CALLOSUM: ONE PATH BETWEEN SEEING, SEARCH, AND SPEAKING

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision-language models (VLMs) increasingly serve as retrieval encoders in deployment systems, where their embedding space becomes a fixed interface for lookup collections, rerankers, and downstream applications. Yet naive generative fine-tuning can silently alter that interface: rank lists drift, thresholds shift, and candidates change despite stable aggregate retrieval scores. Can a unified VLM add generation while keeping retrieval drift within a measured budget? We propose *Callosum*, a one-path adaptation method that treats the original retrieval representation as an anchor. *CAL–LO–SUM* captures our method's three principles: preserving a cache-side anchor, adapting through low-rank operators, and unifying multimodal retrieval with generation. We show that Callosum shifts the retrieval–generation frontier beyond architecture-matched baselines and approaches the generation quality of a frozen-anchor-plus-generator reference. To examine system stability, we compare full rankings, top-10 candidate sets, and threshold decisions, finding Kendall $\tau \approx 0.96$, 93% top-10 overlap, and merely 3.2% threshold crossings. Thus, Callosum enables VLMs to see, search, and speak along one path while preserving the retrieval infrastructure.

1 INTRODUCTION

Today's Artificial intelligence (AI) is expensive, static, and slow to change. To adapt significantly faster, a new structure is required. We envision a form of AI that is able to see, search, and speak through one unified path. It should understand an image, find relevant evidence in a massive multimodal corpus, and use that evidence to answer a question or explain a decision. Recent work has shown the superiority of this new architecture that morphs between modes (Li et al., 2023a; Liu et al., 2024a; Bai et al., 2025; Tschannen et al., 2025). While these functions share a visual-language backbone, they represent fundamentally distinct cognitive modes. Seeing builds a universal representation. Searching enforces a strict, exact match with a vast static memory. Speaking translates evidence into a fluid sequence of tokens. The ultimate challenge for foundation models is harmonizing these capabilities without compromising their individual roles.

This distinction becomes critical once a vast search memory is constructed. The stored image and text representations become a fixed map. Every system built on top of it relies on these exact coordinates to rank candidates and set similarity thresholds. However, a foundation model cannot merely search; it must also speak. When we fine-tune the model to generate fluent text and formulate complex answers, this generative upgrade inevitably shifts the query representations. A stable overall recall score can act as an illusion. It hides the fact that individual queries are now pointing to different coordinates, pulling different results, and silently breaking the original search pipeline. The core training problem is profound: how do we add fluid generative plasticity without shattering a massive, pre-computed search memory?

Most adaptation methods do not make this structural distinction explicit. Parameter-efficient tuning (Hu et al., 2022a; Dettmers et al., 2023; Liu et al., 2024b; Sung et al., 2022), adapter fusion, task arithmetic, and model merging (Pfeiffer et al., 2021; Ilharco et al., 2023; Yadav et al., 2023) treat the issue merely as generic catastrophic forgetting. Continual learning methods (Kirkpatrick et al., 2017; Rusu et al., 2016) aim to protect performance on previous tasks. These are valuable tools, but maintaining a search interface is much stricter than preserving a standard benchmark score. A typical benchmark only checks if a final generated answer is generally correct. Retrieval requires pointing

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

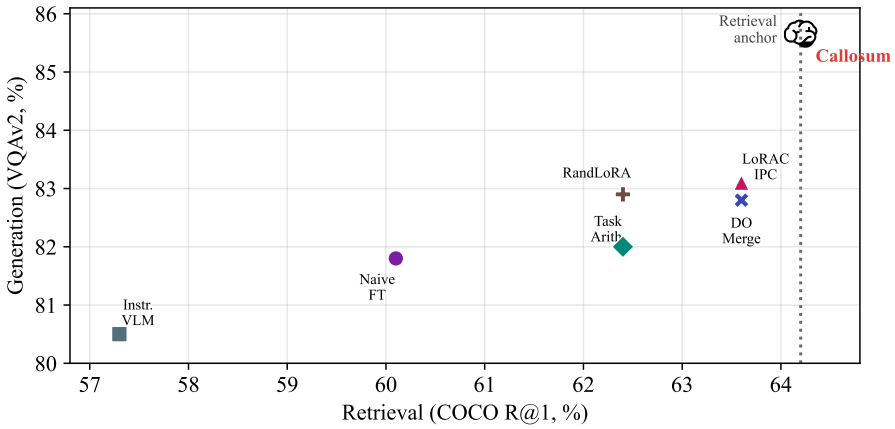

Figure 1: **Shattering the retrieval and generation frontier.** We plot representative methods from the main experimental protocol: the instruction VLM context row, naive fine-tuning (Joint-LoRA), Task Arithmetic, RandLoRA, DO-Merge, LoRAC-IPC, and Callosum. The dotted vertical line marks the strict retrieval anchor at 64.2 COCO R@1. Baselines force a zero-sum compromise, improving generation only by drifting left of the anchor. Callosum breaks this Pareto frontier, significantly elevating generative reasoning while perfectly returning to the exact anchor-level retrieval score.

to the exact same coordinates in a fixed space. If the query shifts, the search fails because the local neighbors change, even if average recall looks healthy. This creates the zero-sum compromise shown in Fig. 1. Training both tasks together causes gradient conflicts that pull the model away from its search anchor, while static merging forces a rigid compromise that limits both capabilities.

We formulate this retrieval-to-generation upgrade as anchor-constrained one-path adaptation. Given a retrieval anchor A and an instruction-tuned generator G from the same model family, our goal is a single model and serving path that preserves the deployed retrieval interface while exposing additional generative capacity through controlled shared and private updates. The name Callosum draws on the corpus callosum, which connects specialized hemispheres while coordinating unified cognition. We summarize this design as *CAL–LO–SUM*: *CAL* for cache-side anchoring, *LO* for low-rank operators, and *SUM* for semantic unification of multimodal retrieval and generation.

