FERVA: FAITHFUL EXPERT RECONSTRUCTION AND VISUAL ALIGNMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

A compact space can represent every expert update yet exclude their best shared reconstruction. Task-specific input correlations reorient expert contributions, so storing every update need not preserve that optimum. For fixed statistics and expert supports, we establish when an input restriction is lossless and identify the unique minimal faithful space. The restriction loss lower-bounds every merge confined to that input span under the same reconstruction objective. Faithful Expert Reconstruction and Visual Alignment (FerVA) turns this distinction into a two-component method: a static language solve retains full input features in compact output coordinates, while a supervised visual calibrator learns through an auxiliary decoder and transfers a valid visual mixture to the constructed model. In a controlled six-task UCIT study, recovering the omitted component raises final average score (FAA) from 69.54 to 76.04. At correction rank 16, objective-guided recovery reaches 75.54 versus 71.18–71.33 for three random bases. The separate complete method leads reproduced merged baselines in FAA across four UCIT/DCL–backbone settings, reaching 76.20 on UCIT–LLaVA, 4.05 points above the strongest comparator. These results link reconstruction-faithful coordinates to task value in multimodal consolidation.

1 INTRODUCTION

Expert merging consolidates specialist capabilities in one multimodal model (Jin et al., 2022; Gao et al., 2026). Preserving responses across tasks (Chen et al., 2024; Guo et al., 2025c) requires more than storing expert updates: a space containing every expert can exclude their best shared reconstruction because task-specific input correlations reorient the shared solution.

An update's column space specifies output changes; its row space specifies the input combinations producing them. Figure 1 shows why restricting both can discard a shared component even when every expert fits the restriction. Which coordinates preserve the shared response?

An expert-spanning output basis is lossless, whereas faithful input coordinates must contain statistically transformed expert supports. Their orientation, not only their dimension, determines whether the shared optimum remains representable. Faithful Expert Reconstruction and Visual Alignment (FerVA) follows that criterion with full-input language reconstruction and task-supervised composition of frozen visual projectors. A calibrator learns through an auxiliary decoder and transfers without refitting, combining adapted visual tokens with one static language model rather than input-dependent language-expert selection.

The omitted component matters for task performance, not just reconstruction loss. On six-task UCIT, restoring it raises final average score (FAA) from 69.54 to 76.04; at rank 16, objective-guided recovery reaches 75.54 versus 71.18–71.33 for three random bases. This equal-rank comparison tests which directions are restored. The separate complete model reaches 76.20 FAA on UCIT–LLaVA, a 4.05-point lead over the strongest reproduced merged baseline, and also leads in FAA in the other three UCIT/DCL–backbone settings. Complete trajectories test retention alongside these final gains. Our contributions are:

- A fidelity criterion identifying the minimal faithful input space, loss bounds for input-restricted merges, and lossless output compression.

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.

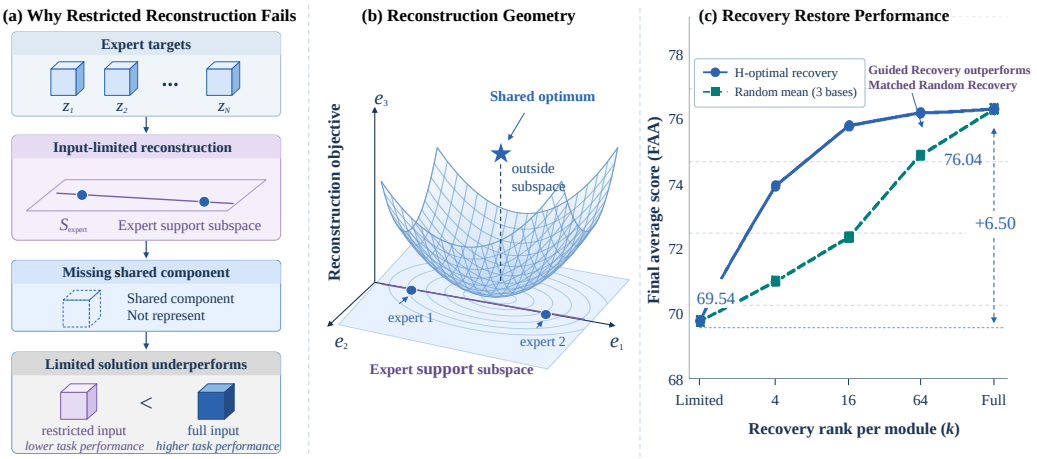

Figure 1: **Reconstruction-faithful coordinates retain task-relevant components. (a)** The expert-support subspace can represent the individual expert targets while excluding a component of their shared reconstruction. **(b)** The geometric schematic places the shared optimum outside that subspace. It illustrates a representational restriction, not a measured neural-network loss surface. **(c)** The recovery comparison uses six UCIT tasks with 3,000 examples each and LLaVA-1.5-7B. Final average score (FAA) is their mean score, with higher values better. Blue circles denote theory-guided recovery in the source second-moment metric, labelled H-optimal; dashed green squares denote the mean of three random bases at the same per-module rank k. Limited and Full denote the input-restricted and full-input endpoints: restoring the omitted component raises FAA from 69.54 to 76.04 (+6.50 points). The comparison tests the task value of the recovered directions, rather than optimality of FAA.

- A two-component construction combining a static full-input language solve with supervised visual composition and simplex deployment.
- Controlled recovery and complete-module tests linking the identified component to task performance, complemented by four continual benchmarks.

2 Related Work

Model merging and functional reconstruction. Model merging consolidates task expertise into a shared model, with different approaches distinguished by the evidence they use to combine parameters. Model soups average fine-tuned weights, task arithmetic combines task vectors, and Fisher-weighted merging incorporates parameter importance (Wortsman et al., 2022; Ilharco et al., 2022; Matena & Raffel, 2022). Alignment and interference-aware methods address complementary aspects of this combination: Git Re-Basin accounts for permutation symmetries, while TIES resolves conflicts among update signs (Ainsworth et al., 2022; Yadav et al., 2023). Functional merging shifts the objective from combining weights to reproducing expert computations. RegMean fits shared layer weights to expert responses using calibration second moments (Jin et al., 2022); IterIS alternates inference and alignment to account for expert and merged-model inputs, while RegMean++ uses representations produced by preceding merged layers (Chen et al., 2025a; Nguyen et al., 2025). MAny applies activation-weighted language reconstruction together with prototype-conditioned visual composition in a multimodal setting (Gao et al., 2026). These approaches expose two distinct design choices: the behavior used to define agreement among experts and the parameter family allowed to realize that agreement. Within response reconstruction, the latter raises a precise question: which compact coordinates preserve the original fitting problem's optimum?

Low-rank representations and reconstruction fidelity. Low-rank adaptation makes specialist updates compact and exposes input- and output-side factors that can be reused during consolidation

(Hu et al., 2021). KnOTS employs a joint SVD to align expert updates in a shared output-side basis while retaining full-width input-side coordinates (Stoica et al., 2025). Core Space builds shared bases on both sides, representing each expert exactly in a smaller core and preserving scalar linear combinations performed consistently in those coordinates (Panariello et al., 2026). Such representations establish which parameter combinations remain expressible; a functional reconstruction additionally depends on how task-specific inputs weight those combinations. In an activation-weighted solve, each task's second moment acts on its update before a shared inverse combines the contributions, so the input directions needed by the optimizer can differ from the original expert row space. Output-Space Projection also makes the feasible merge family explicit, relating output discrepancies to projection bounds and realizable merge coefficients (Evans et al., 2026). FerVA studies this representational question for a full layer-update matrix under task-specific statistics. It characterizes the shared output space that preserves the dense ridge optimum, the minimal faithful input space for fixed expert supports and statistics, and the exact objective penalty of a smaller input space. The resulting criterion connects compact representation to the reconstruction problem it must preserve.

