ATTNCAL: NATIVE ATTENTION CONSTRUCTION AND VISUAL CALIBRATION FOR MULTIMODAL MODEL MERGING

Anonymous authors
Paper under double-blind review

ABSTRACT

Attention logits depend jointly on query and key (Q/K) projections, so preserving each projection separately need not preserve their interaction. In multimodal merging, this interaction must also process visual features from a consolidated interface. AttnCal addresses both requirements through native attention construction and visual-conditioned calibration. It reconstructs joint rotary-position operators under the original head-width constraint, targeting the worst expert-relative reconstruction error within each head. The construction uses expert weights alone, provides computable objective bounds, and controls attention-logit error for fixed hidden inputs uniformly over relative position. AttnCal calibrates the constructed Q/K factors under a fitted visual map, then fixes the language weights while adapting a shared visual residual, retaining one static model and the original visual interface. Across four reproduced benchmark–backbone protocols, AttnCal achieves the highest final average score (FAA) among merged baselines. On six UCIT tasks with LLaVA-1.5-7B and 3,000 examples per task, it reaches 76.07 FAA, 82.72 continual average score (CAA), and 0.33 backward transfer (BWT), compared with 72.15, 81.60, and -1.52 for the strongest FAA baseline. These results connect a native-feasible attention target to the visual state under which the merged model is calibrated and deployed.

1 INTRODUCTION

Merging multimodal experts requires preserving specialist interactions at fixed attention width and adapting them to the visual features supplied at deployment. This couples representation and conditioning (Wang et al., 2024; Guo et al., 2025c). Given compatible experts and source data, we ask which query–key (Q/K) interactions the original architecture can represent, and under which visual state they should be calibrated.

Existing methods combine parameter updates or fit layer responses (Ilharco et al., 2022; Yadav et al., 2023; Jin et al., 2022; Chen et al., 2025a; Xu et al., 2025); however, the preservation target determines which specialist discrepancies the shared model protects (Zhang et al., 2026a). Under the shared rotary positional embedding (RoPE) convention used here, each Q/K row pair represents a rank-at-most-one complex operator. Directly averaging expert operators can leave this native-feasible family, whereas averaging Q/K factors omits the cross-expert covariance of the query and key deviations (Appendix D). These are different failures: operator averaging can exceed the native rank, while factor averaging stays native but changes the interaction being preserved. AttnCal instead constructs native-width factors to target the worst expert-relative operator error across the rotary frequencies of each head. The minimax criterion balances each expert's total error across the head, rather than protecting a different worst-case expert at each frequency. Native representability and expert balance are explicit parts of the construction.

Figure 1(a,b) illustrates the native interaction constraint; panel (c) examines visual conditioning with language weights fixed. Operator reconstruction selects an architecture-compatible starting point, and answer supervision teaches it to use the tasks' visual features.

AttnCal operationalizes this connection in two stages. Stage 1 constructs native Q/K factors, fits an auxiliary visual map, and calibrates Q/K while holding that map fixed. Stage 2 freezes the result-

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.

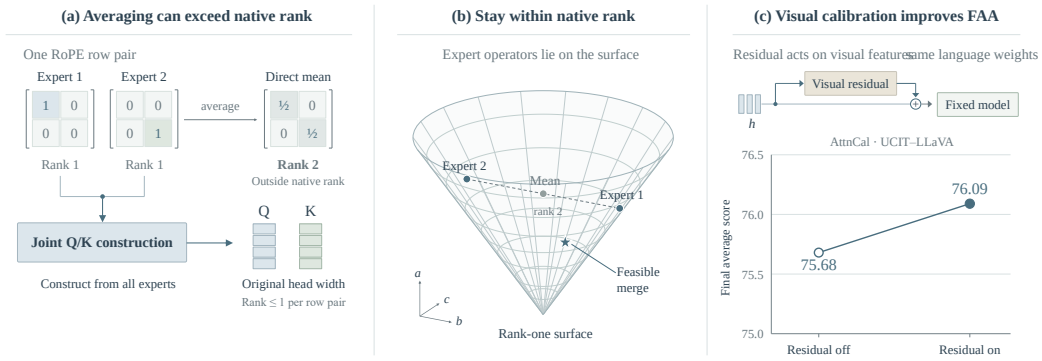


Figure 1: **Native-feasible attention construction and visual calibration address complementary parts of expert merging.** **(a)** For one rotary positional embedding (RoPE) row pair, two rank-one expert operators have a rank-two mean. Joint query–key (Q/K) construction instead retains the original head width and rank-at-most-one operator per pair. **(b)** The illustrative operator-space slice places the feasible merge on the rank-one surface and the direct mean outside it. The coordinates are schematic, not a neural-network loss landscape. **(c)** A residual corrects visual features h while language weights remain fixed. On six UCIT tasks with 3,000 examples each using LLaVA-1.5-7B, enabling this fitted residual raises final average score (FAA), the six-task mean, from 75.68 to 76.09. Residual off disables the final correction only at evaluation; both endpoints use the same task-calibrated language model.

ing language model and resumes visual fitting from the retained map. Across all four reproduced benchmark–backbone protocols, this pipeline obtains the highest FAA among merged baselines; on UCIT–LLaVA it also improves continual average score and backward transfer over the strongest FAA baseline. The residual-off comparison measures the deployed correction, while separate controls assess robustness across visual-state initializations. Our contributions are:

- With the non-Q/K reference weights fixed, we formulate a head-wise worst-expert Q/K operator objective across rotary frequencies, derive fixed-input representability and primal–dual bounds, and recover factors at the backbone's native head width.
- We couple native attention construction to task calibration through a shared visual state, fitting Q/K under that state and adapting one visual residual with language weights frozen, without changing the visual interface.
- The complete model attains the highest FAA among reproduced merged baselines in all four evaluated protocols. On UCIT–LLaVA it reaches 76.07 FAA, 82.72 CAA, and 0.33 BWT. All nine endpoints in the visual-handoff study exceed the strongest baseline, showing robust final performance across the tested visual-state choices.

2 Related Work

Multimodal continual learning and expert consolidation. Continual learning retains earlier capabilities through regularization, replay, and task-specific adaptation around shared parameters (Kirkpatrick et al., 2017; Buzzega et al., 2020; Wang et al., 2022; Hu et al., 2021). CoIN and MCITlib extend this setting to multimodal instruction sequences, with expert-based retention and analyses of superficial forgetting highlighting complementary aspects of capability preservation (Chen et al., 2024a; Guo et al., 2025c; Huai et al., 2025; Chen et al., 2025b). When specialists are already available, model merging turns retention into a consolidation problem. Model soups and task arithmetic combine weights or task vectors, Fisher merging accounts for parameter importance, and TIES resolves conflicting updates (Wortsman et al., 2022; Ilharco et al., 2022; Matena & Raffel, 2022; Yadav et al., 2023). Functional approaches instead specify computations to reproduce: RegMean fits layer responses, IterIS and RegMean++ account for model-dependent inputs, and Output-Space Projection studies output fitting within a declared merge family (Jin et al., 2022; Chen et al., 2025a; Nguyen et al., 2025a; Evans et al., 2026). Adaptive coefficients, specialist-preservation objectives,

and Pareto-oriented merging further examine the trade-off between aggregate quality and individual expertise (Yang et al., 2024; Zhang et al., 2026a; Chen & Kwok, 2024). These directions make the preservation target central to consolidation. AttnCal selects the query–key interaction as that target and studies how to approximate multiple expert interactions within the original attention width.

Attention structure and operator-level merging. Attention parameters, probability maps, and value readouts encode different aspects of the computation, as established by work on attention identifiability and functional equivalence (Brunner et al., 2019; Tran et al., 2026). RoPE adds a structured dependence on relative position through paired query–key rotations (Su et al., 2024), making the positional convention part of any operator representation or function-preserving transformation. MeRoTune learns RoPE-compatible, function-preserving Q/K reparameterizations for tunable two-model merging (Faroz, 2026). These studies connect parameter alignment to the composed attention computation, while leaving room for different approximation objectives and feasible ranks. AttnCal directly constructs expert-derived RoPE kernels under the native factorization: each rotary-frequency pair has a rank-at-most-one complex operator. It aggregates each expert's relative approximation error across the frequencies of a head before minimizing the largest expert error. Frequencies therefore share a head-level error budget instead of being assigned independent worst-case objectives. The accompanying primal–dual analysis characterizes the approximation achieved within that native feasible set, making both representability and expert balance explicit in the construction.

Functional distillation and visual feature calibration. Distillation provides a complementary route from expert computations to a student model. MiniLM and MiniLMv2 transfer attention relations, while ProDistill progressively distills expert behavior during merging (Wang et al., 2020; 2021; Xu et al., 2025). At the visual interface, MAny combines projector outputs through prototype-conditioned weights, and MVP and PMA learn instruction-aware or multimodal projector allocation (Gao et al., 2026; Jin et al., 2025; Zhang et al., 2026b). Low-rank adaptation and bottleneck adapters offer parameter-efficient mechanisms for modifying pretrained computations or their feature representations (Hu et al., 2021; Houlsby et al., 2019). These approaches place supervision at different levels, from intermediate relations to the answer distribution, and separate expert selection from adaptation of the selected features. AttnCal connects these perspectives through the visual conditioning state. Its operator objective defines the initial attention approximation, whereas answer-based calibration adapts Q/K and a shared visual residual in sequence. The visual parameters are retained across that handoff, with the original selection interface fixed. This focuses the comparison on whether language calibration and subsequent visual fitting benefit from sharing the visual state on which the language model was adapted.

3 METHODOLOGY

AttnCal addresses two coupled choices: what attention behavior a fixed-width model can preserve, and which visual representation its language calibration learns to process. It constructs native attention first, then carries a fitted visual state through task calibration (Figure 2). Stage 1 constructs native-width query–key factors with other language projections fixed, then calibrates Q/K under a fitted, fixed auxiliary visual map. Stage 2 freezes the resulting language weights W_E and adapts the retained visual map to that decoder.