Callosum is trained in three stages. First, it fits a low-rank surrogate of the reference retrieval anchor inside the instruction-tuned VLM. Second, it decomposes the adaptation into shared directions aligned with the anchor basis and private residuals that provide task-specific generative capacity. Finally, it calibrates an input-adaptive, layer-wise gate that balances the anchor-aligned and private branches according to the input. This design allows one model to support retrieval and generation while explicitly controlling their interference.

To evaluate both objectives, we pair task-level metrics with a query-level retrieval-contract audit. Aggregate Recall@1 does not reveal whether individual queries receive the same candidates or scores, so we measure full-list rank correlation, top-k overlap, and score-threshold crossings. Across architecture-matched baselines, Callosum reaches 85.6% VQAv2 while matching the reference model's aggregate COCO R1 of 64.2, with Kendall $\tau \approx 0.96$ and 93% top-10 overlap. These results show that Callosum shifts the retrieval–generation frontier without rebuilding the corpus indices. The method targets bounded retrieval drift; exact neighbor or threshold invariance still calls for a two-path system.

Our contributions are threefold:

- We formulate the retrieval-to-generation upgrade as anchor-preserving capability grafting, fundamentally shifting the perspective on retrieval and generation from conflicting tasks to distinct cognitive modes.
- We propose Callosum, a neurologically inspired architecture that mathematically isolates and dynamically unifies representation stability and generative plasticity through a layered, three-stage gating mechanism.

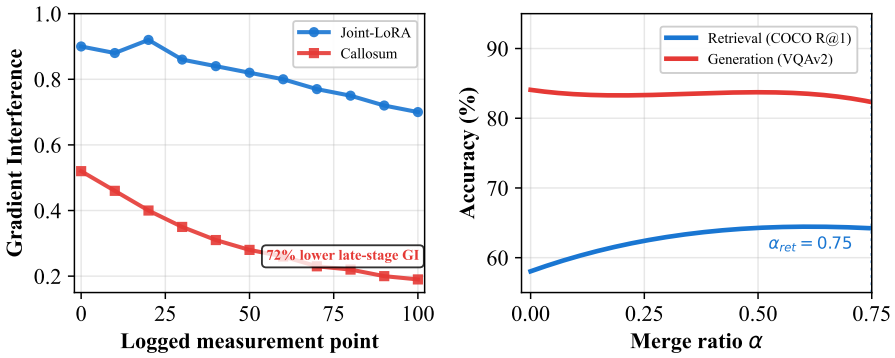

Figure 2: **Gradient interference and merge-ratio trade-offs.** (a) Gradient interference, measured as $1 - \cos(\nabla_\theta \mathcal{L}_{ret}, \nabla_\theta \mathcal{L}_{gen})$ on held-out mixed batches, is lower for Callosum than Joint-LoRA throughout training. (b) Static merge ratios expose a retrieval-generation compromise: retrieval favors higher anchor weight while generation favors lower anchor weight. The shaded interval represents the operating compromise when using one α for all inputs.

- We pioneer a strict evaluation protocol that exposes hidden pipeline failures beyond the illusion of aggregate metrics, establishing a new stability standard for continuously evolving unified models.

2 Investigation of the Retrieval-Generation Trade-off

A retrieval VLM is more than a checkpoint: its image-text embedding space determines which candidates later queries receive. The question we study is practical: how much generation can we add without changing that reference space? We use the architecture-matched baselines in Figs. 1 and 2 as diagnostics rather than only as competitors, since they reveal where current adaptation recipes break.

We start from an architecture-matched retrieval anchor and keep the corpus-side image and text indices frozen. Each baseline updates only the query-side adapted path, then we evaluate retrieval on COCO/Flickr and generation on VQAv2 and hallucination-oriented metrics. For the interference plot, we measure the cosine conflict between retrieval and generation gradients on held-out mixed batches during training. For the merge-ratio plot, we sweep a static anchor-to-generation weight on validation data and record how retrieval and VQA respond. Full directional retrieval, seed variation, and exact sweep values are reported in the supplement.

Fig. 1 shows three recurring failure modes. Naive instruction tuning learns to answer but moves the query embeddings away from the fixed corpus representations. Static adapter merging recovers some retrieval by mixing retrieval and generation adapters, but it must use one global ratio for all inputs. A strict separation penalty protects the anchor span more directly, yet it can also block useful transfer from the retrieval-trained representation. These behaviors form a narrow frontier: better generation usually comes with retrieval drift, while stronger preservation leaves generation underdeveloped.

Fig. 2 explains why the frontier appears. During joint tuning, retrieval and generation gradients pull on incompatible directions. During static merging, retrieval favors a higher anchor weight while generation favors a lower one. The issue is not only insufficient parameter capacity, but also the lack of a structured way to decide which update directions should reuse the anchor and which should remain private. This motivates the design principle behind Callosum: preserve the anchor-aligned subspace, allocate non-anchor capability to private residuals, and recombine the two adaptively for each input.