Continual multimodal learning and visual expert composition. Continual learning preserves earlier capabilities through parameter-importance regularization, replay, and task-specific prompts (Kirkpatrick et al., 2017; Buzzega et al., 2020; Wang et al., 2022). CoIN and MCITlib extend this problem to multimodal instruction sequences, while SEFE examines instruction behavior and the interpretation of forgetting (Chen et al., 2024; Guo et al., 2025c; Chen et al., 2025b). Retaining task experts offers a complementary consolidation setting: specialist responses remain available, but a shared model must combine them at both the language backbone and the visual interface. At that interface, MAny mixes projector outputs using prototype similarity, MVP learns instruction-grounded projector selection, and PMA combines progressive projector expansion with multimodal routing (Gao et al., 2026; Jin et al., 2025; Zhang et al., 2026). Signed aggregation of features across visual-encoder depths explores another form of visual composition (Zhu et al., 2026). Together, these studies distinguish selecting visual expertise from adapting the representation consumed by the language model. FerVA couples reconstruction-faithful language fusion with answer-supervised calibration of frozen projector-output mixtures. Its calibrator is fitted through an auxiliary decoder and transferred to the constructed language model, linking the retained visual expertise to the model that uses it.

3 METHODOLOGY

3.1 EXPERT-RESPONSE RECONSTRUCTION

An expert's parameters matter through the responses they produce on task inputs. FerVA therefore reconstructs those responses, using source examples to identify which input directions the shared model must preserve. Figure 2 separates this language construction from visual calibration. For one language projection across T task experts, let W_0 be the base weight and $D_t \in \mathbb{R}^{d_{\text{out}} \times d_{\text{in}}}$ expert t's native-scaled update. The dimensions count output and input features. For source example i, let X_{ti} contain its nonpadding layer-input activations after image expansion. With n_t examples and equal task weight $a_t = 1/T$, define the uncentered second moment $K_t = n_t^{-1} \sum_i X_{ti}^\top X_{ti}$. These statistics make reconstruction sensitive to task inputs: the same parameter error can matter differently on different experts' examples. Source budgets and collection conventions are in Appendix A.

Choose an orthonormal output basis $Q \in \mathbb{R}^{d_{\text{out}} \times r}$ spanning all expert-update columns. Parameterize the shared increment as $D = QU^\top$, where $U \in \mathbb{R}^{d_{\text{in}} \times r}$ retains every input feature. Compact output coordinates reduce the solve without prematurely excluding useful input combinations. We fit the layer-response changes with ridge coefficient $\lambda \geq 0$:

$$\min_U \sum_t \frac{a_t}{n_t} \sum_i \|X_{ti}U - X_{ti}D_t^\top Q\|_F^2 + \lambda \|U\|_F^2. \tag{1}$$

Define the normal matrix $H = \sum_t a_t K_t + \lambda I$ and right-hand side $N = \sum_t a_t K_t D_t^\top Q$. Solve $HU_* = N$, then form $W_{\text{lang}} = W_0 + \gamma Q U_*^\top$ with a fixed deployment gain γ. The response objective follows activation-weighted reconstruction (Jin et al., 2022); our question is which compact coordinates preserve its solution. Fidelity concerns the unscaled reconstruction, while gain controls deployment strength (Appendix A).

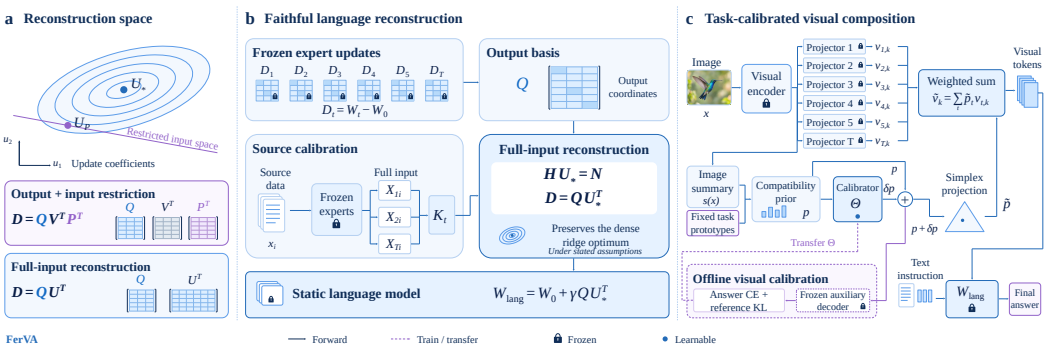

Figure 2: **FerVA combines full-input language reconstruction with visual calibration.** **(a)** The purple input restriction can exclude the best reconstruction, even when both approaches use the same output space. **(b)** Frozen task-expert updates and source examples define a shared output basis and fitted input coefficients. Together they produce one static language update without selecting an expert at inference. The displayed input and output summaries determine those coefficients before deployment scaling. **(c)** Frozen visual projectors supply image tokens. A calibrator uses the image summary and fixed task prototypes to adjust their mixture, learns through a frozen auxiliary decoder, and transfers to the constructed language model without refitting. Its deployment weights form a valid mixture. Lock icons mark frozen components, blue dots mark learned parameters, and dashed arrows mark calibration or transfer rather than forward computation.

3.2 Which coordinates preserve the optimum?

The distinction is between compactly expressing expert outputs and restricting which input combinations the shared model can use.

Proposition 1 (Lossless output-space parameterization)**.** *For expert updates D_t, let $K_t \succeq 0$ be their fixed input second moments, $a_t > 0$ their task weights, and $\lambda \geq 0$ the isotropic ridge coefficient. If the resulting normal matrix $H = \sum_t a_t K_t + \lambda I$ is positive definite, an orthonormal output column-space basis Q spanning every D_t preserves the dense reconstruction optimum under $D = QU^\top$, where U contains the input-coordinate coefficients.*

The output basis reduces the right-hand side to as many columns as the joint expert output rank, while retaining the same normal matrix and optimum.

Proposition 2 (Cost of an input-space restriction)**.** *Let $H \succ 0$ be the second-moment-plus-ridge normal matrix, N its reconstruction right-hand side, and P an orthonormal basis of retained input-feature directions. For the reduced quadratic $J(U) = \frac{1}{2}\operatorname{tr}(U^\top HU) - \operatorname{tr}(U^\top N)$, the unrestricted minimizer is $U_* = H^{-1}N$; restricting its columns to $\operatorname{range}(P)$ gives*

$$U_P = P(P^\top HP)^{-1}P^\top N, \qquad J(U_P) - J(U_*) = \tfrac{1}{2}\|H^{1/2}(U_P - U_*)\|_F^2.$$

It is lossless exactly when $\operatorname{range}(U_*) \subseteq \operatorname{range}(P)$.

Here J is half the fitting objective up to a constant. Every candidate confined to this input space incurs at least the stated gap. When P spans the experts' joint row space, this includes fixed left-weighted combinations of their updates (Appendix C.6). The restricted optimum therefore bounds what any such combination can achieve in the layer objective. Rescaling U_P cannot restore its omitted directions. A two-feature example has $U_* = (1, -1/6)^\top$ and restriction cost $1/24$ (Appendix C.2).