Because attention depends on query–key interactions, we reconstruct their joint operator rather than treating the two projections independently. For expert i and rotary frequency f, pair the two real Q/K rows as $q_{if}, k_{if} \in \mathbb{C}^d$ and write the native rank-one operator $M_{if} = q_{if}k_{if}^*$, where d is the input-feature dimension and $*$ denotes conjugate transpose. A candidate $S_f = q_f k_f^*$ is chosen by the head-coupled minimax objective

$$L_i(S) = \frac{\sum_f \|S_f - M_{if}\|_F^2}{N_i}, \qquad \min_{\mathrm{rank}(S_f)\le 1\ \forall f} \max_i L_i(S), \quad N_i = \sum_f \|M_{if}\|_F^2 > 0.$$

Normalization balances expert scales; minimizing worst-expert error prevents good average agreement from masking a poorly preserved expert. The bilinear interaction explains why factor and operator averages differ: their exact difference is the cross-expert covariance $m^{-1}\sum_i (q_i - \bar{q})(k_i - \bar{k})^*$,

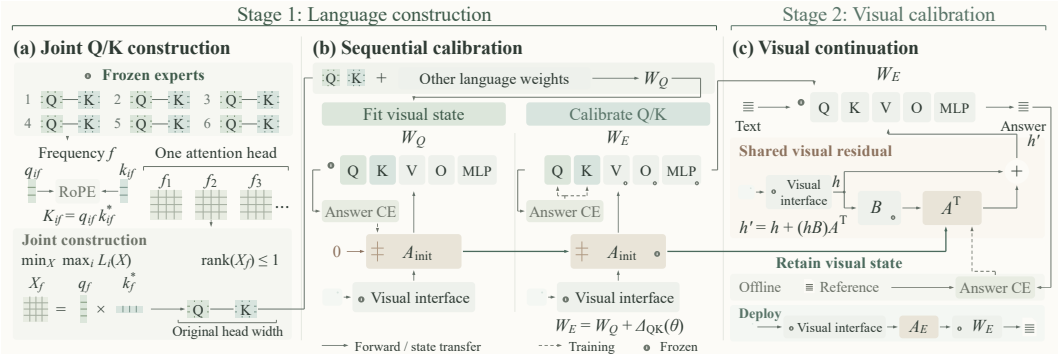


Figure 2: **AttnCal constructs native-width attention and carries a visual state through calibration. (a)** Frozen experts provide query and key factors at each rotary frequency. A shared construction balances their reconstruction errors while retaining the backbone's attention-head width. The small matrices depict rotary attention operators, not measured attention maps. **(b)** The constructed language weights first support an auxiliary visual fit. That visual state is then held fixed while the query and key weights are calibrated for the tasks. These two steps form Stage 1. **(c)** Stage 2 freezes the calibrated language model and starts a new visual fit from the retained visual state. It adds a learned correction to the existing visual features without replacing the visual interface. Answer loss trains the visual fits; the DCL–InternVL variant also uses a frozen-reference regularizer. Locks indicate frozen components, solid arrows show computation or state transfer, and dashed arrows show training.

where m is the expert count and bars denote factor means. Averaging factors discards this term. Operator averaging retains it, but the mean of two orthogonal rank-one projectors has rank two: its best native rank-one approximation loses half its squared Frobenius energy (Appendices D and E). To assess construction quality, let P_H be the optimal head objective and λ nonnegative expert weights summing to one. Define the normalized weight sum $a = \sum_i \lambda_i / N_i$ and weighted target $B_f = \sum_i \lambda_i M_{if} / N_i$. For any native-feasible candidate $\widehat{S}$, the largest singular value σ_1 gives a dual lower bound:

$$g_H(\lambda) = 1 - \frac{\sum_f \sigma_1(B_f)^2}{a} \;\le\; P_H \;\le\; \max_i L_i(\widehat{S}).$$

Completing the square yields the lower bound; candidate feasibility gives the upper bound. Their exact-arithmetic gap bounds objective excess for the constructed initialization, without assuming an exact finite-solver optimum. Under fixed hidden inputs, operator error also bounds attention-logit error uniformly over relative position. Appendix A gives the proofs, numerical checks and architecture-specific conditions.

Answer supervision then adapts this initialization within the same native Q/K parameterization to the visual features supplied to the decoder. For fixed visual features h, a source-derived orthonormal basis B, and fitted coefficients A, the Stage 2 residual is

$$h' = h + (hB)A^\top,$$

and A is fitted from the retained initialization with answer-token cross-entropy (CE) plus the protocol-specific frozen-reference Kullback–Leibler (KL) term. Retaining the visual initialization preserves the condition under which language was calibrated; the final residual then adapts features to that frozen decoder. Deployment retains one static language model and one shared visual residual, with no test-time teacher query or optimization. The exact objectives, loss weights, schedules, state-interaction identity and handoff controls are reported in Appendices A and C.

4 Experiments

4.1 Evaluation design

We ask what the native-width construction preserves, what the calibrated model achieves, and how its final visual correction affects answers. Operator and fixed-state attention diagnostics address the

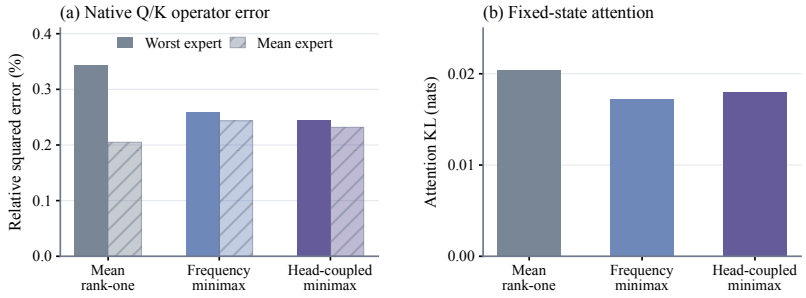


Figure 3: **Q/K construction targets and attention responses differ.** The three native-width query–key (Q/K) constructions are measured before task calibration; lower is better. **(a)** Solid and hatched bars show worst- and mean-expert relative squared operator error, averaged over 32 heads and 32 layers; each operator includes 64 rotary frequencies. Head-coupled minimax, the AttnCal construction, has the lowest worst-expert error, while normalized mean rank-one has the lowest mean error. **(b)** Mean expert-to-construction attention Kullback–Leibler (KL) divergence in nats on fixed states uses the first 128 ordered examples from each of six UCIT tasks and at most 16 answer positions per example. Frequency minimax has the lowest KL. Neither panel measures final task accuracy.

first question; complete task evaluations address the second; frozen-language residual controls address the third. The complete pipeline is thus evaluated separately from its initial Q/K construction. Method selection uses a distinct development population. Experiments and baseline reproductions use a shared pool of 32 NVIDIA A100 GPUs, with allocation varying by job.

Native-width feasibility does not decide which expert discrepancy a merge should protect. Figure 3 compares three Q/K constructions at the operator level and then probes their attention distributions on fixed states. Head-coupled minimax has the lowest measured worst-expert operator error, whereas normalized mean rank-one has the lowest mean-expert error. Frequency minimax gives the lowest attention divergence. These rankings distinguish AttnCal's worst-expert operator target from attention behavior at fixed hidden states.

4.2 Backbones and datasets

We evaluate LLaVA-1.5-7B (Liu et al., 2023; 2024) with CLIP (Radford et al., 2021) and InternVL (Chen et al., 2024b) on UCIT and DCL. UCIT draws 3,000 specified examples per task from ImageNet-R (Hendrycks et al., 2021), ArxivQA (Li et al., 2024), VizWiz-Captions (Gurari et al., 2020), IconQA (Lu et al., 2021), CLEVR-Math (Lindström & Abraham, 2022), and Flickr30k (Young et al., 2014). Caption scores use MCITlib's corpus-level suite, including BLEU and CIDEr (Papineni et al., 2002; Vedantam et al., 2015), rather than classification accuracy. Tables 1–4 separate the four benchmark–backbone settings; Appendix I details their populations and scoring.

4.3 Metrics

For the score $R_{i,j}$ on task j after stage i, with T tasks in total, the final average score (FAA) averages the final-stage task scores. Write $\bar{R}_i = i^{-1} \sum_{j=1}^{i} R_{i,j}$ for the mean task score at stage i. The continual average (CAA) averages these stage means, and backward transfer (BWT) measures subsequent change on earlier tasks. We use

$$\mathrm{FAA} = \frac{1}{T} \sum_{j=1}^{T} R_{T,j}, \quad \mathrm{CAA} = \frac{1}{T} \sum_{i=1}^{T} \bar{R}_i, \quad \mathrm{BWT} = \frac{1}{T-1} \sum_{j=1}^{T-1} (R_{T,j} - R_{j,j}).$$

CAA averages stage means, not all lower-triangular cells equally. BWT is in percentage points; all three metrics are higher-is-better. Positive BWT is an average improvement on earlier tasks from their first-learned to final checkpoints, not a comparison to separate Individual experts. Missing stage evaluations remain unavailable; Appendix I details aggregation and evaluation assets. Displayed results are rounded to two decimal places; reported differences are calculated from the unrounded scores.

4.4 BASELINES AND COMPARISON SCOPE

The four main tables compare our evaluated results within each benchmark–backbone setting. Alongside the common baselines, the LLaVA UCIT panel includes local IterIS, CSP, RegMean++, and PASs-MoE evaluations. AttnCal's complete UCIT and DCL matrices directly supply FAA, stage-balanced CAA, and BWT rather than inferring trajectory metrics from final endpoints. These comparisons assess the deployed system, while the stage-wise control below more directly examines the fitted visual residual.

Table 1: **UCIT performance on LLaVA-1.5-7B.** Rows use 3,000 evaluation examples per task. IN-R, Arxiv, VizWiz, Icon, CLEVR, and Flickr denote ImageNet-R, ArxivQA, VizWiz-caption, IconQA, CLEVR-Math, and Flickr30k; Source cites the original method. Final average score (FAA) averages final-task scores, continual average score (CAA) averages stage-wise means, and backward transfer (BWT) measures later change on earlier tasks. Scores are percentages and BWT is in points; higher is better. Daggers mark the unmerged zero-shot mean (†) and separate task experts (‡), for which CAA and BWT do not apply. Blanks are unavailable under this protocol.

Method	Source	IN-R	Arxiv	VizWiz	Icon	CLEVR	Flickr	FAA ↑	CAA ↑	BWT ↑
Zero-shot	(Guo et al., 2025c)	17.07	52.93	39.12	20.93	19.90	40.95	31.82†	—	—
Individual	(Guo et al., 2025c)	91.33	91.47	60.72	77.40	76.60	57.61	75.86‡	—	—
LoRA-FT	(Hu et al., 2021)	65.60	77.03	44.44	70.93	58.43	57.46	62.32	76.73	-16.57
O-LoRA	(Wang et al., 2023)	82.97	92.07	40.04	64.23	51.10	52.39	63.80	78.21	-7.77
MoELoRA	(Chen et al., 2024a)	61.50	77.57	43.77	70.77	59.83	57.32	61.79	76.90	-17.43
CL-MoE	(Huai et al., 2025)	61.60	77.77	44.22	68.63	54.37	57.65	60.71	75.86	-18.35
HiDE	(Guo et al., 2025a)	84.23	88.83	48.12	66.60	46.83	53.73	64.72	78.59	-7.14
DISCO	(Guo et al., 2025b)	87.53	92.63	46.62	71.23	60.90	56.23	69.19	81.39	-7.92
SiGMA	(Park & Kim, 2026)	82.43	91.13	54.47	70.30	58.50	56.02	68.81	80.61	-6.34
MAny	(Gao et al., 2026)	89.47	91.97	58.86	67.67	68.07	56.88	72.15	81.60	-1.52
IterIS	(Chen et al., 2025a)	69.90	87.23	43.23	65.47	63.13	49.45	63.07	75.44	-5.36
CSP	(Sun et al.)	85.60	92.03	46.73	71.80	65.20	56.95	69.72	78.45	-1.22
RegMean++	(Nguyen et al., 2025b)	75.83	88.80	44.70	64.90	54.83	54.78	63.97	76.25	-4.73
PASs-MoE	(Hou et al., 2026)	76.27	78.67	45.41	65.83	52.63	57.19	62.67	76.93	-12.00
AttnCal	Ours	89.37	92.10	64.79	75.13	76.53	58.51	76.07	82.72	0.33

Table 1 reports 76.07 FAA, 82.72 CAA, and 0.33 BWT for AttnCal, compared with 72.15, 81.60, and -1.52 for the matched MAny reproduction. Thus the 3.92-point endpoint gain is accompanied by a 1.12-point increase in the stage-balanced mean and a 1.85-point BWT improvement, showing that the advantage persists across the trajectory rather than appearing only in the final row. In particular, positive BWT means that earlier tasks improve on average between their first-learned and final checkpoints (Lopez-Paz & Ranzato, 2017; Riemer et al., 2018). The endpoint gain is also substantive rather than a single-task replacement: CLEVR-Math rises from 68.07 to 76.53, VizWiz-caption and IconQA also improve, and ImageNet-R remains nearly unchanged. This is the strongest end-to-end evidence for AttnCal's construction–calibration pipeline. The controls assign specific roles: Figure 3 tests the protected operator criterion, while Table 5 holds the constructed language state fixed and attributes a 0.40-point FAA gain to enabling the fitted final visual residual.