3 Callosum

Callosum formulates VLM adaptation as a constrained retrieval-to-generation problem under a fixed corpus-side retrieval interface. We initialize from a reference retrieval model A and an instruction-tuned generator G from the same model family. The reference parameters and corpus indices

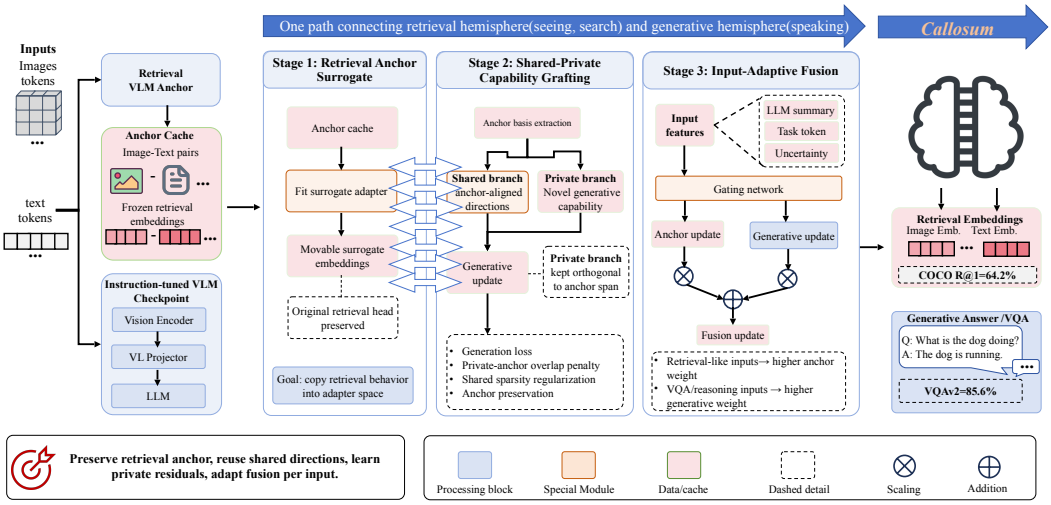

Figure 3: **An overview of Callosum and a Callosum deployment example.** The retrieval and generation paths are coordinated through a shared anchor-aware interface. Stage 1 trains a surrogate to the frozen retrieval anchor (eq. (2)). Stage 2 trains shared and private updates around its basis with an anchor-separation penalty (eqs. (3), (4) and (6)). Stage 3 trains an input-adaptive gate while branch updates remain frozen (eqs. (8) to (10)).

define fixed targets for optimization and evaluation; Callosum's trainable parameters are learned in three consecutive stages, and the resulting model is evaluated against those references. Retrieval is evaluated through two fixed fixed corpus-side indices: text queries search the image index for text-to-image retrieval, while image queries search the text index for image-to-text retrieval. Only the query-side computation is adapted during training; the corpus indices and their embeddings remain fixed in every comparison. At the model level, Callosum comprises four components: a vision encoder E_v, a text/language path E_t, modality projection heads P_v, P_t, and a large language model LLM. We adapt attention, projection, merger, and query-side modules while keeping external anchor heads and corpus embeddings frozen.

3.1 Stage 1: Retrieval Anchor Surrogate

Given the reference checkpoint A, we snapshot its image-text embeddings into a reference cache $\mathcal{D}_a = \{(I_i, T_i, z^a_{\text{img}}(I_i), z^a_{\text{txt}}(T_i))\}_{i=1}^{N_a}$. Before comparison, every image and text embedding is L2-normalized, $\bar{z} = z/\|z\|_2$. We measure anchor matching with

$$\mathcal{D}_{\text{anchor}}(z) = \frac{1}{2}\left(\|\bar{z}_{\text{img}} - \bar{z}^a_{\text{img}}\|_2^2 + \|\bar{z}_{\text{txt}} - \bar{z}^a_{\text{txt}}\|_2^2\right). \tag{1}$$

Next, we fit ΔW^a_l inside the instruction-tuned VLM adapter space by matching the surrogate embeddings z^s to the frozen anchor embeddings on the cache:

$$\mathcal{L}_{\text{sur}} = \mathcal{D}_{\text{anchor}}(z^s). \tag{2}$$

This surrogate is a movable copy of the reference representation inside the unified adapter. The external reference anchor supplies fixed cache targets and is not a runtime branch. The fitted surrogate itself is the internal *anchor branch*, with $\Delta W^a_l = B^a_l A^a_l$, and participates in the final forward computation in Eq. (9). The phrase *reference anchor* denotes the external checkpoint and fixed corpus-side indices used for comparison.

3.2 Stage 2: Shared-Private Capability Grafting

Callosum decomposes the generative update at layer l as

$$\Delta W_l^g = U_l S_l + R_l, \tag{3}$$

where $\Delta W_l^a \in \mathbb{R}^{d_{\text{out}} \times d_{\text{in}}}$ is the fitted anchor update (Stage 1), $U_l \in \mathbb{R}^{d_{\text{out}} \times r_a}$ is an orthonormal basis spanning the dominant left-singular subspace of ΔW_l^a, $S_l \in \mathbb{R}^{r_a \times d_{\text{in}}}$ is a learnable *shared* coefficient matrix whose left dimension matches the anchor surrogate rank r_a (a hyperparameter of the SVD truncation in Stage 1), and $R_l = B_l^p A_l^p$ is a *private* residual with $B_l^p \in \mathbb{R}^{d_{\text{out}} \times r_p}$ and $A_l^p \in \mathbb{R}^{r_p \times d_{\text{in}}}$. We compute U_l independently for each adapted projection and keep it frozen during graft training.

The decomposition is trained under two complementary constraints:

$$\mathcal{L}_{\text{priv}} = \frac{1}{|\mathcal{L}|} \sum_{l \in \mathcal{L}} \frac{\|U_l^\top R_l\|_F^2}{r_a d_{\text{in}}}, \tag{4}$$

$$\mathcal{L}_{\text{share}} = \frac{1}{|\mathcal{L}|} \sum_{l \in \mathcal{L}} \frac{\|S_l\|_1}{r_a d_{\text{in}}}. \tag{5}$$

The first term keeps the private branch from overlapping the anchor basis. The second encourages the shared branch to reuse only a sparse subset of anchor directions rather than collapsing into unrestricted overlap.