Proposition 3 (Minimal faithful input space for fixed expert supports)**.** *For T experts, fix orthonormal bases P_t and Q_t specifying each expert's* allowed *input and output supports, with at least one column in Q_t. The coefficients C_t in $D_t = Q_t C_t^\top P_t^\top$ vary independently and may be zero or rank deficient; individual family members need not have the same exact rank. The expert input second moments $K_t \succeq 0$, positive task weights a_t, and ridge coefficient $\lambda \geq 0$ remain fixed, as does the positive-definite normal matrix $H = \sum_t a_t K_t + \lambda I$. Collect the statistically weighted input supports in the matrix $B = [a_1 K_1 P_1, \ldots, a_T K_T P_T]$. An input space preserves the unrestricted*

reconstruction optimum for every such coefficient choice if and only if it contains

$$\mathcal{L}_* = \mathrm{range}(H^{-1}B), \qquad \dim \mathcal{L}_* = \mathrm{rank}(B) \le \min\{d_{\mathrm{in}}, \textstyle\sum_t \dim(P_t)\}. \tag{2}$$

Thus $\mathcal{L}_*$ is the unique minimal faithful input space for this fixed-statistic support family, not the minimum space required by one already fixed collection of experts.

The faithful space can be compact; task statistics determine its orientation. Full-input fitting includes it without first solving for another basis. Proofs and numerical analysis are in Appendices C.4, C.3 and C.7.

3.3 Task-calibrated visual composition

Preserving language responses does not determine which visual features best support an answer. We therefore calibrate the frozen projector mixture through answer supervision, adapting its composition rather than retraining the experts. For input $x = (I, q)$, let $s(x)$ be the frozen image summary, μ_t a task prototype, and $v_{t,k}(x)$ projector t's output at image-token position k. Prototype compatibility supplies the prior $p_t \propto \exp(\cos(s(x), \mu_t)/\eta)$ with temperature η (Gao et al., 2026). For six experts, a zero-initialized matrix Θ learns bounded adjustments and a centered coefficient correction:

$$h_{1:5} = \tanh([1; s(x)]^\top \Theta), \quad h_6 = 0, \qquad \delta p_t = p_t(h_t - \textstyle\sum_j p_j h_j). \tag{3}$$

The zero-sum correction redistributes weight around the prototype prior, rather than learning another visual expert. Fitting uses the affine combination $\sum_t (p_t + \delta p_t) v_{t,k}$; deployment projects its coefficients onto the probability simplex without refitting:

$$\widetilde{p} = \arg\min_{w \ge 0,\, \mathbf{1}^\top w = 1} \|w - (p + \delta p)\|_2^2, \qquad \widetilde{v}_k = \sum_t \widetilde{p}_t v_{t,k}. \tag{4}$$

Only Θ is trained. For target answer token y_{ik} at answer position $k \in \mathcal{A}_i$, the fitted distribution π_Θ and fixed reference π_0 receive identical teacher-forced prefixes. Their task- and example-balanced loss combines answer supervision with a reference penalty of weight ω_{KL}:

$$\mathcal{L}_{\mathrm{vis}} = \mathbb{E}_i \frac{1}{|\mathcal{A}_i|} \sum_{k \in \mathcal{A}_i} \big[-\log \pi_\Theta(y_{ik}) + \omega_{\mathrm{KL}} D_{\mathrm{KL}}(\pi_0 \| \pi_\Theta) \big]. \tag{5}$$

The deployed tokens lie in the convex hull of the frozen projector outputs. All encoders, projectors and language weights remain frozen. The omitted component also defines the auxiliary decoder: under a fixed rank budget, it restores directions with the largest reconstruction benefit. Visual adjustments are learned in this context, with the unadjusted prior as reference. The fitted coefficients then transfer to the full-input decoder, separating visual adaptation from changes to the final language weights. Appendix C.8 derives the recovery, and Appendices A.1 and D describe the calibration procedure and source data. Inference composes visual tokens once and runs one static language model.

4 Experiments

4.1 Evaluation protocol

We test consolidation, omitted-direction recovery, and deployment sensitivity on six-task UCIT and five-domain DCL using LLaVA-1.5-7B and InternVL-Chat-7B with MCITlib-native decoding and scoring (Guo et al., 2025c; Liu et al., 2024). UCIT uses 3,000 examples per task, whereas DCL uses its complete native populations. Calibration and development use separate source data. Appendix B specifies the datasets, native task scorers and baselines. Experiments and baseline reproductions use a shared pool of 32 NVIDIA A100 GPUs, with allocation varying by job.

For task score $R_{i,j}$ after stage i on task j, let $\bar{R}_i = i^{-1} \sum_{j \le i} R_{i,j}$. Final average score (FAA), continual average (CAA) and backward transfer (BWT) are

$$\mathrm{FAA} = \tfrac{1}{T} \sum_j R_{T,j}, \quad \mathrm{CAA} = \tfrac{1}{T} \sum_i \bar{R}_i, \quad \mathrm{BWT} = \tfrac{1}{T-1} \sum_{j<T} (R_{T,j} - R_{j,j}).$$

All are higher-is-better. BWT compares the model with its own acquisition stages. Complete 21-cell UCIT and 15-cell DCL matrices supply CAA and BWT. Contrasts use unrounded scores; tables retain two decimals. Zero-shot and Individual are unmerged references, excluded from method rankings.

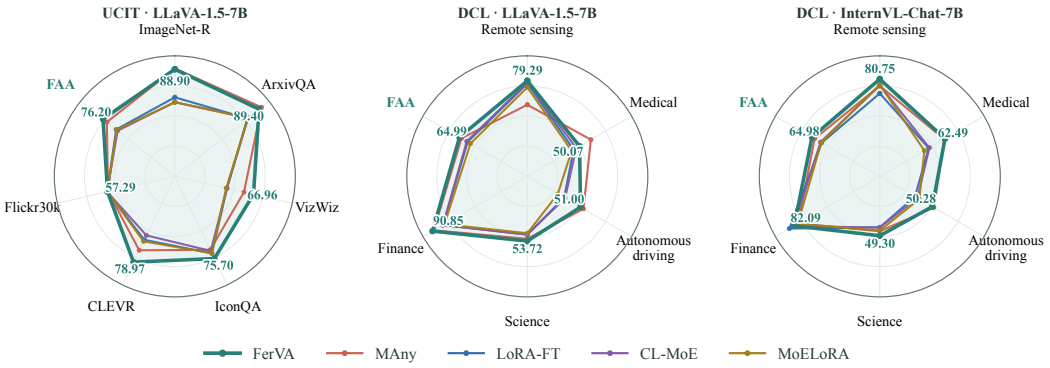

Figure 3: **Final-task profiles across three benchmark–backbone settings.** Each radar plots task scores and their final average score (FAA) on a common 0–100 percentage scale. The panels show six-task UCIT with LLaVA-1.5-7B, five-domain DCL with LLaVA-1.5-7B, and five-domain DCL with InternVL-Chat-7B, selected by FerVA's largest available FAA leads over the strongest other merged method in each setting. The four comparison lines are local MAny, LoRA-FT, CL-MoE, and MoELoRA evaluations. Numeric labels mark FerVA's values only. The DCL–InternVL panel uses the final five-task endpoint from its native stage matrix.

Table 1: **UCIT performance on LLaVA-1.5-7B.** Each task uses 3,000 evaluation examples. IN-R, Arxiv, VizWiz, Icon, CLEVR and Flickr denote ImageNet-R, ArxivQA, VizWiz-caption, IconQA, CLEVR-Math and Flickr30k. Source identifies the method. Final average score (FAA) averages final tasks; continual average score (CAA) averages stage means. Backward transfer (BWT) is the subsequent change on earlier tasks. Higher is better; scores are percentages and BWT is in points. Daggers mark unmerged zero-shot (†) and separate-expert (‡) means; their CAA/BWT are inapplicable.

