divided by the residual norm at the entrance to the decoder layer, before input normalization. Message and output-state classification each use a separately fitted nearest-centroid classifier. The maximum relative mean absolute errors in attention-output reconstruction are 0.00421, 0.00349 and 0.00287 for the control, visual reader and external reference, respectively. Image mass depends on how attention is spread over the whole sequence, so we compare it only between the control and the visual reader, which share the same backbone attention.

Table 11: **The visual reader increases visual writing despite lower mixed-attention image mass.** All entries except layer and model are percentages. Message and text accuracy use the 100-class held-out diagnostic.

Layer	Model	Image mass	Write/residual	Message acc.	Text acc.
14	Control	3.509	1.300	38.6	24.8
14	Visual reader	0.516	10.646	28.6	32.4
14	External	69.513	11.679	38.4	37.6
21	Control	0.301	0.219	58.2	40.0
21	Visual reader	0.126	14.052	63.2	57.8
21	External	91.020	29.018	57.6	51.2
28	Control	0.795	1.495	55.6	32.0
28	Visual reader	0.294	13.075	63.6	66.4
28	External	78.699	4.043	49.8	53.2

The reader's image-only normalization allows a large visual update even when mixed attention assigns little probability to image tokens. Text-state classification improves at each of these three depths, while message classification is lower at layer 14. These measurements describe image-class information after joint model adaptation. They do not isolate the direction of the output projection or test recovery of individual color, text or relation facts.

M LAYER CUTS IN THE TRAINED READER

We evaluate the same control, visual reader and external checkpoints on a fixed 12-image functional set from the training data source, with possible training overlap. Each example uses its target caption with either the correct image or a fixed cyclic replacement. NLL is weighted by the number of supervised tokens. The image gap is shuffled-image NLL minus correct-image NLL.

A prefix cut removes text-query to image-key attention in layers $[0, d)$, and a suffix cut removes it in $[d, 28)$, using zero-based indexing. Image rows and text-to-text attention remain active. Both the standard and additional reader paths receive the same mask. Each condition reconstructs the mask and uses math SDPA. We verify the intact model against an independent forward call and verify image invariance under full cuts for predictions made by text rows. The first caption token may be predicted by the last image row and is excluded from that invariance check.

Table 12: **Image dependence uses both halves of the trained backbone.** Each pair gives correct-image NLL and image gap on the fixed 12-image set.

Removed image access	Control	Visual reader	External
None	2.0437 / 0.0359	1.8977 / 0.4127	1.9184 / 0.3567
First 7 layers	2.0531 / 0.0304	1.9938 / 0.2923	2.0234 / 0.2929
First 14 layers	2.0862 / 0.0097	2.3689 / 0.0744	2.1244 / 0.1688
First 21 layers	2.0849 / 0.0001	2.4931 / 0.0279	2.2098 / 0.0650
Last 14 layers	2.0696 / 0.0159	2.4213 / 0.0503	4.2291 / 0.0293
All layers	2.0875 / −0.0006	2.5329 / 0.0005	2.2906 / −0.0003

The visual reader's gap falls from 0.4127 to 0.0744 when only the late half can read the image, and to 0.0503 when only the early half can do so, 18% and 12% of its intact gap. Its intact image dependence thus uses a path distributed across layers. With 20,000 image-level bootstrap resamples, the intact gap's 95% intervals are [0.0241, 0.0522] for control, [0.3541, 0.4903] for the visual reader and [0.3045, 0.4208] for external. These intervals measure variation among the 12 images, not training-seed variation. The cuts alter a trained model without retraining, and their effects need not add across layers.

N LAYER-WISE VISUAL REPRESENTATIONS

Figure 10 reports attentive probes fitted to frozen image-token states after the shared final RMSNorm. Each layer uses the full ImageNet-1K training and validation sets (1,281,167 and 50,000 images), one training epoch, learning rate 10^{-3}, global batch size 4,096 and seed 0. This protocol differs from the 100-class nearest-centroid diagnostic in Appendix L.