The generative learning objective becomes

$$\mathcal{L}_{\text{graft}} = \mathcal{L}_{\text{LM}} + \lambda_{\text{priv}} \mathcal{L}_{\text{priv}} + \lambda_{\text{share}} \mathcal{L}_{\text{share}} + \lambda_{\text{anchor}} \mathcal{L}_{\text{anchor}}. \tag{6}$$

where $\mathcal{L}_{\text{LM}}$ is the standard next-token prediction loss on the instruction-tuning dataset. To preserve the reference retrieval outputs during training, we add an explicit anchor-preservation term:

$$\mathcal{L}_{\text{anchor}} = \mathcal{D}_{\text{anchor}}(z^u). \tag{7}$$

3.3 Stage 3: Input-Adaptive Fusion

Static merging assumes every layer and every example should use the same anchor-to-generation trade-off. We instead learn a layer-local gating network

$$\alpha_l(x) = \sigma\big(w_l^\top \phi_l(x) + b_l\big) \in (0, 1), \tag{8}$$

where $\phi_l(x)$ is a two-layer GELU MLP. Its input combines pooled prefix language features, pooled vision features, a prompt-format code, a cache-uncertainty feature, and a modality mask. Answer tokens are excluded from pooling. The gate is computed once per input during prefill and held fixed during decoding. Missing-modality features are represented by zero vectors together with the corresponding modality mask. The exact pooling, cache-uncertainty, and feature-normalization settings are given in the supplementary reproducibility section.

The final Callosum update is implemented by weighting branch outputs rather than dynamically materializing a different matrix for each example:

$$y_l = W_l h_l + \alpha_l(x) B_l^a A_l^a h_l + (1 - \alpha_l(x)) \left(U_l S_l h_l + B_l^p A_l^p h_l\right). \tag{9}$$

This form is batch-compatible because both branch outputs are computed once and the scalar gate is broadcast over the hidden dimension. All three stages are training stages. Stage 1 trains ΔW^a to the anchor cache. Stage 2 trains S_l and R_l with the gate fixed at $\alpha_l(x) = 0.50$. Stage 3 freezes the branches and trains the gate on

$$\mathcal{J}_{\text{val}} = -\mathcal{L}_{\text{LM,val}} - \eta \cdot \mathcal{L}_{\text{ITC,val}} - \tau \cdot \text{Drift}_{\text{ret,val}}. \tag{10}$$

which conservatively penalizes validation language loss, retrieval contrastive loss, and drift from the reference retrieval outputs. For a validation batch B of image and text queries, the drift term is the differentiable normalized embedding distance

$$\text{Drift}_{\text{ret,val}} = \frac{1}{|B|} \sum_{x \in B} \left(1 - \bar{z}^u(x)^\top \bar{z}^a(x)\right), \tag{11}$$

where the reference representation supplies the fixed comparison target. This validation loss uses a differentiable embedding-distance surrogate rather than discrete retrieval decisions, so gradients can train the gate. The numerical weights and implementation details are reported in the supplementary reproducibility section. The fixed reference model defines training targets and validation comparisons, not a second runtime branch.

Table 1: Main architecture-matched results. F_{anchor} and F_p measure retrieval drop relative to the reference anchor and pre-graft reference, respectively.

	RETRIEVAL		GEN. / ROBUST.					DROP	
METHOD	COCO	FLICKR	VQAV2	GQA	POPE-R	POPE-P	POPE-A	F_{ANCHOR}	F_p
Reference (not an adapted baseline)									
INSTR. VLM (CTX.)	57.3	73.2	80.5	62.5	85.9	–	–	–	–
Exact-interface control									
FR-HEAD ($r = 43$)	64.2	79.8	84.1	66.0	87.7	85.0	80.3	0.0	3.7
Architecture-matched baselines									
JOINT-LORA	60.1	75.0	81.8	64.3	85.4	83.0	77.1	6.4	9.9
SEQ.-LORA	61.7	76.2	81.5	64.1	85.7	83.4	77.6	3.9	7.5
ORTHO-LORA	63.0	77.6	82.3	64.6	85.9	83.7	77.9	1.9	5.5
TASK ARITH.	62.4	77.0	82.0	64.8	86.0	83.6	78.0	2.8	6.4
TIES-MERGE	63.1	78.1	82.7	65.3	86.3	84.1	78.7	1.7	5.4
DORA	61.5	76.8	82.8	65.2	86.2	83.9	78.3	4.2	7.8
RANDLORA	62.4	77.5	82.9	65.4	86.5	84.3	78.8	2.8	6.4
DO-MERGE	63.6	78.6	82.8	65.1	86.4	84.2	78.5	0.9	4.6
PROGLORA	63.3	78.3	82.6	64.9	86.1	83.8	78.1	1.4	5.1
PEARL	63.4	78.5	82.7	64.8	86.2	84.0	78.2	1.2	4.9
LORAC-IPC	63.6	78.8	83.1	65.5	85.8	83.5	77.6	0.9	4.6
CALLOSUM	**64.2**	**79.8**	**85.6**	**68.1**	**89.3**	**87.2**	**82.5**	**0.0**	**3.7**

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Benchmarks. Retrieval is measured on COCO and Flickr30K (Lin et al., 2014; Young et al., 2014; Plummer et al., 2015), with FashionIQ used as a product-domain stress test (Wu et al., 2021). Generation is evaluated on VQAv2, GQA, and ChartQA (Goyal et al., 2017; Hudson & Manning, 2019; Masry et al., 2022). Robustness and hallucination are evaluated on VizWiz, OK-VQA, POPE, and MMHal. Exact split choices, evaluator details, and overlap checks are reported in the supplement.