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., 2024)	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
RegMean++	(Nguyen et al., 2025)	75.83	88.80	44.70	64.90	54.83	54.78	63.97	76.25	-4.73
CSP	(Sun et al.)	85.60	92.03	46.73	71.80	65.20	56.95	69.72	78.45	-1.22
IterIS	(Chen et al., 2025a)	69.90	87.23	43.23	65.47	63.13	49.45	63.07	75.44	-5.36
PASs-MoE	(Hou et al., 2026)	76.27	78.67	45.41	65.83	52.63	57.19	62.67	76.93	-12.00
FerVA	Ours	88.90	89.40	66.96	75.70	78.97	57.29	76.20	82.83	0.16

4.2 Continual Performance Across Backbones

Tables 1–4 retain the complete local comparison in all four settings. The Source column identifies each method under the same evaluation protocol.

On UCIT–LLaVA, FerVA reaches 76.20 FAA, 82.83 CAA and 0.16 BWT, improving on the strongest other merged FAA baseline by 4.05 points. Relative to that comparator, CAA and BWT improve by 1.23 and 1.68 points, so the advantage extends beyond the final checkpoint. VizWiz, IconQA and CLEVR-Math improve by 8.10, 8.03 and 10.90 points, while ImageNet-R and ArxivQA are slightly lower. Positive BWT is consistent with beneficial backward transfer (Lopez-Paz & Ranzato, 2017; Riemer et al., 2018). The gains span captioning and reasoning rather than a sin-

Table 2: **UCIT performance on InternVL-Chat-7B.** Evaluations use 3,000 examples per task and InternVL 1.0 preprocessing. IN-R, Arxiv, VizWiz, Icon, CLEVR and Flickr denote ImageNet-R, ArxivQA, VizWiz-caption, IconQA, CLEVR-Math and Flickr30k. Source identifies the method. Final average score (FAA) averages final tasks; continual average score (CAA) averages stage means. Backward transfer (BWT) is the subsequent change on earlier tasks. Higher is better; scores are percentages and BWT is in points. Daggers mark unmerged zero-shot (†) and separate-expert (‡) means; their CAA/BWT are inapplicable.

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., 2024)	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
FerVA	Ours	92.67	95.07	62.82	73.27	59.70	58.63	73.69	84.38	-0.54

Table 3: **DCL performance on LLaVA-1.5-7B.** Each domain uses its complete native evaluation population. RS, Med, AD, Sci and Fin denote remote sensing, medical, autonomous driving, science and finance. Source identifies the method. Final average score (FAA) averages final tasks; continual average score (CAA) averages stage means. Backward transfer (BWT) is the subsequent change on earlier tasks. Higher is better; scores are percentages and BWT is in points. Daggers mark unmerged zero-shot (†) and separate-expert (‡) means; their CAA/BWT are inapplicable.

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., 2024)	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
FerVA	Ours	79.29	50.07	51.00	53.72	90.85	64.99	64.56	1.46

gle task type; the controlled tests below separate reconstruction fidelity from the two-component system.

Table 2 summarizes final tasks and metrics from the 21-cell InternVL trajectory. Relative to the strongest FAA baseline, FAA rises from 73.01 to 73.69, CAA from 84.12 to 84.38, and BWT from -1.69 to -0.54. The final-average gain is accompanied by less forgetting over the full sequence. This tests the two-component system on a second backbone; the matched LLaVA controls below isolate its construction choices.

For DCL–LLaVA, FerVA leads the strongest other merged model by 1.89 FAA points and attains positive BWT of 1.46: earlier domains improve on average after later experts are incorporated. Its CAA of 64.56 is below CL-MoE's 66.16, distinguishing final retention from average performance during acquisition. Relative to the strongest FAA comparator, RS gains 19.85 points while Med loses 10.84; consolidation reallocates task strengths.

On DCL–InternVL, FerVA leads the strongest other merged models by 2.51 FAA and 1.36 CAA points, reaching 64.98 and 68.12. Its -0.24 BWT is close to HiDE's -0.15 while retaining a much

Table 4: **DCL performance on InternVL-Chat-7B.** Each domain uses its complete evaluation population and InternVL 1.0 preprocessing. RS, Med, AD, Sci and Fin denote remote sensing, medical, autonomous driving, science and finance. Source identifies the method. Final average score (FAA) averages final tasks; continual average score (CAA) averages stage means. Backward transfer (BWT) is the subsequent change on earlier tasks. Higher is better; scores are percentages and BWT is in points. Daggers mark unmerged zero-shot (†) and separate-expert (‡) means; their CAA/BWT are inapplicable.

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., 2024)	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
FerVA	Ours	80.75	62.49	50.28	49.30	82.09	64.98	68.12	-0.24

stronger final score. Figure 3 locates these aggregate gains across domains; the following controls isolate reconstruction.

4.3 Task value of the reconstructed component

Figure 4 restores equal-rank corrections while holding expert statistics, visual interface and deployment gain fixed. At rank 16, second-moment-directed recovery reaches 75.54 FAA versus 71.18–71.33 for three random bases: direction choice matters beyond added capacity. The restricted and full-input endpoints differ by 6.50 points, with VizWiz and CLEVR-Math improving most and ArxivQA decreasing. Equal rank controls added capacity, while the theory determines which directions to restore. Correction energy can differ, and these fixed-interface endpoints precede visual calibration.

Table 5 tests complete-module removals without retraining. Equal expert language updates give 61.28 FAA, equal projector outputs give 66.73, and replacing both gives 51.50. Both modules therefore contribute without adapting the surviving module. The visual contrast replaces both the prototype-based composition and its learned correction.

Table 5: **Both FerVA components contribute to final UCIT performance.** Language and Visual indicate active learned components. Removal replaces one composition by equal expert contributions without retraining the other. All rows use the same six UCIT tasks, 3,000 examples per task and LLaVA-1.5-7B. Final average score (FAA) is averaged over tasks. FerVA uses affine-trained visual calibration with nearest-simplex deployment.

Configuration	Language	Visual	FAA ↑
FerVA	On	On	76.20
Remove language module	Off	On	61.28
Remove visual module	On	Off	66.73
Remove both modules	Off	Off	51.50

4.4 Deployment and source-budget sensitivity

Figure 5 separates rescaling a fitted solution from refitting it with a different source budget. Full-input reconstruction leads throughout the displayed gain window 1.2–1.6. Outside this window, the original scan reverses at 1.8 and 2.0: FAA is 66.54 and 60.54 versus 70.33 and 70.57 for the restricted control. The separate full-input budget scan changes little over 64–512 source examples

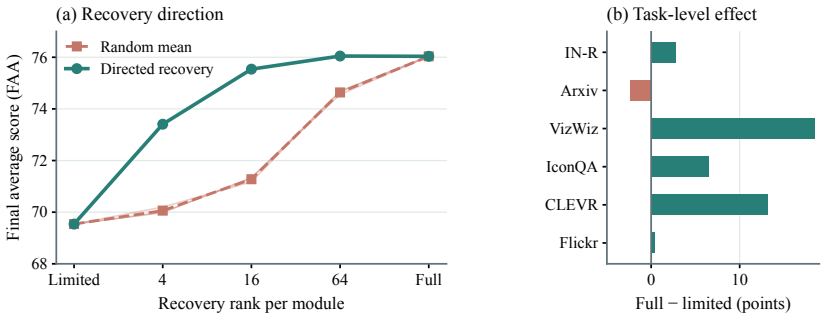

Figure 4: **Recovering the omitted reconstruction component reveals its task value. (a)** Equal-rank corrections use second-moment-optimal directions or three nested random bases (thin curves; dashed mean). Rank counts directions per module; Limited and Full are shared endpoints. **(b)** Full-input minus restricted task scores average +6.50 points, with an ArxivQA decrease. IN-R, VizWiz, CLEVR and Flickr denote ImageNet-R, VizWiz-caption, CLEVR-Math and Flickr30k. Both panels share six experts, 256 source examples per task, visual interface, gain 1.5 and native UCIT–LLaVA scoring on 3,000 examples per task. Final average score (FAA) is averaged over tasks; quadratic-optimal directions need not maximize FAA.

per task. Thus the two scans probe estimation and scaling, while the recovery experiment tests coordinate fidelity. Each addresses a different part of the construction, separately from the complete FerVA endpoint.