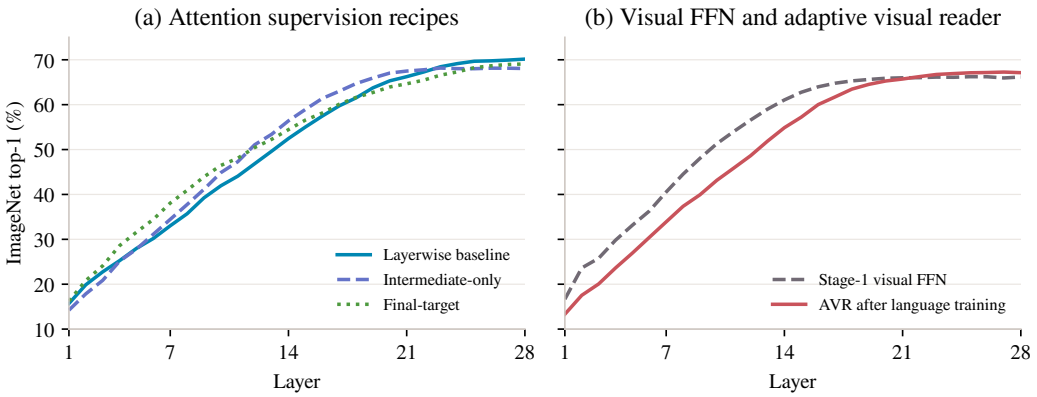


Figure 10: **Visual supervision changes where class information becomes readable.** (a) Three visual attention supervision recipes. (b) The distilled visual-FFN checkpoint and the adaptive visual reader after language training. Panel (b) shows different branches and training histories, not a before-and-after comparison.

At layer 14, intermediate-only supervision reaches 56.398%, compared with 52.460% for the layerwise baseline. At layer 28, their order reverses, to 68.056% and 70.150%. Compared with the layerwise baseline, the final-target recipe also raises earlier-layer accuracy while lowering final-layer accuracy. Supervision can therefore change the depth distribution of class information.

The adaptive visual reader checkpoint reaches 67.126% top-1 and 88.044% top-5 at layer 28, with peak top-1 of 67.262% at layer 27. It was initialized from the earlier final-only visual attention distilled checkpoint, whose full layer-wise curve is not available here. The layerwise-distilled baseline in panel (a) is a different source and cannot serve as its pre-training reference.

O LIMITATIONS

Our controlled experiments use a Qwen3-0.6B backbone, with a Qwen3-1.7B backbone for the direct native training pilot, a single SigLIP2 teacher and one training seed per configuration. The reading gap and the benefit of the adaptive visual reader may change with backbone scale, teacher choice and longer multimodal training, and we do not report variation across seeds. The reader ablations also change the output mapping and the size of the visual update together (Table 11), and the final learned gains were not recorded, so they do not separate the magnitude of the visual message from the direction in which it is written.

The analysis of public checkpoints relies on two proxies. ImageNet probes measure class information in image-token states, not all visual information that language may need. The attention cuts use MMVP, which has 150 image pairs, and some models answer few pairs correctly when intact, so retained-accuracy ratios are sensitive to small changes in correct-pair counts. Removing attention

also measures what the model loses under intervention, not what it would learn if trained without those connections, and alternative paths through prompt tokens can partly compensate for removed connections.

Image dependence is measured on a small fixed evaluation set drawn from the training source, and its bootstrap intervals reflect variation among images rather than across training runs. CAPEval scores depend on the judge model, and we replace the original judge. Our generation evaluation covers captioning, not question answering or other downstream benchmarks.

The adaptive visual reader narrows but does not close the gap to the matched external reference, and it adds parameters and decoding cost (Table 9). Finally, optic aphasia serves as an analogy for the dissociation between visual features and their use in generation. We do not claim that the model shares the mechanisms of the neurological condition.