Metrics. Retrieval quality is reported with Recall@1 (R@1), generation with task accuracy, and hallucination robustness with POPE/MMHal scores. We distinguish the reference-anchor drop from the stronger pre-graft reference gap:

$$F_{\text{anchor}} = \frac{R_{\text{anchor}} - R_{\text{after}}}{R_{\text{anchor}}} \times 100\%, \qquad F_{\text{pre}} = \frac{R_{\text{pre}} - R_{\text{after}}}{R_{\text{pre}}} \times 100\%.$$

Here $R_{\text{anchor}} = 64.2$ is the frozen-reference retrieval score and $R_{\text{pre}} = 66.7$ is a stronger retrieval-only same-family reference, not the context-only instruction VLM row in Table 1.

Frozen-index protocol. COCO uses the Karpathy 5K test split and Flickr30K uses the standard 1K test split, both with bidirectional R@1. Corpus-side image and text indices are fixed reference indices for evaluation. After training, only the query-side encoder changes. Hyperparameter selection and gate training use disjoint validation data and do not access evaluation splits.

Baselines. The main comparison uses architecture-matched, same-budget methods: Joint-LoRA, Sequential-LoRA, Orthogonal-LoRA, Task Arithmetic, TIES-Merging, DoRA, RandLoRA, DO-Merging, ProgLoRA, PEARL, and LoRAC-IPC. The off-the-shelf instruction-only VLM context row is not a drop-comparable adapted baseline. FR-Head is a bit-exact, parameter-matched frozen-retriever control with a side decoder; the separate Two-Path system and further matched variants appear in the supplement.

Statistical reporting. Main architecture-matched rows use five seeds ($\{42, 123, 2024, 7, 999\}$). Table 1 reports compact means, with confidence intervals, paired bootstrap summaries, and cache-feature matched comparisons in the supplement.

Implementation. The main model uses a retrieval-tuned anchor and a same-family instruction generator. The two corpus-side indices are frozen throughout evaluation. Anchor/shared/private ranks are 8/8/24, and the trainable modules cover the query-side attention and projection updates plus

Table 2: Retrieval reference gap under fixed references. F_{anchor} measures reference-anchor preservation ($R_{\text{anchor}} = 64.2$). F_p is a reference gap relative to the pre-graft upper bound ($R_{\text{pre}} = 66.7$), since the pre-graft checkpoint was never exposed to generative adaptation.

Method	R_a	R_p	R'_a	$F_{\textbf{anchor}}$	F_p
Frozen anchor only	64.2	66.7	64.2	0.0%	3.7%
Joint-LoRA	64.2	66.7	60.1	6.4%	9.9%
Sequential-LoRA	64.2	66.7	61.7	3.9%	7.5%
Orthogonal-LoRA	64.2	66.7	63.0	1.9%	5.5%
DoRA	64.2	66.7	61.5	4.2%	7.8%
TIES-Merge	64.2	66.7	63.1	1.7%	5.4%
Callosum	**64.2**	**66.7**	**64.2**	**0.0%**	**3.7%**

the layer-wise gate. The backbone, external anchor, corpus embeddings, and retrieval heads remain frozen after initialization. Training follows the three consecutive stages described in the method, with 2,000, 6,000, and 2,000 updates, global batch size 32, AdamW, bf16 mixed precision, gradient checkpointing, and gradient clipping at 1.0. The 38.914M unique parameters trained across stages are 0.54% of the 7.2B backbone or 0.469% of the full 8.292B checkpoint. Gate features are computed from the prefix during prefill and then reused during decoding, so the gate does not inspect answer tokens or introduce a second runtime branch.

4.2 Main Results

Table 1 shows that Callosum restores reference-level aggregate retrieval performance (COCO R1 = 64.2) while improving generation and hallucination robustness. Among the adapted baselines, DO-Merging best preserves retrieval (63.6 COCO R@1, $F_{\text{anchor}} = 0.9$) and LoRAC-IPC gives the highest VQAv2 score (83.1). Callosum reaches 85.6 VQAv2, 2.5 points above LoRAC-IPC and 1.5 points above the bit-exact FR-Head control. This is a measured retrieval–generation trade-off, not an absolute evaluation-dominance claim. Tables 30 and 31 report cache-matched, two-path, and continual-learning alternatives.

The margin is meaningful because the comparison keeps the retrieval-side resources fixed while changing only the query path. Callosum therefore improves generation without receiving credit for a refreshed corpus representation. The result also separates capacity from routing: matched baselines can recover part of the retrieval score through static mixing, but they do not reach the same VQAv2 level while retaining the anchor-level aggregate score. The exact-interface controls remain useful as a reference point, since their unchanged retrieval outputs come from keeping retrieval computation separate rather than learning a compromise inside one model.

Beyond the aggregate scores in Table 1, the supplementary retrieval analysis shows why the headline is still limited: rounded aggregate matching coexists with rank, threshold, and reranker changes. FR-Head and Two-Path remain bit-exact controls. Tables 23 and 24 report the detailed query-level comparison, while Tables 4, 5, 12 and 21 provide held-out robustness checks, hallucination-source decomposition, gate behavior under format perturbations, and second-backbone/domain stress tests. Fig. 2 adds rank-allocation sensitivity.

Table 2 separates three distinct effects that explain *why* Callosum achieves this result. Freezing the anchor preserves the retrieval interface exactly but does not provide a unified generative upgrade. A strict separation penalty reduces retrieval loss compared to Joint-LoRA but leaves generation under-optimized. The shared-private decomposition recovers transfer, and the adaptive gate further reduces retrieval drift by routing retrieval-style inputs toward the fitted surrogate branch.