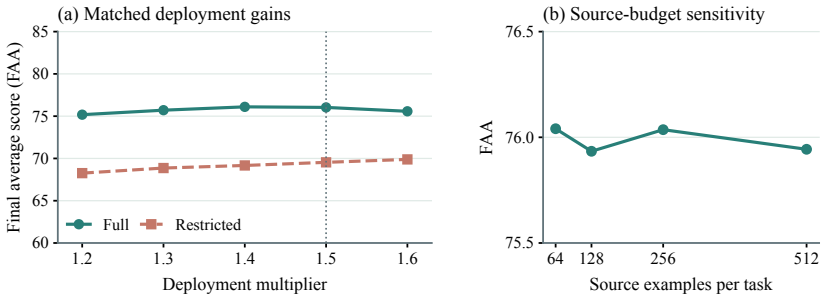

Figure 5: **Deployment scaling and source-sample budget probe different sensitivities. (a)** Fixed full-input and restricted solutions are rescaled over a post-hoc window; the guide marks gain 1.5. The gain-1.3 setting supplements the original scan; its unplotted 1.8 and 2.0 settings reverse the ordering. **(b)** Full-input construction is refitted with nested source budgets at gain 1.5, without a restricted-solution comparison. Both panels use native LLaVA scoring on 3,000 examples per UCIT task and report final average score (FAA) over six tasks; higher is better.

5 Conclusion

FerVA distinguishes a compact representation of experts from a faithful representation of their shared reconstruction. A lossless output basis and full input coordinates preserve the fixed-statistic optimum, while supervised visual composition aligns frozen projector responses with the constructed decoder. Controlled recovery yields 6.50 FAA points, with equal-rank tests distinguishing the recovered directions' value from added capacity. The complete system also improves final performance and retention across four evaluated settings. This provides a criterion for choosing compression, rather than expansion for its own sake: the retained coordinates must accommodate the solution required by task statistics. Recovery tests then connect this geometric criterion to the value of the omitted directions. The design principle is to compress outputs without losing the optimum and retain the statistically required input directions, rather than equating expert representability with faithful reconstruction.

REFERENCES

Samuel K Ainsworth, Jonathan Hayase, and Siddhartha Srinivasa. Git re-basin: Merging models modulo permutation symmetries. *arXiv preprint arXiv:2209.04836*, 2022.

Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pp. 65–72, 2005.

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, 2024.

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.

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

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.

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.

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.

Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pp. 74–81, 2004.

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.

Ashin Mukherjee and Ji Zhu. Reduced rank ridge regression and its kernel extensions. *Statistical analysis and data mining: the ASA data science journal*, 4(6):612–622, 2011.

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*, 2025.

Aniello Panariello, Daniel Marczak, Simone Magistri, Angelo Porrello, Bartłomiej Twardowski, Andrew Bagdanov, Simone Calderara, and Joost van de Weijer. Accurate and efficient low-rank model merging in core space. *NeurIPS*, 38:61793–61825, 2026.

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.

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.

George Stoica, Pratik Ramesh, Boglarka Ecsedi, Leshem Choshen, and Judy Hoffman. Model merging with svd to tie the knots. In *ICLR*, volume 2025, pp. 4501–4519, 2025.

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 *Forty-third ICML*.

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

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 *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.

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

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.

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

Xuanyu Zhu, Yan Bai, Yang Shi, Yihang Lou, Yuanxing Zhang, Jing Jin, and Yuan Zhou. Beyond the last layer: Multi-layer representation fusion for visual tokenization. *arXiv preprint arXiv:2605.10780*, 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 connects the construction, its analysis and its evaluation. Appendix A specifies how to construct and calibrate FerVA, while Appendix D records the numerical and fitting settings. Appendix C proves when coordinates preserve the shared optimum and quantifies the cost of a restriction. Appendix B gives evaluation and component-control details; the visual and transfer studies examine calibration under a fixed reference and generalization to new domains.

A CONSTRUCTION AND CALIBRATION DETAILS

Construction and deployment. Each language module uses an orthonormal basis of the joint expert-update column space. Expert activations define the full-input least-squares system, and cached calibration statistics determine regularization. A fixed gain scales the solved increment before inference. The LLaVA matched comparisons use deployment gain $\gamma = 1.5$; InternVL–UCIT uses 1.2. The fidelity theorem characterizes the solution before deployment scaling. For six rank-16 experts, the shared output basis needs at most 96 coordinates. The same experts and statistics define the auxiliary decoder used to fit the affine projector coefficients. Those coefficients transfer without refitting and are projected onto the simplex for deployment with the full-input language model. The following details specify the shared statistics first, then the visual fit and transfer rule.

For a collection of T task experts and one attention or feed-forward linear projection in the language backbone, let W_0 denote its shared base weight matrix and $D_t = W_t - W_0 \in \mathbb{R}^{d_{\text{out}} \times d_{\text{in}}}$ the increment of expert t, whose layer weights are W_t. The dimensions d_{in} and d_{out} count input and output features. Here t indexes experts, i indexes calibration examples, and n_t is the number of examples for expert t. For calibration example i, $X_{ti} \in \mathbb{R}^{\ell_{ti} \times d_{\text{in}}}$ contains its nonpadding token input activations to that linear projection under expert t, after image-token expansion, with ℓ_{ti} the number of retained token positions. For each task, we collect its uncentered input second-moment matrix K_t and assign the equal task weight a_t before solving for the merged weights. Both remain fixed during that solve:

$$K_t = \frac{1}{n_t} \sum_i X_{ti}^\top X_{ti}, \qquad a_t = \frac{1}{T}.$$

We use the uncentered second moment K_t, retaining both the mean and correlation structure of the layer inputs. Examples share the same outer coefficient within a task; token contributions are summed within each example, so longer expanded sequences can contribute more. The evaluated UCIT construction uses six experts and 256 examples per task. We suppress the layer index below; each attention and feed-forward linear map is constructed independently. For each expert, the collection forward loads its native-scaled language update and uses the same frozen six-projector, prototype-conditioned visual composition. Thus the language experts are measured under a common visual composition throughout statistics collection. The fixed source conversation template supplies teacher-forced answer prefixes; after image-token expansion, all nonpadding layer-input positions, including image, prompt and answer-context positions, enter X_{ti}. The aligned answer mask is checked separately and is not the statistics mask.

We fix a source-scaled ridge coefficient before each module solve and use the same value across the compared parameterizations. Regularization is applied to the full-input normal system, independently of the coordinates used to express it. The theoretical statements requiring a positive-definite normal matrix apply when regularization or the source Gram ensures that condition. Statistics, construction and inference are separate numerical stages; the comparison protocols below identify their precision.