This distinction separates system performance from component attribution. The benchmark comparison asks whether the consolidated model retains useful capabilities across classification, question answering and captioning. The fixed-language control then asks whether visual calibration improves that same decoder, without changing its attention factors. Read together, they evaluate the complete construction and the final interface adjustment at their respective levels.

Table 2 reports 75.82 FAA, 84.95 CAA, and -0.57 BWT for AttnCal, compared with 73.01, 84.12, and -1.69 for local MAny. The simultaneous gains of 2.81 FAA, 0.83 CAA, and 1.12 BWT points show that the framework transfers beyond LLaVA while respecting InternVL's native attention and visual interface. What transfers is the construction and retained-state calibration procedure, not a shared set of fitted Q/K factors: each backbone receives its own native solution. The endpoint benefit is task-selective, led by a 15.70-point CLEVR-Math gain; the smaller IconQA and Flickr30k differences in the opposite direction explain why the aggregate claim is appropriately made at the

Table 2: **UCIT performance on InternVL-Chat-7B.** Rows use the stated InternVL 1.0 checkpoint, preprocessing, and 3,000 examples per task. IN-R, Arxiv, VizWiz, Icon, CLEVR, and Flickr denote ImageNet-R, ArxivQA, VizWiz-caption, IconQA, CLEVR-Math, and Flickr30k; Source cites the original method. Final average score (FAA) averages final scores, continual average score (CAA) averages stage-wise means, and backward transfer (BWT) measures later change on earlier tasks. Scores are percentages and BWT is in points; higher is better. Daggers mark the unmerged zero-shot model (†) and separate task experts (‡), for which CAA and BWT do not apply. Blanks are unavailable under this protocol.

Method	Source	IN-R	Arxiv	VizWiz	Icon	CLEVR	Flickr	FAA ↑	CAA ↑	BWT ↑
Zero-shot	(Guo et al., 2025c)	20.80	62.80	40.83	28.33	20.13	46.19	36.51†	—	—
Individual	(Guo et al., 2025c)	94.70	93.83	62.08	75.30	61.30	58.92	74.36‡	—	—
LoRA-FT	(Hu et al., 2021)	70.23	78.17	45.54	71.07	74.40	58.32	66.29	79.08	-9.45
O-LoRA	(Wang et al., 2023)	78.47	92.63	44.33	62.13	65.73	54.68	66.33	78.25	-4.56
MoELoRA	(Chen et al., 2024a)	54.97	84.47	46.47	63.13	61.83	55.98	61.14	73.68	-10.05
CL-MoE	(Huai et al., 2025)	73.17	78.27	44.57	72.00	71.03	58.02	66.18	78.33	-9.54
HiDE	(Guo et al., 2025a)	86.47	94.17	53.34	69.07	71.17	56.07	71.71	81.86	-0.08
MAny	(Gao et al., 2026)	88.07	95.13	62.44	73.47	60.20	58.77	73.01	84.12	-1.69
AttnCal	Ours	89.83	94.77	65.25	71.80	75.90	57.34	75.82	84.95	-0.57

benchmark level. Even with that heterogeneity, AttnCal leads reproduced merged baselines in both FAA and CAA for this protocol.

Table 3: **DCL performance on LLaVA-1.5-7B.** Rows use the complete native evaluation population. RS, Med, AD, Sci, and Fin denote remote sensing, medical, autonomous driving, science, and finance; Source cites the original method. Final average score (FAA) averages final scores, continual average score (CAA) averages stage-wise means, and backward transfer (BWT) measures later change on earlier tasks. Scores are percentages and BWT is in points; higher is better. Daggers mark the unmerged zero-shot model (†) and separate task experts (‡), for which CAA and BWT do not apply. Zero-shot and O-LoRA use the corrected native Fin endpoint; blanks are unavailable under this protocol.

Method	Source	RS	Med	AD	Sci	Fin	FAA ↑	CAA ↑	BWT ↑
Zero-shot	(Guo et al., 2025c)	32.27	28.26	15.61	35.41	37.69	29.85†	—	—
Individual	(Guo et al., 2025c)	68.67	62.07	53.69	53.20	81.80	63.89‡	—	—
LoRA-FT	(Hu et al., 2021)	77.03	44.39	35.29	48.37	80.81	57.18	64.92	-10.03
O-LoRA	(Wang et al., 2023)	79.41	39.58	30.53	40.40	87.14	55.41	61.48	-8.71
MoELoRA	(Chen et al., 2024a)	73.92	42.50	29.22	47.19	79.44	54.45	64.29	-14.28
CL-MoE	(Huai et al., 2025)	77.90	46.94	36.17	48.30	80.91	58.04	66.16	-10.22
HiDE	(Guo et al., 2025a)	79.92	52.38	36.79	48.25	74.44	58.36	64.72	-3.80
DISCO	(Guo et al., 2025b)	72.00	44.50	52.97	51.61	82.80	60.78	63.27	-3.75
MAny	(Gao et al., 2026)	59.44	60.91	53.81	51.79	89.57	63.10	63.41	-3.79
AttnCal	Ours	68.31	61.62	53.87	52.34	88.27	64.88	65.52	-1.68

In the native five-domain matrix, Table 3 gives AttnCal 64.88 FAA, 65.52 CAA, and -1.68 BWT. Relative to the strongest other merged method on each metric, FAA is 1.78 points higher and BWT is 2.07 points higher, while CAA remains within 0.64 points of the best comparator. AttnCal therefore provides the strongest final-domain aggregate and the least-negative transfer among reproduced merged baselines, yielding a more favorable endpoint–retention trade-off across the five domains. The BWT gain means earlier-domain scores decline less on average from their acquisition checkpoints as new experts are incorporated. Together with the FAA lead, this supports preservation of previously acquired capabilities in the final consolidated model.

Table 4: **DCL performance on InternVL-Chat-7B.** Rows use the complete native population and the stated InternVL 1.0 setup. RS, Med, AD, Sci, and Fin denote remote sensing, medical, autonomous driving, science, and finance; Source cites the original method. Final average score (FAA) averages final scores, continual average score (CAA) averages stage-wise means, and backward transfer (BWT) measures later change on earlier tasks. Scores are percentages and BWT is in points; higher is better. Daggers mark the unmerged zero-shot model (†) and separate task experts (‡), for which CAA and BWT do not apply; blanks are unavailable. AttnCal selects one visual-strength coefficient on all five tasks' source-validation data and fixes it across prefixes, so its prefix curves use global rather than strictly online hyperparameter selection.

Method	Source	RS	Med	AD	Sci	Fin	FAA ↑	CAA ↑	BWT ↑
Zero-shot	(Guo et al., 2025c)	32.05	35.90	14.39	34.90	42.83	32.01†	—	—
Individual	(Guo et al., 2025c)	73.50	66.64	54.25	53.52	82.49	66.08‡	—	—
LoRA-FT	(Hu et al., 2021)	68.85	46.86	33.72	42.26	86.65	55.67	63.66	-12.83
O-LoRA	(Wang et al., 2023)	76.91	28.38	28.95	31.99	67.10	46.67	55.05	-4.82
MoELoRA	(Chen et al., 2024a)	76.05	42.74	37.84	45.02	78.68	56.07	62.92	-8.96
CL-MoE	(Huai et al., 2025)	75.46	47.52	36.10	42.52	82.44	56.81	66.76	-13.07
HiDE	(Guo et al., 2025a)	80.36	59.23	51.20	23.91	53.97	53.73	64.34	-0.15
MAny	(Gao et al., 2026)	75.05	61.78	51.76	45.53	78.23	62.47	65.68	-0.65
AttnCal	Ours	75.81	62.39	50.53	49.37	79.37	63.49	66.10	-1.21

Table 4 reports 63.49 FAA, 66.10 CAA, and -1.21 BWT. Its 1.02-point FAA lead over the strongest other merged method extends AttnCal's endpoint advantage to a second backbone and domain sequence, while CAA remains within 0.66 points of the best comparator. Relative to the strongest FAA comparator, Science contributes a 3.84-point gain, whereas autonomous driving differs by -1.23 points, again locating the advantage in the aggregate rather than every domain. One globally selected visual-strength coefficient is fixed across all prefixes. Earlier prefixes are fitted with their available experts and source examples under this common deployed configuration; the shared coefficient is selected using the five-task source-validation pool.

The InternVL–DCL HiDE row uses a controlled inference variant: task-free anchor top-1 selection chooses the matching expert in every layer and that task's trained projector. This differs from the public inference code. Its FAA is 53.73 under this configuration.

4.5 Task profiles and visual-residual controls

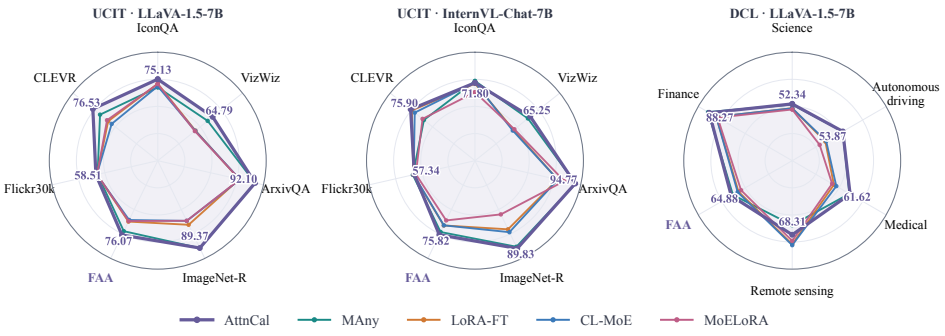


Figure 4: **Final-task profiles for AttnCal and four merged comparators.** The panels show six-task UCIT on LLaVA-1.5-7B, six-task UCIT on InternVL-Chat-7B, and five-domain DCL on LLaVA-1.5-7B—the three settings with AttnCal's largest final average score (FAA) leads. All radars use the same 0–100 scale; lines are local MAny, LoRA-FT, CL-MoE, and MoELoRA results, and labels show AttnCal's values. These endpoints expose task-level concentration, not stage-wise transfer or an individual component.

Figure 4 shows that AttnCal's final-average gains are concentrated rather than uniform: over MAny, UCIT–LLaVA gains are clearest on VizWiz-caption, IconQA, and CLEVR-Math while ImageNet-R and ArxivQA remain nearly unchanged; DCL–LLaVA combines a large remote-sensing gain with smaller changes elsewhere. The largest task gains recur in meaningful places across settings: CLEVR-Math improves by 8.46 points on LLaVA and 15.70 on InternVL, while DCL remote sensing improves by 8.87 points. Thus the FAA margins reflect task-specific capability preservation across backbones, not a uniform score shift; the shared scale makes that structure visible alongside the trajectory metrics.