Reference-gap interpretation. In this paper, F_{pre} is termed a *reference gap* because $R_{\text{pre}} = 66.7$ is a pre-graft retrieval-adapted checkpoint that was never exposed to generative adaptation, so the 3.7% difference does not constitute forgetting of a capability held by the reference model. Rather, it measures how far the grafted model falls relative to a stronger retrieval-only reference from the same family. We study adding generation while keeping the legacy embedding space fixed, rather than preserving the best retrieval checkpoint available in hindsight. Using the stronger R_{pre} checkpoint would require a separate interface migration and likely corpus re-indexing. The reference anchor

($R_{\text{anchor}} = 64.2$) and the pre-graft reference ($R_{\text{pre}} = 66.7$) are both fixed. Callosum approximately matches the reference anchor on rounded aggregate R@1 while remaining 3.7% below the pre-graft reference. F_{anchor} reflects seed-level post-graft R@1 relative to the fixed reference anchor, then formatted to one decimal place in the table. It is the relevant metric for the frozen-index legacy-interface setting.

4.3 Ablation Study

Fig. 4 shows that each component contributes differently to the final retrieval-generation trade-off under five-seed, independently re-tuned ablations. Removing the anchor surrogate distillation drops retrieval to 59.4 COCO R@1 and raises the pre-graft retrieval drop to 10.9%. Removing the anchor loss mainly hurts retrieval preservation while generation stays near-optimal at 83.6 VQAv2. Removing shared-private decomposition hurts both: w/o shared branch drops VQAv2 to 82.4, while w/o private branch drops VQAv2 to 81.5. The separately retrained fixed-gate ablation reaches 64.0 COCO R@1 and 83.1 VQAv2, while the independent global-merge sweep in Table 14 reports its own values for each fixed α. The companion panel shows why this matters: static merging forces one compromise across inputs, whereas adaptive fusion tracks the retrieval-generation balance more flexibly. The ablations support a cumulative mechanism interpretation: no single removed component fully explains the full-model trade-off, while the largest retrieval degradation comes from removing anchor-surrogate distillation and the largest generation degradation comes from removing the private branch. Tables 8, 15 and 16 in appendix O provide full numeric ablations and mechanistic diagnostics including CKA, nearest-neighbor overlap, and anchor embedding cosine analysis.

4.4 Robustness, Hallucination, and Generalization

Callosum improves held-out robustness and hallucination resistance by combining anchor reuse with private generative capacity. The full VizWiz, OK-VQA, ChartQA, and MMHal results, together with the component-wise hallucination analysis, are in Tables 4 and 5.

The gate is input-adaptive within the tested prompt distribution, but it is not invariant to its features. Removing prompt-format features lowers VQAv2 from 85.6 to 81.8 and COCO R@1 from 64.2 to 63.6. Retraining with mixed paraphrased templates recovers 83.9 VQAv2 and 64.1 COCO R@1. Additional format, language, query-perturbation, backbone, and domain-shift tests are reported in Tables 12 and 21. Together, these results show that anchor preservation and adaptive fusion are complementary, while performance remains dependent on the anchor and the calibration distribution.

The same pattern holds across the stress tests: the unified path retains its advantage when retrieval and generation must share computation, while exact-interface controls retain their advantage when every candidate and score must remain unchanged. This makes robustness a property of the use setting and calibration data, rather than a single scalar score.

This qualification matters for interpreting the headline numbers. The prompt-feature ablation changes both accuracy and retrieval, which indicates that the gate uses task-format information as part of its routing decision. Recovery after paraphrase retraining shows that this dependence can be reduced by matching the training distribution to the expected prompts. The backbone and product-domain tests preserve the method ordering but reduce the margin, so anchor quality and calibration population remain practical determinants of the final trade-off.

The extended analyses also clarify what the method is not claiming. Callosum does not make the adapted query path identical to the frozen anchor, nor does it eliminate the need for calibration when exact thresholds matter. Instead, the evidence points to a practical middle ground: the single model keeps aggregate retrieval at the anchor level, reduces the rank and threshold churn seen in other single-model baselines, and still improves generation and hallucination robustness.

Exact retrieval can be retained with FR-Head or Two-Path. At matched capacity, FR-Head reaches 84.1 VQAv2 versus Callosum's 85.6 (paired 95% CI for the gain $[1.31, 1.69]$), while their linked-request times are 130 and 138 ms; Two-Path takes 189 ms and 46.1 GB. In 2,000 linked retrieve–answer–retrieve episodes, Callosum reaches 62.9% end-to-end success versus 61.4% for FR-Head and 62.8% for Two-Path. The fixed-HNSW comparison finds 2.1% upgrade top-1 churn for Callosum versus 3.4% for matched DO-Merge. Protocols and complete results are in appendix R.5.

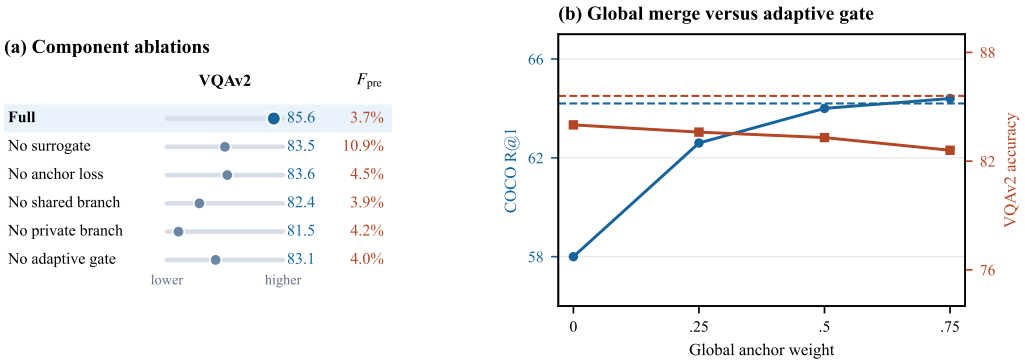

Figure 4: Component ablations and merge-ratio sensitivity. (a) The connected-dot plot compares VQAv2 accuracy (blue) and pre-graft retrieval gap F_{pre} (orange) across ablations; Callosum is highlighted. (b) Static anchor weights trade generation against retrieval; the sweep ends at the confirmed best retrieval weight $\alpha = 0.75$.