A.1 VISUAL CALIBRATION AND DEPLOYMENT DETAILS

For multimodal input $x = (I, q)$, comprising image I and instruction q, let $s(x)$ be the frozen image encoder's classification-token representation, μ_t a fixed task prototype, and $v_{t,k}(x)$ the image-token embedding produced by frozen vision-to-language projector t at position k. Each projector maps a visual-encoder feature vector to the language decoder's embedding dimension, producing

the visual tokens supplied to the decoder. The index t denotes the same task expert as in language reconstruction. We retain prototype compatibility $p_t(x) \propto \exp(\cos(s(x), \mu_t)/\eta)$ as a prior over projector responses, with $\sum_t p_t = 1$ and a fixed temperature η (Gao et al., 2026). The learned component adjusts this prior to combine frozen projector outputs. Because it is trained through the decoder's answer loss, the adjustment is guided by how the visual representations support answers, not just by how well they identify a task.

For the six-expert implementation, initialize the visual calibration matrix $\Theta \in \mathbb{R}^{1025\times 5}$ to zero. Its bounded task adjustments h_t induce a centered correction δp_t to the projector prior:

$$h_{1:5} = \tanh([1; s(x)]^\top \Theta), \quad h_6 = 0, \qquad \delta p_t = p_t\big(h_t - \textstyle\sum_j p_j h_j\big). \tag{6}$$

We use an anchored five-output parameterization with $h_6 = 0$. The centered correction is invariant to an unconstrained common offset; the bounded $\tanh$ map fixes an anchored five-output parameterization. The sixth projector remains available through its prior and centered correction. Centering gives $\sum_t \delta p_t = 0$. Here δp is the visual coefficient correction, distinct from the language-space dimension r. Fitting uses the affine combination $\sum_t (p_t + \delta p_t) v_{t,k}$, shared across image positions. At deployment, project the same fitted coefficients onto the probability simplex to obtain the valid projector weights $\widetilde{p}$ and their combined visual embedding $\widetilde{v}_k$:

$$\widetilde{p} = \arg\min_{w \geq 0,\ \mathbf{1}^\top w = 1} \|w - (p + \delta p)\|_2^2, \qquad \widetilde{v}_k = \sum_t \widetilde{p}_t v_{t,k}. \tag{7}$$

This standard Euclidean projection produces nonnegative weights without refitting. Define the unadjusted embedding as $v_k^0 = \sum_t p_t v_{t,k}$. The training mixture is equivalently $v_k^0 + \sum_t \delta p_t v_{t,k}$; deployment uses $v_k^0 + \sum_t (\widetilde{p}_t - p_t) v_{t,k}$. Both recover v_k^0 at $\Theta = 0$. Together, the compatibility prior, learned correction and deployment projection form the visual module.

Training objective and decoder context. Only the visual coefficients Θ are trained. For the answer-token positions $\mathcal{A}_i$ of example i, let π_Θ be the answer distribution under the fitted visual mixture and π_0 the frozen auxiliary decoder's answer distribution under the unadjusted visual prior. At each answer position, we train on the correct next token and penalize deviation from the fixed reference distribution using Kullback–Leibler (KL) divergence, with reference-penalty weight $\omega_{\mathrm{KL}} = 0.1$. Both distributions receive the same preceding tokens. With y_{ik} denoting the target answer token and $y_{i,<k}$ its prefix, the objective is

$$\begin{aligned} \mathcal{L}_{\mathrm{vis}} = \mathbb{E}_{\text{task, example}} \frac{1}{|\mathcal{A}_i|} \sum_{k \in \mathcal{A}_i} \Big[& -\log \pi_\Theta(y_{ik} \mid x_i, y_{i,<k}) \\ & + 0.1\, D_{\mathrm{KL}}(\pi_0(\cdot \mid x_i, y_{i,<k}) \,\|\, \pi_\Theta(\cdot \mid x_i, y_{i,<k})) \Big]. \end{aligned} \tag{8}$$

Tasks and examples are weighted equally. Teacher and student receive the same prefix and aligned answer positions; image, prompt and padding positions do not receive supervision. The vision encoder, projectors, prototypes and language decoder are frozen. Access to source answers makes this supervised offline calibration.

To fit the visual mixture, we use an auxiliary decoder that partially restores the reconstruction component omitted by the input-restricted solution. We then transfer the fitted visual coefficients to the final language weights W_{lang} without refitting. The auxiliary decoder uses the same language statistics and output basis, with P spanning the joint row space of the expert weight updates. Starting from U_P, denote the omitted coefficient component by $E = U_* - U_P$. The matrix V_{16} contains up to sixteen leading eigenvectors of $E^\top H E$, selecting the directions that recover the most reconstruction loss under a rank-16 budget. The auxiliary update coordinates are $U_P + E V_{16} V_{16}^\top$, with the same deployment gain. The fixed teacher π_0 uses this decoder and the unadjusted visual prior. Before applying the deployment multiplier, the correction is the best rank-16 recovery of the omitted coefficient component for the declared fixed-statistic quadratic, as derived in Appendix C.8. The auxiliary decoder supplies the calibration context, and the learned visual coefficients are then transferred to the full-input language decoder. Both decoders are constructed within FerVA. This transfer and deployment rule is fixed for the complete evaluation; the reconstruction result characterizes the auxiliary decoder, while the module tests below assess the deployed visual composition as a whole.

Deployment and component tests. Inference composes the visual tokens once per image and runs the static constructed language model. Complete visual removal replaces the entire image-conditioned composition by the equal mean of the same projector outputs, keeping the vision encoder and image input. Complete language removal uses the frozen native-scale equal mean of expert increments. The resulting four-way test is reported in Table 5; retained modules are not retrained. This complete visual removal replaces both the prototype-based composition and its fitted correction; it measures the contribution of the whole visual module, not the incremental effect of the correction alone. The matched language comparison instead varies the input restriction while holding source statistics and the visual interface fixed.

B Additional Evaluation Details

B.1 Benchmarks and evaluation protocol

We evaluate LLaVA-1.5-7B and InternVL-Chat-7B on UCIT and DCL. UCIT evaluations use 3,000 ordered examples per task, the explicitly identified visual interface for each method, strict ArxivQA whole-string case-insensitive matching without trimming or option extraction, and the frozen official seven-metric corpus caption aggregation. Development slices serve method development, while reported UCIT results use the complete evaluation population.

Our backbone follows LLaVA visual instruction tuning and its 1.5 configuration (Liu et al., 2023; 2024). ImageNet-R evaluates visual recognition (Hendrycks et al., 2021), and ArxivQA evaluates scientific figures (Li et al., 2024). VizWiz evaluates accessible-image captioning (Gurari et al., 2020), while IconQA tests abstract diagram reasoning (Lu et al., 2021). CLEVR-Math tests compositional arithmetic (Lindström & Abraham, 2022), and Flickr30k evaluates image description (Young et al., 2014). The frozen UCIT subsets and scoring implementation define the evaluation protocol. Caption scores aggregate BLEU-1/2/3/4 (Papineni et al., 2002), METEOR (Banerjee & Lavie, 2005), ROUGE-L (Lin, 2004), and CIDEr (Vedantam et al., 2015). We report their frozen corpus-level aggregate.

B.2 Metrics

Let $R_{i,j}$ be the task score after learning stage i, evaluated on task j, and let T be the number of tasks. Write $\bar{R}_i = i^{-1} \sum_{j=1}^{i} R_{i,j}$ for the mean task score at stage i. The final average score (FAA) averages final-stage task scores, the continual average (CAA) averages these stage means, and backward transfer (BWT) measures subsequent change on earlier tasks. We use the following definitions:

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

BWT is measured in percentage points, with positive values indicating backward transfer. Positive BWT compares the final shared model with its own first-learned task scores on average. All three metrics are higher-is-better. Metric equivalence requires the same task sequence, task weighting and scoring protocol. Only FAA requires final endpoints alone. The other metrics require the relevant stage-wise evaluations. For caption tasks, $R_{i,j}$ is the frozen corpus-score aggregate rather than an exact-match accuracy. Displayed results are rounded to two decimal places; reported differences are calculated from the unrounded scores.