Figure 5 then isolates the deployment strength of the final visual correction. Residual-off gives 75.68 FAA, scale 0.5 reaches 76.30, native scale gives 76.09, and scale 1.5 gives 74.68. This nonmonotone scan shows that the fitted direction is useful within a calibrated strength range rather than improving monotonically with magnitude. At native scale, gains led by ArxivQA and Flickr30k offset smaller task decreases, explaining how the residual raises the mean without acting as a uniform score offset. Table 5 reports the separate 76.07 main-table endpoint and fixed-W_E control.

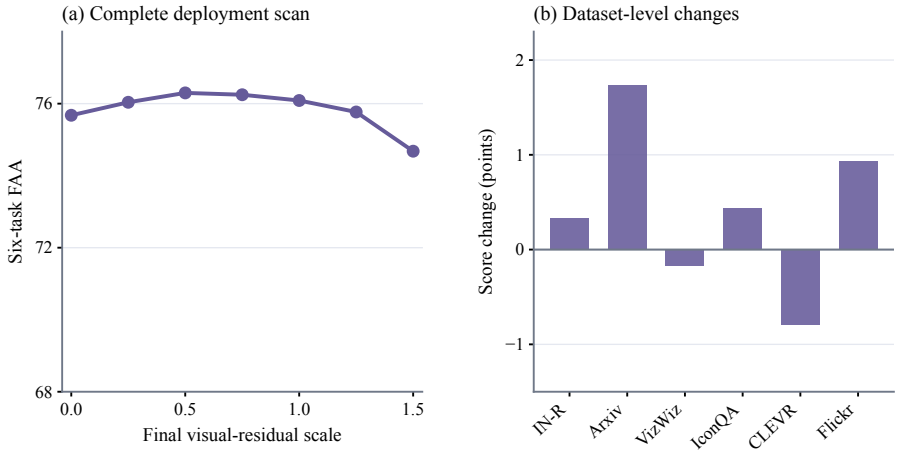


Figure 5: **AttnCal's visual correction has scale- and task-dependent effects.** One frozen AttnCal state is evaluated without retraining on six UCIT tasks with 3,000 examples each and the native LLaVA-1.5-7B scorer. Final average score (FAA) is the six-task mean. **(a)** Residual scales zero, 0.5, one (native), and 1.5 give 75.68, 76.30, 76.09, and 74.68 FAA; the separate main-table evaluation gives 76.07. **(b)** Bars show native-minus-off task changes. IN-R, VizWiz, CLEVR, and Flickr mean ImageNet-R, VizWiz-caption, CLEVR-Math, and Flickr30k. This is a within-model analysis.

Appendix B reports the fixed-W_E control and paired handoffs: the residual raises FAA from 75.68 to 76.07 (0.40 before rounding). All three visual-state handoffs retain strong final performance, with keeping the fitted state giving the highest mean across the repeats.

5 Conclusion

AttnCal casts expert attention consolidation as an explicit native-width Q/K operator problem and provides fixed-input bounds on the worst relative expert error. It calibrates the constructed Q/K factors under a fitted visual map, then freezes language weights for visual fitting. On full UCIT, the complete endpoint reaches 76.07 FAA versus 75.68 with its final residual disabled; across all four benchmark–backbone protocols, it has the highest FAA among reproduced merged baselines. Operator diagnostics identify which expert discrepancies the construction protects, fixed-state residual tests measure the effect of visual calibration, and complete trajectories establish the behavior of the continual model. Together, these complementary results connect the attention-level objective to deployment: native-width construction reconciles expert interactions in a shared attention mechanism, while visual calibration adapts the resulting language state to task inputs. This provides a principled route from attention consolidation to multimodal model merging while retaining the backbone's attention width and original visual interface.

REFERENCES

Joshua Ainslie, James Lee-Thorp, Michiel De Jong, Yury Zemlyanskiy, Federico Lebrón, and Sumit Sanghai. Gqa: Training generalized multi-query transformer models from multi-head checkpoints. In *EMNLP*, pp. 4895–4901, 2023.

Gino Brunner, Yang Liu, Damian Pascual, Oliver Richter, Massimiliano Ciaramita, and Roger Wattenhofer. On identifiability in transformers. *arXiv preprint arXiv:1908.04211*, 2019.

Pietro Buzzega, Matteo Boschini, Angelo Porrello, Davide Abati, and Simone Calderara. Dark experience for general continual learning: a strong, simple baseline. *NeurIPS*, 33:15920–15930, 2020.

Cheng Chen, Junchen Zhu, Xu Luo, Heng T Shen, Jingkuan Song, and Lianli Gao. Coin: A benchmark of continual instruction tuning for multimodel large language models. *NeurIPS*, 37:57817–57840, 2024a.

Hongxu Chen, Zhen Wang, Runshi Li, Bowei Zhu, and Long Chen. Iteris: Iterative inference-solving alignment for lora merging. In *CVPR*, pp. 4829–4838. IEEE, 2025a.

Jinpeng Chen, Runmin Cong, Yuzhi Zhao, Hongzheng Yang, Guangneng Hu, Horace Ho Shing Ip, and Sam Kwong. Sefe: Superficial and essential forgetting eliminator for multimodal continual instruction tuning. *arXiv preprint arXiv:2505.02486*, 2025b.

Weiyu Chen and James Kwok. Pareto merging: Multi-objective optimization for preference-aware model merging. *arXiv preprint arXiv:2408.12105*, 2024.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *CVPR*, pp. 24185–24198, 2024b.

Samuel N Cohen and Eliana Fausti. Hyperbolic contractivity and the hilbert metric on probability measures. *arXiv preprint arXiv:2309.02413*, 2023.

Bethan Evans, Benjamin Etheridge, Stephen Roberts, and Jared Tanner. Model merging by output-space projection. *arXiv preprint arXiv:2605.29101*, 2026.

Salman Faroz. Merotune: Rope-safe merging with a tunable dial. *arXiv preprint arXiv:2609.07971*, 2026.

Zijian Gao, Wangwang Jia, Xingxing Zhang, Pengfei Qian, Tao Sun, Bo Ding, Yong Dou, Huaimin Wang, and Kele Xu. Many: Merge anything for multimodal continual instruction tuning. *arXiv preprint arXiv:2604.14016*, 2026.

Haiyang Guo, Fanhu Zeng, Ziwei Xiang, Fei Zhu, Da-Han Wang, Xu-Yao Zhang, and Cheng-Lin Liu. Hide-llava: Hierarchical decoupling for continual instruction tuning of multimodal large language model. In *ACL*, pp. 13572–13586, 2025a.

Haiyang Guo, Fanhu Zeng, Fei Zhu, Wenzhuo Liu, Da-Han Wang, Jian Xu, Xu-Yao Zhang, and Cheng-Lin Liu. Federated continual instruction tuning. In *ICCV*, pp. 1325–1335. IEEE, 2025b.

Haiyang Guo, Fei Zhu, Hongbo Zhao, Fanhu Zeng, Wenzhuo Liu, Shijie Ma, Da-Han Wang, and Xu-Yao Zhang. Mcitlib: Multimodal continual instruction tuning library and benchmark. *arXiv preprint arXiv:2508.07307*, 2025c.

Danna Gurari, Yinan Zhao, Meng Zhang, and Nilavra Bhattacharya. Captioning images taken by people who are blind. In *ECCV*, pp. 417–434. Springer, 2020.

Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *ICCV*, pp. 8320–8329. IEEE, 2021.

Alex Henry, Prudhvi Raj Dachapally, Shubham Shantaram Pawar, and Yuxuan Chen. Query-key normalization for transformers. In *Findings of EMNLP 2020*, pp. 4246–4253, 2020.

Zhiyan Hou, Haiyun Guo, Haokai Ma, Yandu Sun, Yonghui Yang, and Jinqiao Wang. Pass-moe: Mitigating misaligned co-drift among router and experts via pathway activation subspaces for continual learning. In *ACL*, pp. 31959–31972, 2026.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *ICML*, pp. 2790–2799. PMLR, 2019.

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. *arXiv preprint arXiv:2106.09685*, 2021.

Tianyu Huai, Jie Zhou, Xingjiao Wu, Qin Chen, Qingchun Bai, Ze Zhou, and Liang He. Cl-moe: Enhancing multimodal large language model with dual momentum mixture-of-experts for continual visual question answering. In *CVPR*, pp. 19608–19617. IEEE, 2025.

Canbin Huang, Tianyuan Shi, Xiaojun Quan, Jingang Wang, Jianfei Zhang, and Qifan Wang. When model merging breaks routing: Training-free calibration for moe. *arXiv preprint arXiv:2606.03391*, 2026.

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Suchin Gururangan, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. Editing models with task arithmetic. *arXiv preprint arXiv:2212.04089*, 2022.

Hyundong Jin, Hyung Jin Chang, and Eunwoo Kim. Instruction-grounded visual projectors for continual learning of generative vision-language models. In *ICCV*, pp. 3466–3476. IEEE, 2025.

Xisen Jin, Xiang Ren, Daniel Preotiuc-Pietro, and Pengxiang Cheng. Dataless knowledge fusion by merging weights of language models. *arXiv preprint arXiv:2212.09849*, 2022.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526, 2017.

Lei Li, Yuqi Wang, Runxin Xu, Peiyi Wang, Xiachong Feng, Lingpeng Kong, and Qi Liu. Multimodal arxiv: A dataset for improving scientific comprehension of large vision-language models. In *ACL*, pp. 14369–14387, 2024.

Adam Dahlgren Lindström and Savitha Sam Abraham. Clevr-math: A dataset for compositional language, visual and mathematical reasoning. *arXiv preprint arXiv:2208.05358*, 2022.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *NeurIPS*, 36:34892–34916, 2023.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *CVPR*, pp. 26286–26296. IEEE, 2024.

David Lopez-Paz and Marc'Aurelio Ranzato. Gradient episodic memory for continual learning. *NeurIPS*, 30, 2017.

Pan Lu, Liang Qiu, Jiaqi Chen, Tony Xia, Yizhou Zhao, Wei Zhang, Zhou Yu, Xiaodan Liang, and Song-Chun Zhu. Iconqa: A new benchmark for abstract diagram understanding and visual language reasoning. *arXiv preprint arXiv:2110.13214*, 2021.

Michael S Matena and Colin Raffel. Merging models with fisher-weighted averaging. *NeurIPS*, 35: 17703–17716, 2022.

The-Hai Nguyen, Dang Huu-Tien, Takeshi Suzuki, and Le-Minh Nguyen. Regmean++: Enhancing effectiveness and generalization of regression mean for model merging. *arXiv preprint arXiv:2508.03121*, 2025a.

The-Hai Nguyen, Dang Huu-Tien, Takeshi Suzuki, and Le-Minh Nguyen. Regmean++: Enhancing effectiveness and generalization of regression mean for model merging. *arXiv preprint arXiv:2508.03121*, 2025b.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *ACL*, pp. 311–318, 2002.

Keonhee Park and Gunhee Kim. Sigma: Sign-guided merging and adaptation for multimodal continual instruction tuning. *arXiv preprint arXiv:2607.20511*, 2026.

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

Matthew Riemer, Ignacio Cases, Robert Ajemian, Miao Liu, Irina Rish, Yuhai Tu, and Gerald Tesauro. Learning to learn without forgetting by maximizing transfer and minimizing interference. *arXiv preprint arXiv:1810.11910*, 2018.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

Wenju Sun, Qingyong Li, Tiancheng Li, Yangliao Geng, and Boyang Li. Revisiting the role of pretrained weights in model merging: On near-optimality within the core subspace. In *ICML*.