4.5 Choosing an Upgrade Strategy: Callosum or Two-Path

Although Callosum creates a potential scaling benefit for future VLM systems, two-Path is preferable when neighbor identity, score thresholds, and downstream reranking must remain unchanged. This exactness costs a second execution path, 46.1 GB, and 189 ms in our linked-request measurement. It also keeps retrieval and generation as separate computations, which can be useful when an existing search service has strict compatibility requirements. Callosum is preferable when one path, shared representations, and lower resource use matter and a bounded amount of drift can be calibrated. It uses 23.8 GB and 138 ms, reaches 85.6 VQAv2 versus 85.1 for Two-Path, and retains aggregate COCO R@1 at 64.2.

The unified path is deliberately a measured compromise rather than an exact-interface replacement. Callosum has 93.2% top-10 overlap with the anchor, 3.2% threshold crossings, and -0.70 nDCG change under the frozen reranker. In 2,000 linked episodes, it reaches 79.9% answer accuracy with returned evidence and 62.9% end-to-end success, compared with 79.4% and 62.8% for Two-Path. The added stress tests show why calibration remains part of the choice: a FashionIQ cache lowers COCO R@1 to 62.5 and raises rank-1 changes to 4.0%, while multilingual recalibration recovers 84.2 VQAv2 and 64.1 COCO R@1. Thus Callosum fits unified retrieval-generation workflows with monitored, calibratable drift. When any such change is unacceptable, Two-Path remains the appropriate design. Full cache, language, approximate-index, and domain-shift results are reported in appendix R.5 and the supplement.

5 Conclusion

We present Callosum for upgrading retrieval-oriented VLMs into unified retrieval-and-generation models while keeping the corpus representations fixed during training and evaluation. It improves generation and hallucination metrics over architecture-matched single-model baselines, remains competitive under prompt-robust gate retraining, and transfers to a second backbone and product-domain retrieval stress test with smaller but consistent margins. FR-Head and Two-Path remain exact-interface alternatives. Callosum gains 1.5 VQAv2 points over parameter-matched FR-Head and reaches 62.9% success in linked retrieve–answer–retrieve episodes. The practical boundary is clear: Callosum is appropriate when unified serving and bounded, calibratable drift matter, while Two-Path is preferable when exact neighbors or thresholds must remain invariant.

AI USE STATEMENT

Generative AI tools assisted with manuscript drafting and editing, literature search and synthesis, and LaTeX formatting. The authors take responsibility for the final content of this work, including AI-assisted text and artifacts.

REFERENCES

Paul Albert, Frederic Z. Zhang, Hemanth Saratchandran, Cristian Rodriguez-Opazo, Anton van den Hengel, and Ehsan Abbasnejad. RandLoRA: Full-rank parameter-efficient fine-tuning of large models. In *International Conference on Learning Representations*, 2025.

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, Junyang Lin, et al. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025.

Prashant Shivaram Bhat, Shakib Yazdani, Elahe Arani, and Bahram Zonooz. Parameter efficient continual learning with dynamic low-rank adaptation. *Transactions on Machine Learning Research*, 2026.

Hyuntak Cha, Jaeho Lee, and Jinwoo Shin. Co2L: Contrastive continual learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9516–9525, 2021.

B Cheng and J Lu. Multi-modal parameter-efficient fine-tuning via graph neural network. *arXiv preprint arXiv:2408.00290*, 2024.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. QLoRA: Efficient finetuning of quantized LLMs. In *Advances in Neural Information Processing Systems*, pp. 10088–10115, 2023.

Y Gao, J Lu, S Li, N Ma, S Du, Y Li, and Q Dai. Action recognition and benchmark using event cameras. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(12), 2023.

Y Gao, J Lu, S Li, Y Li, and S Du. Hypergraph-based multi-view action recognition using event cameras. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(10):6610–6624, 2024.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6904–6913, 2017.

Jinwei Han, Zhiwen Lin, Zhongyisun Sun, Yingguo Gao, Ke Yan, Shouhong Ding, Yuan Gao, and Gui-Song Xia. Anchor-based robust finetuning of vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26919–26928, 2024.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022a.

Weihua Hu, Rajas Bansal, Kaidi Cao, Nikhil Rao, Karthik Subbian, and Jure Leskovec. Learning backward compatible embeddings. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2022b.

Drew A. Hudson and Christopher D. Manning. GQA: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6700–6709, 2019.

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Suchin Gururangan, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. Editing models with task arithmetic. In *International Conference on Learning Representations*, 2023.

Young Kyun Jang and Ser-nam Lim. Towards cross-modal backward-compatible representation learning for vision-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1783–1792, 2025.

Y Jiang, W Lou, L Wang, Z Tang, S Feng, J Lu, H Sun, Y Pan, S Gu, H Su, et al. SCP: Accelerating discovery with a global web of autonomous scientific agents. *arXiv preprint arXiv:2512.24189*, 2025.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, 2017.

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*, 2023a.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 292–305, 2023b.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In *European Conference on Computer Vision*, pp. 740–755, 2014.

Y Lin, J Zhang, S Li, J Xiao, D Xu, W Wu, and J Lu. Event uskt: U-state space model in knowledge transfer for event cameras. *arXiv preprint arXiv:2411.15276*, 2024.

Y Lin, J Zhang, S Yu, J Xiao, and J Lu. FFEvent: Fast fourier-based knowledge transfer for event cameras. *Expert Systems with Applications*, pp. 130055, 2025.

Shimou Ling, Liang Zhang, Jiangwei Zhao, Lili Pan, and Hongliang Li. LoRA-based continual learning with constraints on critical parameter changes. *Pattern Recognition*, 171:112086, 2026.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26286–26296, 2024a.