B.3 Baselines and comparison scope

Tables 1–4 compare local evaluations in four benchmark–backbone settings: UCIT and DCL on LLaVA and InternVL. Within each setting, the methods use the data ordering, decoding and scoring specified above. The Source column cites the original method, or the benchmark configuration for Zero-shot and Individual. Zero-shot evaluates the unadapted base model, while Individual evaluates each task's separate expert. These two reference rows do not compete as merged models.

The LLaVA UCIT panel additionally includes local evaluations of RegMean++, CSP, IterIS, and PASs-MoE. These rows use the same 3,000-example-per-task UCIT evaluation setting.

The baseline selection follows MCITlib and MAny (Guo et al., 2025c; Gao et al., 2026). DCL retains its native evaluation populations and score definitions rather than inheriting UCIT's 3,000-example protocol. Each of FerVA's four settings has a complete continual trajectory: FAA measures its final endpoint, while CAA and BWT use the full stage matrix.

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 inference configuration. LLaVA Zero-shot and O-LoRA use the native Finance scoring protocol.

B.4 Complete-module removal controls

Table 5 crosses the presence and absence of language construction and complete visual composition on the six UCIT tasks. Removing L replaces the constructed language update with the fixed equal mean of the same native-scaled expert updates; removing V replaces its complete input-dependent composition with the fixed equal mean of the same projector outputs. The backbone, vision encoder and expert assets remain available in every cell. No alternative learned module is introduced in a removal cell. Each evaluated cell has a fixed language state, visual state and removal rule. The retained module uses the same frozen state in all applicable cells, isolating removal at deployment. With s_{lv} denoting FAA for cell (l, v), we report the language contribution $s_{11} - s_{01}$, the visual contribution $s_{11} - s_{10}$, and their interaction $s_{11} - s_{10} - s_{01} + s_{00}$. These contrasts measure the complete modules.

Retaining the language construction improves FAA by 14.92 points when the visual module is present. Retaining visual composition improves FAA by 9.48 points when the language module is present. Both complete components therefore contribute to the final system under the specified removal rule. The factorial interaction is -0.31 points: the two substantial module contributions are approximately additive under these removals. Appendix E reports a complementary visual-calibration comparison using a separately frozen language configuration. It addresses the effect of adding supervised visual calibration to that configuration.

The FerVA row reports the frozen full-input language construction combined with affine-trained visual calibration and nearest-simplex deployment. Its own six-task predictions yield 76.20 FAA on 3,000 examples per task. The deployed visual coefficients are nonnegative. The comparison table uses this single fixed configuration throughout.

The complete InternVL results establish final performance and retention on a second backbone; the module-removal comparisons isolate component effects on LLaVA-1.5-7B.

C Derivations and Scope

C.1 Exact reconstruction in expert-output-feature coordinates

We establish the conditions under which an expert-update column-space parameterization preserves the full reconstruction solution. Let $D_t \in \mathbb{R}^{d_{\text{out}} \times d_{\text{in}}}$ be the actual expert updates to one linear sublayer of the language backbone: it maps an input activation $x \in \mathbb{R}^{d_{\text{in}}}$ to an output feature $D_t x \in \mathbb{R}^{d_{\text{out}}}$. Thus its column space is an output-feature space and its row space is an input-activation space. Let $K_t = n_t^{-1} \sum_i X_{ti}^\top X_{ti}$ be the same example-normalized second-moment matrices as in the main text. Rows of X_{ti} are the collected input activations for example i. These activations are held constant during this solve, but are not assumed identical across examples. The matrix $Q \in \mathbb{R}^{d_{\text{out}} \times r}$ satisfies $Q^\top Q = I$ and $QQ^\top D_t = D_t$ for every expert. With the same nonnegative task weights a_t in both formulations and isotropic regularization, define

$$H = \sum_t a_t K_t + \lambda I, \qquad N_{\text{dense}} = \sum_t a_t K_t D_t^\top, \qquad N = N_{\text{dense}} Q,$$

and assume H is positive definite. The unrestricted ridge solution is $D_{\text{dense}} = (H^{-1} N_{\text{dense}})^\top$. The output-span solve gives $U_* = H^{-1} N$ and $D_{\text{span}} = Q U_*^\top$. Since $N_{\text{dense}} Q Q^\top = N_{\text{dense}}$, we have

$$D_{\text{span}} = Q Q^\top N_{\text{dense}}^\top H^{-\top} = N_{\text{dense}}^\top H^{-\top} = D_{\text{dense}}.$$

Thus the reduced representation retains the matched dense optimum. For identical expert statistics, task weights and isotropic ridge, this is the dense ridge response-reconstruction solution associated with RegMean (Jin et al., 2022). FerVA's coordinate criterion specifies when a compact solve preserves that solution and exactly what an input restriction loses. The identity concerns exact coordinates under matched statistics and regularization. Approximate bases, finite precision and deployment gains introduce the numerical effects analyzed below. The numerical output basis is obtained from the concatenated LoRA output factors by retaining singular values $s > \max(10^{-12}, 10^{-7}s_{\max})$; construction requires a nonempty retained basis. The proposition applies to exact spans, while this threshold specifies their numerical approximation.

Normalization also needs care. Multiplying all reconstruction terms by a common positive constant leaves an unregularized solution unchanged; with a fixed ridge coefficient it changes the effective regularization instead. Equal numbers of source rows per task do not make token-sum aggregation equal to a per-row token mean. We therefore hold aggregation and precision fixed in representation comparisons and evaluate source re-estimation separately.

C.2 Selecting constraints by the capacity they preserve

The preceding output-feature-span result establishes a representation benefit: if the expert output features lie in an r-dimensional span, the matched solution can be expressed with $d_{\text{in}}r$ rather than $d_{\text{in}}d_{\text{out}}$ coefficients. The compact solve and the deployed model therefore have different parameter counts, and reconstruction fidelity should be assessed separately from task accuracy.

In those output-feature coordinates, consider $J(U) = \operatorname{tr}(U^\top HU)/2 - \operatorname{tr}(U^\top N)$ with $H \succ 0$. For an orthonormal input basis P, define $U_* = H^{-1}N$ and $U_P = P(P^\top HP)^{-1}P^\top N$. Completing the square and using the restricted normal equations gives

$$J(U_P) - J(U_*) = \tfrac{1}{2}\|H^{1/2}(U_* - U_P)\|_F^2, \qquad P^\top H(U_* - U_P) = 0.$$

Thus an input restriction is lossless precisely when U_* lies in the range of P. Otherwise, removing it restores the missing component $U_* - U_P$ in the reconstruction metric. The design implication is to choose constraints by what they preserve: an output-feature span containing the expert responses preserves this optimum, whereas an input-activation span must additionally contain the fitted coefficient map to do so. Approximate output spans, different objectives and anisotropic penalties change the reconstruction problem and its preservation conditions.

Architectural scope of the argument. The restriction-cost identity holds for any compatible matrix dimensions and positive-definite reconstruction system, independently of a particular backbone's width, depth or attention layout. Applying the output-span result additionally requires its stated response-containment and penalty assumptions. The statements apply at the module level with fixed statistics. Each model supplies its own activations, expert bases and fitted weights; complete-model experiments assess the resulting inter-layer and visual interactions.