Viet-Hoang Tran, Vinh Khanh Bui, Van-Hoan Trinh, Tan Lai Ngoc, and Tan M Nguyen. Functional equivalence in attention: A comprehensive study with applications to linear mode connectivity. *arXiv preprint arXiv:2606.17830*, 2026.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *ICCV*, pp. 4566–4575, 2015.

Liyuan Wang, Xingxing Zhang, Hang Su, and Jun Zhu. A comprehensive survey of continual learning: Theory, method and application. *TPAMI*, 46(8):5362–5383, 2024.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *NeurIPS*, 33: 5776–5788, 2020.

Wenhui Wang, Hangbo Bao, Shaohan Huang, Li Dong, and Furu Wei. Minilmv2: Multi-head self-attention relation distillation for compressing pretrained transformers. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pp. 2140–2151, 2021.

Xiao Wang, Tianze Chen, Qiming Ge, Han Xia, Rong Bao, Rui Zheng, Qi Zhang, Tao Gui, and Xuan-Jing Huang. Orthogonal subspace learning for language model continual learning. In *Findings of EMNLP 2023*, pp. 10658–10671, 2023.

Zifeng Wang, Zizhao Zhang, Chen-Yu Lee, Han Zhang, Ruoxi Sun, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer Dy, and Tomas Pfister. Learning to prompt for continual learning. In *CVPR*, pp. 139–149. IEEE, 2022.

Mitchell Wortsman, Gabriel Ilharco, Samir Ya Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, et al. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In *ICML*, pp. 23965–23998. Pmlr, 2022.

Jing Xu, Jiazheng Li, and Jingzhao Zhang. Scalable model merging with progressive layer-wise distillation. *arXiv preprint arXiv:2502.12706*, 2025.

Prateek Yadav, Derek Tam, Leshem Choshen, Colin A Raffel, and Mohit Bansal. Ties-merging: Resolving interference when merging models. *NeurIPS*, 36:7093–7115, 2023.

Enneng Yang, Zhenyi Wang, Li Shen, Shiwei Liu, Guibing Guo, Xingwei Wang, and Dacheng Tao. Adamerging: Adaptive model merging for multi-task learning. In *ICLR*, volume 2024, pp. 22743–22763, 2024.

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

Dengming Zhang, Xiaowen Ma, Zhenliang Ni, Zhenkai Wu, Han Shu, Xin Jiang, and Xinghao Chen. Expert merging: Model merging with unsupervised expert alignment and importance-guided layer chunking. In *ICLR*, volume 2026, pp. 21780–21803, 2026a.

Duzhen Zhang, Yahan Yu, Qiaoyi Su, Jiahua Dong, and Tielin Zhang. Progressive multimodal alignment for continual instruction tuning. *arXiv preprint arXiv:2607.26947*, 2026b.

Adam Zweiger, Xinghong Fu, Han Guo, and Yoon Kim. Fast kv compaction via attention matching. *arXiv preprint arXiv:2602.16284*, 2026.

AI USE STATEMENT

Generative AI tools assisted with manuscript language polishing and experimental code development and debugging. The authors take responsibility for the final manuscript and code.

The appendix develops the construction and tests the individual choices behind the complete pipeline. Appendix A specifies the native Q/K solve and both calibration stages, with parameter and data settings in Appendix C. Appendix B separates the deployed residual's effect from the choice of visual-state initialization. Appendices F and G give the operator bounds and architectural conditions. These analyses explain different parts of the method; complete-model performance remains the question answered by the main benchmark tables.

A DETAILED CONSTRUCTION AND CALIBRATION

Figure 2 traces one shared visual state through native attention construction, Q/K task calibration, and final visual adaptation. The sequence has three transitions: construct a native attention initialization, calibrate it under a fitted visual map, and adapt that map with the language model frozen. The inputs are a pretrained backbone, compatible task experts, the retained visual interface ϕ_0, and source calibration data $\mathcal{D}_{\mathrm{fit}}$.

Stage 1 first fixes the declared reference weights outside the query–key (Q/K) projections and constructs native-width Q/K factors that protect the largest expert-relative operator error within each attention head. It then holds the resulting intermediate language state W_Q fixed while an auxiliary visual fit starts from zero and learns the initial visual coefficients A_{init} through answer cross-entropy (CE) and, where specified, a reference Kullback–Leibler (KL) term. With that visual condition fixed, the final part of Stage 1 fits only the Q/K residual parameters θ_E using answer CE, producing the calibrated language model W_E. Stage 2 freezes that model and continues the same visual coefficients from A_{init} to their final state A_E under the visual-fitting objective in Equation 3. The visual basis and preceding interface ϕ_0, including any retained expert-projector selection, remain unchanged in matched controls. The fitted visual state is carried forward, with fresh optimizer moments at the stage boundary. This sequence gives language calibration a concrete visual condition and carries that condition into the final model.

For the selected stage-wise ablation, Module 1 is the complete language construction: it combines fixed non-Q/K projections, native Q/K construction, and visual-conditioned Q/K task calibration. The ablation holds the resulting W_E fixed and tests the effect of enabling the final visual correction at deployment. In Table 5, "final residual off" describes only the evaluation endpoint: the auxiliary visual map was still used during Module 1. Construction histories and total costs therefore remain distinct even when the final visual-fitting schedules are matched.

A.1 STAGE 1: SOLVING THE MERGED LANGUAGE MODEL

The joint Q/K solve is the construction component developed here. The remaining projections are assembled by task-balanced, second-moment-weighted response reconstruction and held fixed in the matched Q/K comparisons. For InternVL, prototype-conditioned weights combine frozen projector outputs (Gao et al., 2026). AttnCal constructs the native Q/K factors, calibrates them under a fitted visual map, and fits a shared residual after this interface. These operations define the model directly from its task experts.

Representable Q/K targets. The inputs are task-adapted query and key projections that share a backbone, head ordering, and rotary-frequency convention. For a head, expert i, and rotary frequency f, pair the corresponding real rows into complex vectors $q_{if} = q_{if}^{(1)} + \mathrm{i}q_{if}^{(2)}$ and $k_{if} = k_{if}^{(1)} + \mathrm{i}k_{if}^{(2)}$ in $\mathbb{C}^d$, where the superscripts identify the two real rows in a rotary pair and $*$ denotes conjugate transpose. For the native split-half convention with head width d_h, these rows are f and $f + d_h/2$, not adjacent rows. Weight export concatenates real and imaginary parts in that same order. Thus the complex construction preserves the backbone's rotary layout rather than introducing a new pairing. With rotation $e^{\mathrm{i}\omega_f m}$ at position m, set $\Delta = m_{\mathrm{query}} - m_{\mathrm{key}}$. The bilinear query–key operator is $M_{if} = q_{if}k_{if}^{*}$. It acts on hidden-state pairs; it is not a token–token attention

matrix or a softmax attention distribution. For real hidden states x, y and relative position Δ, its contribution to the unscaled attention logit is $\mathrm{Re}(e^{\mathrm{i}\omega_f \Delta} x^T M_{if} y)$, using the complex representation of RoPE (Su et al., 2024). This target is invariant to a common phase change of q_{if} and k_{if}.

This representation makes native realizability explicit: each row pair of the original architecture represents a matrix of rank at most one. An unconstrained average of expert operators can exceed that rank, while separate averages of q_{if} and k_{if} introduce cross-expert products. We therefore optimize the shared operator within the native rank constraint and recover Q/K factors from the resulting approximation.

Head-coupled minimax reconstruction. Let $S_f = q_f k_f^*$ denote a candidate operator in the native rank-one family, and let $N_i = \sum_f \|M_{if}\|_F^2 > 0$ measure expert i's operator energy within that head. The relative reconstruction error $L_i(S)$ and its best worst-expert value P_H are

$$L_i(S) = \frac{\sum_f \|S_f - M_{if}\|_F^2}{N_i}, \qquad P_H = \min_{\mathrm{rank}(S_f)\le 1\ \forall f} \max_i L_i(S).$$

The normalization expresses reconstruction error relative to each expert's head energy. Protecting the worst relative error prevents a favorable average from concealing a poorly reconstructed expert. This is a robustness criterion for the native attention operator. Summing across frequencies before taking the maximum permits the construction to distribute error within a head while protecting that head's worst expert-relative error. Heads are solved independently, so the objective does not allocate a shared error budget across heads. Frequency-wise protection imposes a different allocation constraint; it is a matched objective ablation. Zero-energy heads require a separately declared convention and are outside the present $N_i > 0$ construction.

This objective has an explicit fixed-input interpretation. For the same hidden states x, y, F rotary frequencies and head dimension d_h, the reconstructed and expert attention logits satisfy

$$|\widehat{s} - s_i| \le \frac{\|x\|_2 \|y\|_2}{\sqrt{d_h}} \sqrt{F N_i L_i(S)}.$$

The bound holds uniformly over relative position. It connects operator reconstruction to local attention-logit control; propagation through later hidden states and generated answers is assessed empirically.

Low-rank factorization and primal–dual bounds. Thin QR factorizations restrict computation to the small joint spans of the expert factors, avoiding dense $d \times d$ operator materialization. For simplex expert weights λ, let a be the normalized weight sum, B_f the weighted target at frequency f, and $g_H(\lambda)$ a dual lower bound:

$$a = \sum_i \lambda_i / N_i, \qquad B_f = \sum_i \lambda_i M_{if} / N_i, \qquad g_H(\lambda) = 1 - \frac{\sum_f \sigma_1(B_f)^2}{a}.$$

Here a collects the normalized expert weights and B_f is their frequency-specific weighted operator. For a fixed simplex vector, the best rank-one approximation of each B_f supplies a native-factor candidate. Completing the square gives $g_H(\lambda) \le P_H$. Thus a feasible candidate $\widehat{S}$ and any simplex point bracket the optimum:

$$g_H(\lambda) \le P_H \le \max_i L_i(\widehat{S}).$$

This bracket also has a diagnostic decomposition. For the same simplex weights and feasible candidate, distinguish an inner-solve gap $\epsilon_{\mathrm{inner}}$ from an expert-balance gap $\epsilon_{\mathrm{balance}}$:

$$\epsilon_{\mathrm{inner}} = \sum_i \lambda_i L_i(\widehat{S}) - g_H(\lambda),$$

$$\epsilon_{\mathrm{balance}} = \max_i L_i(\widehat{S}) - \sum_i \lambda_i L_i(\widehat{S}).$$

Both are nonnegative and their sum is the displayed upper–lower gap, so

$$0 \le \max_i L_i(\widehat{S}) - P_H \le \epsilon_{\mathrm{inner}} + \epsilon_{\mathrm{balance}}. \tag{1}$$

The first term measures reconstruction at the chosen expert weights; the second measures whether the candidate still favors some experts over the worst-served expert. The bracket holds in exact arithmetic. Its proof and a singular-value-tie counterexample are in Appendix F. A finite pair-exchange solver uses one simplex vector shared across frequencies to construct candidates. A global dual maximizer with separated leading singular values yields an exact saddle solution under the appendix's conditions; finite iterations are assessed through their primal and dual values. Factoring each candidate operator recovers query and key rows at the original width, producing one static Q/K replacement with no added router. The exact-arithmetic bounds and numerical residual checks are reported separately.

Appendix G gives the corresponding feasibility conditions for nonrotary and shared-key attention.

Q/K controls hold the visual configuration and other weights fixed to isolate attention construction. The complete method additionally includes the following task calibration and visual continuation, which are evaluated together.