Shih-Yang Liu, Chien-Yi Wang, Hongxu Yin, Pavlo Molchanov, Yu-Chiang Frank Wang, Kwang-Ting Cheng, and Min-Hung Chen. DoRA: Weight-decomposed low-rank adaptation. In *International Conference on Machine Learning*, 2024b.

Haodong Lu, Chongyang Zhao, Minhui Xue, Lina Yao, Kristen Moore, and Dong Gong. Adaptive rank, reduced forgetting: Knowledge retention in continual learning vision-language models with dynamic rank-selective LoRA. *arXiv preprint arXiv:2412.01004*, 2024a.

J Lu, F Yan, X Zhang, Y Gao, and S Zhang. PathoTune: Adapting visual foundation model to pathological specialists. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2024b.

J Lu, J Shi, Y Lin, F Yan, Y Gao, S Zhang, and X Wang. Hypergraph mamba for efficient whole slide image understanding. *arXiv preprint arXiv:2505.17457*, 2025.

J Lu, Z Kong, Y Wang, R Fu, H Wan, C Yang, W Lou, H Sun, L Wang, et al. Beyond static tools: Test-time tool evolution for scientific reasoning. *arXiv preprint arXiv:2601.07641*, 2026a.

J Lu, K Wang, Y Wang, Q Tang, H Zeng, X Chen, J Pi, S Deng, L Chen, et al. FinToolBench: Evaluating LLM agents for real-world financial tool use. *arXiv preprint arXiv:2603.08262*, 2026b.

Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. ChartQA: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the Association for Computational Linguistics: ACL*, pp. 2263–2279, 2022.

Jonas Pfeiffer, Aishwarya Kamath, Andreas Rücklé, Kyunghyun Cho, and Iryna Gurevych. AdapterFusion: Non-destructive task composition for transfer learning. In *Proceedings of the European Chapter of the Association for Computational Linguistics*, pp. 487–503, 2021.

Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2641–2649, 2015.

Vivek Ramanujan, Pavan Kumar Anasosalu Vasu, Ali Farhadi, Oncel Tuzel, and Hadi Pouransari. Forward compatible training for large-scale embedding retrieval systems. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

Andrei A. Rusu, Neil C. Rabinowitz, Guillaume Desjardins, Hubert Soyer, James Kirkpatrick, Koray Kavukcuoglu, Razvan Pascanu, and Raia Hadsell. Progressive neural networks. *arXiv preprint arXiv:1606.04671*, 2016.

Yantao Shen, Yuanjun Xiong, Wei Xia, and Stefano Soatto. Towards backward-compatible representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6368–6377, 2020.

Yi-Lin Sung, Jaemin Cho, and Mohit Bansal. VL-Adapter: Parameter-efficient transfer learning for vision-and-language tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5227–5237, 2022.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, et al. SigLIP 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025.

H Wan, C Yang, J Yu, M Tu, J Lu, D Yu, J Cao, B Gao, J Xie, A Wang, et al. DeepResearch Arena: The first exam of LLMs' research abilities via seminar-grounded tasks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2026.

Xiao Wang, Tianze Chen, Qiming Ge, Han Xia, Rong Bao, Rui Zheng, Qi Zhang, Tao Gui, and Xuanjing Huang. Orthogonal subspace learning for language model continual learning. *arXiv preprint arXiv:2310.14152*, 2023.

Y Wang, R Zhou, R Fu, S Cao, H Zeng, J Lu, S Fan, J Zhao, and L Pan. ASA: Training-free representation engineering for tool-calling agents. *arXiv e-prints*, 2026. arXiv:2602.04935.

Hui Wu, Yupeng Gao, Xiaoxiao Guo, Ziad Al-Halah, Steven Rennie, Kristen Grauman, and Rogerio Feris. Fashion IQ: A new dataset towards retrieving images by natural language feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

Prateek Yadav, Derek Tam, Leshem Choshen, Colin Raffel, and Mohit Bansal. TIES-merging: Resolving interference when merging models. In *Advances in Neural Information Processing Systems*, pp. 7093–7115, 2023.

F Yan, J Wu, J Li, W Wang, Y Chen, L Wei, J Lu, W Chen, Z Gao, J Li, H Li, et al. Pathorchestra: A comprehensive foundation model for computational pathology with over 100 diverse clinical-grade tasks. *npj Digital Medicine*, 8(1):695, 2025.

C Yang, J Lu, H Wan, J Yu, and F Qin. From what to why: A multi-agent system for evidence-based chemical reaction condition reasoning. In *International Conference on Learning Representations*, 2026.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014.

Yahan Yu, Duzhen Zhang, Yong Ren, Xuanle Zhao, Xiuyi Chen, and Chenhui Chu. Progressive LoRA for multimodal continual instruction tuning. In *Findings of the Association for Computational Linguistics: ACL*, pp. 2779–2796, 2025.

Shenghe Zheng, Hongzhi Wang, Chenyu Huang, Xiaohui Wang, Tao Chen, Jiayuan Fan, Shuyue Hu, and Peng Ye. Decouple and orthogonalize: A data-free framework for LoRA merging. *arXiv preprint arXiv:2505.15875*, 2025.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Haoqiang Duan, Hao Tian, Weijie Su, Jie Shao, Zhangwei Zhang, Mingze Gao, Erfei Li, Junting Yan, Yu Qiao, Lewei Wang, and Jifeng Dai. InternVL3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025a.

Yuhan Zhu, Xiangyu Zeng, Chenting Wang, Xinhao Li, Chunxu Liu, Yicheng Xu, Ziang Yan, Yi Wang, and Limin Wang. Freeret: Mllms as training-free retrievers. *arXiv preprint arXiv:2509.24621*, 2025b.