Visual-conditioned language calibration. The selected construction combines the expert-derived Q/K solution with task calibration of its query and key projections. Let W_Q be the language model assembled by the preceding Q/K solve and the declared reference rule for all remaining weights. With the previously fitted visual map A_{init} held fixed, this calibration updates only the Q/K residual parameters θ:

$$\theta_E \leftarrow \mathrm{Fit}_\theta \ \frac{1}{|\mathcal{D}_{\text{fit}}|} \sum_{(x,y)\in\mathcal{D}_{\text{fit}}} \ell_{\text{CE}}\big(f_{W_Q+\Delta_{QK}(\theta)}(x;\phi_0,A_{\text{init}}),y\big)\,, \qquad W_E = W_Q + \Delta_{QK}(\theta_E). \tag{2}$$

Here Fit denotes the finite optimization schedule used for calibration. The visual map, all non-Q/K language weights, the backbone, and the preceding visual interface remain frozen during this step. The retained visual state is then passed to Stage 2. Thus language calibration and visual continuation share the fitted visual state at their boundary. Figure 1(c) tests the final correction within this constructed language state. Disabling that correction at evaluation leaves the visual conditioning used during language calibration unchanged.

The operator solve supplies a native-feasible, expert-balanced initialization; task calibration adds a rank-8 Q/K correction. The objective bounds characterize the constructed initialization, and task scores evaluate the calibrated model.

A.2 Stage 2: shared visual residual calibration

Implemented calibration. For visual features $h \in \mathbb{R}^{n_v\times d_v}$ produced by the fixed visual interface, let $B \in \mathbb{R}^{d_v\times r_v}$ denote the fixed, source-derived orthonormal basis and $A \in \mathbb{R}^{d_v\times r_v}$ the fitted coefficients; h' denotes the calibrated features. The answer-token CE loss ℓ_{CE} trains the visual coefficients through the frozen language predictor f_{W_E}. The reference KL divergence D_{KL} compares answer distributions with and without the residual and is weighted by ω_{KL}. The executed shared calibration is

$$\begin{aligned} h' &= h + (hB)A^\top, \\ A_E \leftarrow \mathrm{Fit}_{A;A_{\text{init}}} \ &\frac{1}{|\mathcal{D}_{\text{fit}}|} \sum_{(x,y)\in\mathcal{D}_{\text{fit}}} \Big[\ell_{\text{CE}}(f_{W_E}(x;h'),y) \\ &\quad + \omega_{\text{KL}} D_{\text{KL}}(p_{W_E}(\cdot\mid x;h)\,\|\,p_{W_E}(\cdot\mid x;h'))\Big]. \end{aligned} \tag{3}$$

When included, both loss terms are averaged over answer-token positions in each source example. The KL teacher uses the same frozen language model and unmodified routed visual features h at temperature one, with detached logits. The auxiliary fit starts at zero, whereas the final fit starts at A_{init}. Initialization and calibration settings are given in Appendix C. The frozen LLaVA recipes and the UCIT–InternVL port use $\omega_{\text{KL}} = 0$ (KL is diagnostic where computed); the selected DCL–InternVL recipe uses $\omega_{\text{KL}} = 1$. The shared linear residual keeps the adaptation space fixed across comparisons. Its compact form makes the retained visual state, rather than a change in adapter expressivity, the object of the stage comparison. The fitting operator uses finite optimization initialized at A_{init}. The language weights and the preceding visual interface are frozen. The same map is used

across tasks; no new inference router is trained. Current matched experiments retain the existing expert-selection procedure; the shared residual adapts its output rather than replacing the interface.

What retaining a visual state preserves. Let $R(\theta, a)$ denote answer CE on the same weighted source set, where θ are Q/K parameters and a are coefficients of a *fixed* visual basis and preceding interface. Let θ_0 be the constructed Q/K state, θ_1 the task-calibrated state, and $a_0 = A_{\text{init}}$. Define the language improvement under visual state a as $G(a) = R(\theta_0, a) - R(\theta_1, a)$ and the visual–language interaction as

$$I = R(\theta_1, a_0) - R(\theta_0, a_0) - R(\theta_1, 0) + R(\theta_0, 0).$$

Then $G(0) = G(a_0) + I$ exactly. Thus the language gain measured under the fitted visual condition need not be the gain after resetting that condition. AttnCal preserves the map at the stage boundary by passing A_{init}, the same basis and interface to the final fit while resetting optimizer moments. Appendix F gives a conditional sensitivity bound under common answer CE; the handoff experiments evaluate the final fitted states.

Deployment and architecture scope. AttnCal combines an explicitly constrained attention construction with source-supervised calibration. The operator analysis describes the initial Q/K approximation, and matched evaluation assesses the resulting construction–calibration sequence. The output has one fixed language model and one shared residual calibration map after the retained visual interface, with no test-time teacher query or optimization. Extending construction to visual attention requires the visual architecture's own positional and shared-key constraints (Appendix G).

Module interface and controls. Module 1 constructs the complete language state W_E from the fixed non-Q/K reference weights and the proposed Q/K solve and calibration. The first scored endpoint evaluates this state with the common reference interface ϕ_0 and its final shared residual disabled. Module 2 continues A_{init} under the same frozen W_E; the complete method evaluates the resulting visual state. The two endpoints keep the same source roles, architecture and scorer, and construction costs include the auxiliary visual fit used during language calibration.

B Supporting Diagnostics and Controls

Table 5 compares two configurations evaluated on UCIT with 3,000 examples per task. They share the pretrained backbone, expert collection, calibration roles, reference connector ϕ_0, and the same completed language-construction module W_E. That module combines fixed non-Q/K projections with AttnCal's native Q/K construction and visual-conditioned calibration. The comparison changes only whether the final fitted shared visual residual is enabled at deployment. The original interface's expert-selection mechanism remains active in both rows, isolating the final residual within this fixed language state and interface.

Table 5: **The final visual correction improves AttnCal's task-calibrated language model.** Both UCIT rows use the same constructed and calibrated query–key weights, the same remaining language weights, and 3,000 evaluation examples per task. The table calls this fixed task-calibrated language state W_E. The first row switches the shared visual correction off only for evaluation, after construction has already used an auxiliary visual fit. The second row uses the final fitted correction initialized from that retained fit. Final average score (FAA) averages the six final task scores, and FAA change is the difference between the two rows under the same native scorer.

Configuration	Final visual residual	FAA	FAA change
W_E construction, residual off	Off	75.68	
+ Visual adaptation (complete)	On	76.07	+0.40

The complete model exceeds the same constructed language model with its final residual disabled by 0.40 FAA points (76.07 versus 75.68). With the task-calibrated language state W_E fixed, this comparison measures the final residual's deployment benefit. Both rows already include Q/K construction and auxiliary visual fitting, so continuation from that fitted initialization is examined separately in the paired handoff study below.

These controls hold the backbone, expert collection, source roles, retained interface and scorer fixed, and include the auxiliary visual fit in construction cost. Across backbones, the complete pipeline respects each model's native attention and visual-interface constraints.

Across three paired UCIT–LLaVA repeats, keeping the fitted visual state without further updates, continuing from it, and training the residual coefficients from zero all achieve state-of-the-art final average score (FAA) relative to the reproduced merged baselines in Table 1. This consistently strong performance across initialization and continuation choices indicates that AttnCal is robust to its final visual-state handoff. Appendix C.1 reports all nine endpoints and their means. Keeping the fitted state without continuation has the highest mean, followed by resetting and continuing from the retained state. The evidence for robustness is the strong performance of all three choices; continued updates offer no further gain over keeping the fitted state in this comparison.

C Evaluation and Numerical Protocol

Selected UCIT–LLaVA construction and fitting. The language reference uses task-equal covariance-weighted reconstruction. AttnCal keeps its 160 non-Q/K projections fixed and constructs the 64 query and key projections across 32 layers.

Calibration uses AdamW with batch size one and 62 source examples per task. The auxiliary visual fit, Q/K task calibration and final visual continuation each use the same 372-update budget. Q/K corrections have rank eight, with non-Q/K and visual parameters frozen during their fit. The visual correction has one fixed shared rank-eight basis and one trainable coefficient matrix of shape 4096×8, rather than separate corrections for the six projectors. All three fits optimize answer-token cross-entropy. Reference-logit KL is diagnostic, not a training loss in this setting. Final continuation retains the auxiliary visual state and basis with fresh optimizer moments.

Source-derived visual basis. Visual-basis extraction converts source images to RGB and uses the native LLaVA processor and frozen visual routing path. It collects the resulting projector-output tokens before the language decoder. For each example, it averages the tokens and their outer products separately; it then averages these first and second moments within each task and equally across the six tasks. Subtracting the outer product of the pooled mean from the pooled second moment gives the centered covariance. The covariance is symmetrized, and an FP32 CUDA eigendecomposition supplies the eight largest-eigenvalue directions in descending order as the fixed basis. This weighting gives every example equal mass within its task rather than giving longer token sequences more weight.

Evaluation population and scoring. The full evaluation uses 3,000 examples per task across ImageNet-R, ArxivQA, VizWiz-caption, IconQA, CLEVR-Math, and Flickr30k. The reported FAA is the equal-task mean of the six frozen task scores. Caption scores use the official seven-metric corpus aggregation; they are not averages of per-example correctness. ArxivQA uses uppercase whole-string comparison, without whitespace stripping or option-letter extraction.

Numerical conditions for controlled comparisons. Q/K construction uses FP32, including complex arithmetic with FP32 real and imaginary components; published weights and inference use FP16. Component comparisons keep the FP64-constructed non-Q/K weights fixed. These experiments therefore compare Q/K choices under a common reference, rather than changing the construction precision of the entire model.

Data roles. The development pool is sampled from UCIT training files and is training-exposed for baselines fitted on those complete files. A path-and-question audit finds no identical image–question pair between those files and the six evaluation sets, although 1,280 of 3,000 ArxivQA questions share an image path with training. This establishes question-level separation, with image-level and semantic deduplication outside the audit. Development and full-test scores use separate populations; full-test outcomes do not retroactively select development configurations.

Visual-strength sensitivity. We select the release strength of the Stage 2 visual correction without using the DCL test population. Holding the experts, Q/K operator construction, conditioned Q/K residual, and all evaluation code fixed, we interpolate the visual coefficients from their retained

Table 6: **Source-validation sensitivity to visual-correction strength.** The first column scales the Stage 2 visual correction by ρ. RS, Med, AD, Sci, and Fin mean remote sensing, medical, autonomous driving, science, and finance. Entries are cross-entropy multiplied by 1,000 for display, and Macro is their five-task mean. Lower is better. Every row uses the same 120 held-out source examples, 24 per task, without formal test scores.

ρ	RS	Med	AD	Sci	Fin	Macro
1.0	50.64	87.36	91.55	98.12	23.22	70.18
1.5	50.57	87.15	91.59	98.06	23.21	70.12
2.0	50.42	87.10	91.54	98.10	23.15	70.06
2.5	50.28	86.80	91.50	98.02	23.14	69.95
3.0	50.23	86.62	91.53	98.04	23.16	69.91
3.5	50.18	86.46	91.62	97.91	23.12	69.86
4.0	50.10	86.21	91.49	97.86	23.09	69.75
5.0	**49.94**	**85.98**	91.50	**97.81**	23.11	**69.67**

initialization A_{init} to their fitted state A_E as $A(\rho) = A_{\text{init}} + \rho(A_E - A_{\text{init}})$. Each candidate is evaluated on the same 24 held-out source examples per task, disjoint from the 62 examples used for fitting. Selection uses mean answer-token cross-entropy on this source-validation set. Table 6 reports source-validation losses; only the selected setting is evaluated on the final test row. For DCL–InternVL, the coefficient is selected once from the five-task source-validation pool and fixed at five across prefixes. Earlier prefixes fit only their available experts and source examples, but inherit that global selection. The matrix therefore evaluates a common deployed configuration rather than strictly online hyperparameter selection.

DCL–InternVL retained-state fitting and initialization. The selected final visual fit performs 372 source-training updates with the language weights frozen. It retains the fitted visual map and fixed basis, using fresh optimizer moments. The visual objective is per-example answer-token cross-entropy plus a unit-weight, temperature-one $D_{\text{KL}}(p_{\text{ref}} \| p_{\text{student}})$ term, where p_{ref} comes from the same frozen language model on unchanged routed visual features and is detached from the gradient. This KL concerns answer logits, not attention maps. The continuation budget is only one part of total construction cost, which also includes Q/K construction and the auxiliary visual and Q/K fits.

C.1 Final visual-residual handoff sensitivity

We compare the visual-residual handoff within three paired source-order repeats, fixing each repeat's constructed language state, visual basis, expert bank, and original visual interface. *Keep* evaluates the fitted auxiliary visual state without a final continuation. *Retain* starts the final 372-update fit from that state with fresh optimizer moments. *Reset* performs the same 372 updates with only the residual coefficients initialized at zero; the visual tower, expert selection, basis, and language weights are unchanged. Every endpoint has an independently scored native six-task evaluation.

Table 7: **Final visual-residual handoff on UCIT–LLaVA-1.5-7B.** R0–R2 are three paired source-order repeats sharing the language state, visual basis, experts, data order, and native scorer within each repeat. Keep evaluates the auxiliary visual state without further updates. Retain continues from that state for 372 updates with a fresh optimizer; Reset uses the same 372-update schedule but initializes only the residual coefficients at zero. Each endpoint is scored on 3,000 examples from each of six UCIT tasks. Final average score (FAA) is their equal-task mean; higher is better. The strongest reproduced merged baseline in the same setting has 72.15 FAA; the separately evaluated Individual experts are not a merged baseline. Means average the three repeats, and displayed values are rounded to two decimal places. The Keep–Reset ordering changes between repeats.

Final visual state	R0 FAA	R1 FAA	R2 FAA	Mean FAA
Keep	76.17	76.16	76.10	**76.14**
Retain	75.96	76.01	75.09	75.68
Reset	76.18	76.01	75.34	75.84

All nine endpoints in Table 7 exceed the strongest reproduced merged baseline. Keep exceeds Retain in every repeat, while Reset slightly exceeds Keep in R0 and is lower in R1 and R2. The ordering therefore varies, but performance remains strong across all three final visual-state choices.

D What Factor Averaging Preserves

At a fixed rotary frequency, let the expert query and key factors be $q_i, k_i \in \mathbb{C}^d$ for $i = 1, \ldots, m$. Their arithmetic means are $\bar{q}$ and $\bar{k}$, and their deviations from those means are $u_i = q_i - \bar{q}$ and $v_i = k_i - \bar{k}$. The operator difference obeys

$$\frac{1}{m}\sum_i q_i k_i^* - \bar{q}\bar{k}^* = \frac{1}{m}\sum_i u_i v_i^*.$$

To see this, expand each product as $\bar{q}\bar{k}^* + \bar{q}v_i^* + u_i\bar{k}^* + u_i v_i^*$. The two cross terms vanish after averaging because the deviations sum to zero. Hence factor averaging matches the operator mean exactly when this cross-expert covariance vanishes. If both sets of deviations are scaled by a common factor, the discrepancy scales quadratically. This identity explains the preservation target: separate factor averaging produces a native rank-at-most-one operator but discards a term retained by operator averaging. AttnCal instead balances discrepancies from the expert operators within that native family.

E A Representability Example

The following rank-one example shows why the mean operator can exceed native width. Let e_1, e_2 be orthonormal real vectors embedded in the shared complex coordinate space. The two valid expert operators are $M_1 = e_1 e_1^*$ and $M_2 = e_2 e_2^*$, each of rank one. Their equal-weight mean is

$$\overline{M} = \tfrac{1}{2}(e_1 e_1^* + e_2 e_2^*), \qquad \operatorname{rank}(\overline{M}) = 2.$$

Every single original-width row pair has form qk^* and rank at most one. The squared Frobenius residual of its best approximation to $\overline{M}$ is $1/4$, obtained by retaining either of the two singular components. Thus, exact mean-operator preservation can be infeasible in this operator class. Operator residuals in trained heads measure the extent of this representability constraint in the constructed model.

F Theoretical Development

Compact coordinates preserve the stated operator problem. At one frequency, let U_f and V_f have orthonormal complex columns that span the left and right factors of all expert operators M_{if}. The map $\mathcal{P}_f(S) = U_f U_f^* S V_f V_f^*$ cannot increase rank and fixes every M_{if}. As an orthogonal projection under the real Frobenius inner product, it satisfies

$$\|S_f - M_{if}\|_F^2 = \|\mathcal{P}_f(S_f) - M_{if}\|_F^2 + \|S_f - \mathcal{P}_f(S_f)\|_F^2.$$

Therefore, restricting each candidate to the joint factor spans loses no optimum of this native-rank Frobenius objective.

Head-coupled construction. For $N_i = \sum_f \|M_{if}\|_F^2 > 0$, write $L_i(S) = \sum_f \|S_f - M_{if}\|_F^2 / N_i$. For simplex weights λ, define

$$a = \sum_i \lambda_i / N_i, \quad B_f = \sum_i \lambda_i M_{if}/N_i, \quad g_H(\lambda) = 1 - \frac{\sum_f \sigma_1(B_f)^2}{a}.$$

Completing the square separately at each frequency gives $g_H(\lambda) = \min_{\operatorname{rank}(S_f)\le 1\ \forall f} \sum_i \lambda_i L_i(S)$, a lower bound on the head minimax objective. The weighted objective is uniformly coercive on the simplex because $a \ge \min_i 1/N_i > 0$.

Proposition 1 (Head minimax bracket and sufficient saddle condition)**.** *For expert i and rotary frequency f, let M_{if} be its target query–key operator and $N_i = \sum_f \|M_{if}\|_F^2 > 0$ its head energy. The normalized expert loss is $L_i(S) = \sum_f \|S_f - M_{if}\|_F^2/N_i$, and the head minimax value P_H minimizes $\max_i L_i(S)$ over native-feasible rank-at-most-one operators S_f. For nonnegative simplex weights λ_i summing to one, define the weighted operator $B_f(\lambda) = \sum_i \lambda_i M_{if}/N_i$ and let $g_H(\lambda)$ be the minimum weighted loss $\min_{\mathrm{rank}(S_f)\le 1} \sum_i \lambda_i L_i(S)$. For every simplex λ and native-feasible $\widehat{S}$, $g_H(\lambda) \le P_H \le \max_i L_i(\widehat{S})$. If λ^* globally maximizes g_H and its weighted native-feasible minimizer S^* is unique, then $\max_i L_i(S^*) = g_H(\lambda^*) = P_H$. A sufficient uniqueness condition is that at each frequency either $B_f(\lambda^*) = 0$ or its largest singular value is separated from the second.*

Proof. The first inequality follows since a simplex-weighted loss never exceeds the worst expert loss; the second follows from feasibility. The weighted minimizers remain bounded near λ^* by coercivity. Under uniqueness, the envelope directional derivative toward simplex vertex e_i is $L_i(S^*) - \sum_j \lambda_j^* L_j(S^*)$ and is nonpositive at the dual maximum. Thus every expert loss at S^* is at most their weighted mean $g_H(\lambda^*)$; their maximum is also at least that mean. Weak duality gives equality with P_H. A separated top singular value yields a unique best rank-one matrix; when $B_f = 0$, the unique weighted minimizer is zero. Uniqueness here concerns the operator matrix, not its Q/K factorization. □

For any feasible $\widehat{S}$, the nonnegative inner and balance terms in Equation 1 telescope to $\max_i L_i(\widehat{S}) - g_H(\lambda)$. Weak duality therefore bounds the candidate's objective excess above the unknown P_H by their sum, without requiring a global dual maximizer from the finite solver. At a singular-value tie, arbitrary factor recovery can still be poor: with one frequency, $M_1 = e_1 e_1^*$ and $M_2 = e_2 e_2^*$, equal simplex weights give $g_H = 3/4$. Choosing the tied matrix $\widehat{S} = \mathrm{diag}(1/2, 0)$ yields worst loss $5/4$; the feasible rank-one matrix $\frac{1}{4}\mathbf{1}\mathbf{1}^\top$ gives both losses $3/4$. The former has zero inner error but balance error $1/2$. This is a primal-recovery choice, not a positive duality gap. The bracket is exact-arithmetic; floating-point values need explicit error bounds before they can be called certified enclosures.

When all target frequencies have positive energy, uniformly bounding every frequency-relative squared error by ϵ implies $L_i \le \epsilon$ by weighted averaging, but the converse fails. For scalar target frequencies with squared norms 1 and 100, deleting the first and preserving the second gives $L_i = 1/101$ and first-frequency error 1. Thus head protection permits error to be distributed across frequencies according to their contribution to total head energy. For identical hidden inputs and F frequencies, the fixed-input argument below and Cauchy–Schwarz give

$$|\widehat{s} - s_i| \le \frac{\|x\|_2 \|y\|_2}{\sqrt{d_h}} \sqrt{F N_i L_i(S)}.$$

The bound controls a single attention logit at fixed hidden inputs.

Frequency-wise special case and saddle argument. We state the exact optimization object separately from finite-precision solver diagnostics. Let nonzero expert operators be $M_1, \dots, M_m$, set $n_i = \|M_i\|_F^2$, and write

$$\ell_i(S) = \frac{\|S - M_i\|_F^2}{n_i}, \qquad P = \min_{\mathrm{rank}(S)\le 1} \max_i \ell_i(S).$$

Complex matrices use the real Frobenius inner product. For simplex weights λ, define

$$a = \sum_i \lambda_i / n_i, \quad B = \sum_i \lambda_i M_i / n_i, \quad g(\lambda) = 1 - \sigma_1(B)^2/a.$$

Completing the square shows that g is the minimum weighted loss over rank-at-most-one matrices. Thus $\max_{\lambda \in \Delta_m} g(\lambda) \le P$ even without strong duality. If $B = u\Sigma v^*$ has a separated leading singular value, the unique minimizing matrix is $S_\lambda = \sigma_1(B) u_1 v_1^*/a$.

Proposition 2 (A sufficient condition for an exact saddle solution)**.** *For nonzero expert query–key operators M_i at one rotary frequency, let $n_i = \|M_i\|_F^2$ be their energies and $\ell_i(S) = \|S - M_i\|_F^2/n_i$ their normalized reconstruction losses. Write $P = \min_{\mathrm{rank}(S)\le 1} \max_i \ell_i(S)$ for the worst-expert*

optimum. For simplex weights λ_i, define the weighted operator $B(\lambda) = \sum_i \lambda_i M_i/n_i$, the scalar $a(\lambda) = \sum_i \lambda_i/n_i$, and the weighted rank-one optimum $g(\lambda) = 1-\sigma_1(B(\lambda))^2/a(\lambda)$. When the leading singular value is separated, write u_1, v_1 for its singular vectors and $S_\lambda = \sigma_1(B(\lambda))u_1v_1^*/a(\lambda)$ for the unique weighted rank-one minimizer. Suppose λ^* globally maximizes g on the simplex and $\sigma_1(B(\lambda^*)) > \sigma_2(B(\lambda^*))$. Then S_{λ^*} solves the worst-expert primal problem and $\max_i \ell_i(S_{\lambda^*}) = g(\lambda^*) = P$.

Proof. Uniqueness of the weighted minimizer permits the simplex-direction envelope derivative $Dg(\lambda^*; e_i - \lambda^*) = \ell_i(S_{\lambda^*}) - \sum_j \lambda_j^* \ell_j(S_{\lambda^*})$. At a simplex maximum this feasible directional derivative is nonpositive. Consequently $\ell_i(S_{\lambda^*}) \le \sum_j \lambda_j^* \ell_j(S_{\lambda^*}) = g(\lambda^*)$ for every i. The maximum loss is at least its weighted mean, so equality holds. Together with weak duality, this proves primal optimality. □

A checkable sufficient spectral condition is $\|M_i - M_0\|_2 \le \delta$ for a common rank-one M_0 with $2\delta < \sigma_1(M_0)$. Indeed, B/a is a convex combination of the M_i, hence its distance to M_0 is at most δ. Singular-value perturbation gives $\sigma_1(B/a) \ge \sigma_1(M_0) - \delta > \delta \ge \sigma_2(B/a)$. This spectral condition can be checked for each constructed head.

Connection to a fixed-input attention logit. Let S_f and M_{if} be candidate and expert operators at frequency f, and define $E_{if} = S_f - M_{if}$. For the same real input states x, y, the scaled head-logit difference satisfies, for every relative position,

$$\begin{aligned} |\widehat{s}(x,y,\Delta) - s_i(x,y,\Delta)| &\le \frac{\|x\|_2\|y\|_2}{\sqrt{d_h}} \sum_f \|E_{if}\|_2 \\ &\le \frac{\|x\|_2\|y\|_2}{\sqrt{d_h}} \sum_f \|E_{if}\|_F \\ &\le \frac{\|x\|_2\|y\|_2}{\sqrt{d_h}} \sqrt{F N_i L_i(S)}. \end{aligned}$$

Here d_h is the real head width. Each rotary factor has unit modulus; the first inequality follows by the triangle and bilinear operator-norm bounds, the second by $\|E\|_2 \le \|E\|_F$, and the last by Cauchy–Schwarz. Individual frequencies may have zero energy; only the total head energy N_i must be positive. This identifies the local functional discrepancy controlled by the objective. The comparison uses identical hidden states; extending it to normalized attention also requires the same visible-key mask and a further argument.

Normalized attention and fixed-value readout. Fix a query, its visible keys and their mask and positions, and let p and q be the expert and constructed softmax distributions. Write $\delta_j = s_j^S - s_j^i$ for each visible key and $r = \max_j \delta_j - \min_j \delta_j$. Then

$$D_{\mathrm{KL}}(p\|q) \le r^2/8, \qquad \mathrm{TV}(p,q) \le \tanh(r/4). \tag{4}$$

For the first bound, set $F(t) = \log \sum_j p_j e^{t\delta_j}$. The KL equals $F(1) - F(0) - F'(0)$; $F''(t)$ is a variance of numbers in an interval of width r, at most $r^2/4$. Taylor's integral remainder proves the inequality. For total variation, write $X_j = q_j/p_j$. Then $\mathbb{E}_p X = 1$ and $\max X/\min X \le e^r$. Bounding the absolute deviation by the chord through its extrema and optimizing that ratio gives $\mathrm{TV}(p,q) \le (e^{r/2}-1)/(e^{r/2}+1)$, the finite-simplex Hilbert-metric comparison (Cohen & Fausti, 2023). If the value vectors v_j are also fixed, with $D_V = \max_{j,k} \|v_j - v_k\|_2$, splitting $p - q$ into its equal-mass positive and negative parts gives

$$\left\| \sum_j (p_j - q_j) v_j \right\|_2 \le D_V \, \mathrm{TV}(p,q). \tag{5}$$

Any logit shift constant across the visible keys has $r = 0$ and leaves softmax unchanged. The bounds apply with hidden inputs, mask, positions and values held fixed.

Visual-state handoff under a common task loss. Let $R(\theta, a)$ be answer CE for one fixed weighted source set, where θ_0, θ_1 are the constructed and task-calibrated Q/K states, $a_0 = A_{\text{init}}$, and the visual basis and preceding interface are fixed. Define $G(a) = R(\theta_0, a) - R(\theta_1, a)$ and

$$I = R(\theta_1, a_0) - R(\theta_0, a_0) - R(\theta_1, 0) + R(\theta_0, 0).$$

Expanding gives $G(0) = G(a_0) + I$ exactly. If R is twice continuously differentiable on the two parameter segments and its mixed derivative has operator norm at most M, the two-variable fundamental theorem of calculus gives

$$|I| \le M\|\theta_1 - \theta_0\|_2 \|a_0\|_2. \tag{6}$$

The same argument between a_0 and any a_1 bounds $|G(a_1) - G(a_0)|$ by $M\|\theta_1 - \theta_0\|_2\|a_1 - a_0\|_2$ when the derivative condition holds along that segment. Retaining the same a_0, fixed basis and visual interface preserves the visual map exactly at the stage boundary; resetting optimizer moments leaves that map unchanged. These identities concern common answer CE; the auxiliary and final visual objectives use their respective frozen reference logits when KL regularization is enabled.

Together, these results specialize low-rank approximation, weak duality, finite-probability comparison and calculus to native attention construction and the visual-state handoff. The finite-step implementation reports a feasible candidate loss and a numerical dual value. These solver diagnostics are distinguished from the exact-arithmetic saddle conditions above. The operator and attention bounds characterize local reconstruction at fixed inputs, while task evaluations measure the complete calibrated model.

G ARCHITECTURE-SPECIFIC OPERATOR FEASIBILITY

The reconstruction principle extends by identifying each architecture's representable operator class. Here we give constructive conditions for ordinary, nonrotary dot-product attention, using matrix factorization and softmax invariance.

Affine projections. For column inputs x, y, let the query and key be $Qx + b_q$ and $Ky + b_k$. Their dot product is

$$[x; 1]^T [Q, b_q]^T K y + (Qx + b_q)^T b_k.$$

The second term is constant across keys for a given query and cancels in softmax under the same allowed-key mask with nonempty support. The effective operator $M = [Q, b_q]^T K$ therefore has rank at most the head width d_h. Conversely, any matrix of rank at most d_h can be factored as AK; the final row of A defines b_q^T, and the remaining rows define Q^T. Padding factors with zeros preserves the original width. This recovers a valid parameterization of the target attention operator. Absolute position embeddings can be included in the fixed inputs; rotary transformations after projection require the frequency-specific construction.

Shared keys. Grouped-query attention (GQA) shares a key/value head across query heads (Ainslie et al., 2023). For a fixed group, stack its effective operators vertically into $\mathcal{M} = [M_1^T, \dots, M_H^T]^T$. With unrestricted affine query factors and one common key factor, joint representability is equivalent to

$$\operatorname{rank}(\mathcal{M}) \le d_h.$$

Necessity follows from $\mathcal{M} = [A_1^T, \dots, A_H^T]^T K$. For sufficiency, factor the stack and partition its left factor into query blocks; each block supplies its query weights and bias as above. Individual rank constraints are insufficient: $\operatorname{diag}(1, 0)$ and $\operatorname{diag}(0, 1)$ each have rank one, but their vertical stack has rank two and cannot share a width-one key. This construction preserves the architecture's existing head-group assignment.

Compact construction and bounds. Let U, V span all target left and right factors, respectively, with orthonormal columns. Projection $S \mapsto UU^T SVV^T$ does not increase rank or Frobenius distance to any target in these spans. Thus restriction to compact coordinates loses no optimum for this rank-constrained objective. For target blocks M_{if}, $N_i = \sum_f \|M_{if}\|_F^2 > 0$, a common feasible rank r, and simplex weights λ, completing the square gives

$$g_r(\lambda) = 1 - \frac{1}{a} \sum_f \sum_{j=1}^{r} \sigma_j(B_f)^2, \qquad a = \sum_i \lambda_i / N_i, \quad B_f = \sum_i \lambda_i M_{if} / N_i.$$

This is a lower bound on the minimax reconstruction objective. A representable candidate gives an upper bound through its worst expert loss. For shared-key reconstruction the target block is the group stack, so normalization is group-level, not independent per-head normalization. Finite dual search and arbitrary choices at singular-value ties do not imply an exact primal optimum.

A boundary that requires a different representation. Postprojection query/key normalization (Henry et al., 2020) is not identified by $Q^T K$ alone. For example, $Q = K = I$ and $Q' = \mathrm{diag}(2, 1/2)$, $K' = \mathrm{diag}(1/2, 2)$ have the same product. With query $(1, 1)^T$ and keys e_1, e_2, their normalized dot products differ. Consequently, the product-only construction above cannot be asserted to preserve that architecture's attention. Likewise, unchanged attention probabilities ensure unchanged readouts only when value inputs and the output projection are held fixed.

H Baseline Evaluation Details

MAny uses FP32 for its core solve. For InternVL, our fixed reference construction uses a deployment multiplier of 3.0, while the MAny baseline uses 2.5. Comparisons between those rows describe complete systems, not the isolated effect of AttnCal's Q/K or visual modules. Each comparison uses the same backbone, benchmark population and scoring protocol. Zero-shot evaluates the base model, whereas Individual evaluates a separate expert for each task; neither is a merged-model baseline.

I Evaluation Assets and Metric Aggregation

InternVL checkpoint provenance. Our evaluation uses the official OpenGVLab InternVL 1.0 checkpoints `InternVL-Chat-ViT-6B-Vicuna-7B` (revision `5a62edecca67848bea28f66156d02e11270cbbd9`) and `InternViT-6B-224px` (revision `084bfb607eb04f373bf0349c49eb6bc3ba094919`), with MCITlib's CLIP-L/14-336 binding.

The final average score (FAA) weights final-stage tasks equally. The continual average (CAA) weights stages equally after averaging the tasks at each stage. Backward transfer (BWT) averages acquisition-to-final changes over the $T-1$ earlier tasks. These quantities use unrounded scores from complete task evaluations.

Zero-shot and Individual means are reference summaries, not sequential acquisition-stage scores. They appear in FAA with † and ‡, respectively, in DCL; Individual additionally denotes separate experts rather than one final model. No missing CAA or BWT is inferred from final endpoints.

J Related Visual Interfaces and Extension Boundaries

This discussion separates the selected shared-residual calibration from possible interface replacements; no replacement is evaluated as AttnCal. MVP and PMA condition visual-projector selection on instruction or multimodal information (Jin et al., 2025; Zhang et al., 2026b), and HARC calibrates routing after MoE merging (Huang et al., 2026). These are relevant to a separate visual-interface extension, not evidence for the static Q/K contribution. Attention-state composition tracks normalized outputs and log-partition values; Attention Matching preserves outputs and mass during KV compaction (Zweiger et al., 2026). An externally weighted mixture of visual-bank readouts would require a matched storage and attention-work comparison against these constructions. Its normalization identity alone would not establish an accuracy advantage.