Why the experts' input-activation row space need not contain the shared solution. A two-feature example isolates the role of task-conditioned input correlations. Take expert updates $D_1 = qe_1^\top$ and $D_2 = 2qe_1^\top$, with unit q, task activation second-moment matrices and illustrative weights $a_1 = a_2 = 1$:

$$K_1 = \begin{bmatrix} 1 & 1/2 \\ 1/2 & 1 \end{bmatrix}, \qquad K_2 = \begin{bmatrix} 1 & -1/2 \\ -1/2 & 1 \end{bmatrix}, \qquad \lambda = 1.$$

Both expert-update row spaces use only input-activation direction e_1. Nevertheless, the shared ridge system has $H = K_1 + K_2 + I = 3I$ and reduced right-hand side $N = (3, -1/2)^\top$. Its solution is $U_* = (1, -1/6)^\top$, whereas restriction to $P = e_1$ gives $U_P = (1, 0)^\top$ and excess objective $1/24$. The additional e_2 dependence is an input direction required by the joint task-conditioned reconstruction problem. Conversely, if $H^{-1}N$ already lies in the input-activation span, removing that restriction gives no improvement in this objective. Both D_1 and D_2 are exactly representable in a shared two-sided expert basis, as in the representation guarantee of Panariello et al. (2026). The positive regression gap therefore concerns the search space of the subsequent second-moment-weighted solve, not loss of either expert during basis alignment.

C.3 SECOND-MOMENT COUPLING AND THE MISSING RECONSTRUCTION COMPONENT

The loss characterization can be sharpened to identify the source of the omitted component. Suppose $[P\ R]$ is an orthogonal input basis, with P spanning retained directions and R spanning omitted directions. Express each expert update as $D_t = QC_t^\top P^\top$, where Q is the shared output basis and C_t contains the expert's coefficients. Write $K_t = n_t^{-1}\sum_i X_{ti}^\top X_{ti}$, $H = \sum_t a_t K_t + \lambda I \succ 0$, and $N = \sum_t a_t K_t P C_t$. Define the retained block A, the coupling block B, and the omitted block C. The restricted solution has coefficients V, while the Schur complement S accounts for the omitted directions after the retained ones are optimized:

$$A = P^\top H P, \quad B = P^\top H R, \quad C = R^\top H R, \quad V = A^{-1}P^\top N, \quad S = C - B^\top A^{-1} B.$$

Proposition 4 (A second-moment-coupling criterion)**.** *Let P and R span the retained and omitted input directions, respectively, and write each expert update as $D_t = QC_t^\top P^\top$ for shared orthonormal output basis Q. For expert input second moments K_t and task weights a_t, define the normal matrix $H = \sum_t a_t K_t + \lambda I \succ 0$ and right-hand side $N = \sum_t a_t K_t P C_t$, with ridge coefficient $\lambda \geq 0$. Define the retained block $A = P^\top H P$, coupling block $B = P^\top H R$, Schur complement $S = R^\top H R - B^\top A^{-1} B$, and restricted coefficients $V = A^{-1}P^\top N$. For the reduced quadratic $J(U) = \frac{1}{2}\operatorname{tr}(U^\top H U) - \operatorname{tr}(U^\top N)$, write $U_* = H^{-1}N$ for the unrestricted optimum. The residual cross-input forcing $b = \sum_t a_t R^\top K_t P(C_t - V)$ then satisfies*

$$R^\top U_* = S^{-1} b, \qquad J(PV) - J(U_*) = \tfrac{1}{2}\operatorname{tr}(b^\top S^{-1} b).$$

In particular, the expert input span preserves the reconstruction optimum if and only if $b = 0$.

Proof. In the $[P\ R]$ coordinates, the full normal equations have block matrix $\left[\begin{smallmatrix} A & B \\ B^\top & C \end{smallmatrix}\right]$. Eliminating the retained coordinates gives $SR^\top U_* = R^\top N - B^\top V = b$; isotropic ridge has no cross block. The retained-coordinate difference is $-A^{-1}BS^{-1}b$. Substitution into the completed-square objective gives the stated excess loss. Since $S \succ 0$, this loss vanishes precisely when $b = 0$. □

The criterion distinguishes the experts' row-space support (over input activations) from the support needed to fit their joint output features: cross-input activation second moments act on the expert-to-shared residuals $C_t - V$.

For any common cross-block map $\mathcal{T}$, write $R^\top K_t P = \mathcal{T}(P^\top K_t P) + E_t$. The restricted normal equation yields

$$b = \sum_t a_t E_t (C_t - V) + \lambda \mathcal{T} V.$$

Thus identical conditional cross-input maps ($E_t = 0$) suffice in the unregularized positive-definite case, but not generally with ridge. This separates task-dependent correlation mismatch from regularization-induced effects. Conversely, zero cross blocks imply $b = 0$ even with different experts. Measuring the aggregate b does not by itself identify which taskwise term dominates. These fixed-statistic conditions guide input-space construction; they do not assert monotonic generated-answer performance or extend automatically to nonlinear visual mixing.

C.4 FIDELITY WITH INDIVIDUAL EXPERT SUPPORTS

To retain the rank structure of each expert, fix orthonormal input-activation and output-feature bases $P_t \in \mathbb{R}^{d \times p_t}$ and $Q_t \in \mathbb{R}^{d_{\text{out}} \times q_t}$, with $q_t \geq 1$. Let Q contain their joint output space, $R_t = Q^\top Q_t$, and write $D_t = Q_t C_t^\top P_t^\top$, where $C_t \in \mathbb{R}^{p_t \times q_t}$. With fixed positive task weights, activation second-moment matrices and $H = \sum_t a_t K_t + \lambda I \succ 0$, the free-input reconstruction has the unique solution

$$U_*(C) = H^{-1} \sum_t a_t K_t P_t C_t R_t^\top .$$

The following statement varies the coefficients within each expert's own supports, rather than allowing every expert to fill the entire joint core.

Proof of Proposition 3. In the notation of the main text, B contains every block $a_t K_t P_t$.

Proof. The displayed solution proves sufficiency. For necessity, set all but one C_t to zero and take $C_t = vw^\top$ with arbitrary v and nonzero w. Since R_t has orthonormal columns, $R_t w \neq 0$. At least one column of U_* is therefore a nonzero multiple of $H^{-1} a_t K_t P_t v$. Varying v and t proves the necessary inclusion. Invertibility of H gives the rank statement. Strict convexity makes containment equivalent to preserving the complete reconstruction optimum. □

Within this fixed-statistic family, six rank-16 input supports require at most 96 dimensions; the needed space may be *rotated* relative to the experts' joint input space rather than larger. Full-input reconstruction includes it without imposing an additional input basis. An alternative basis obtained from $H^{-1}B$, where B concatenates the statistically weighted individual input supports, would still require solving with H. This distinguishes preserving fixed expert matrices (Panariello et al., 2026) from preserving the optimum of their shared, second-moment-weighted reconstruction (Jin et al., 2022).

When both individual supports are fixed, the whitened coefficient map is $[R_t \otimes (H^{-1/2} a_t K_t P_t)]_t$. Restricting input space corresponds to a structured projection $I_r \otimes \Pi_{H^{1/2}\mathcal{L}}$. An unrestricted low-rank approximation of the vectorized map need not have this structure and cannot automatically supply an optimal input basis. The minimal-space proposition does not rely on that relaxation.

Recovering the space from cached task responses. Suppose the effective projected target matrix of task t is $T_t = P_t F_t$, with P_t an orthonormal basis of its actual column space, so F_t has full row rank. A cached task right-hand side is $N_t = a_t K_t T_t$. Then

$$\operatorname{range}(N_t) = \operatorname{range}(a_t K_t P_t), \qquad \mathcal{L}_* = \operatorname{range}\big(H^{-1}[N_1, \ldots, N_T]\big).$$

Indeed, $N_t = a_t K_t P_t F_t$ gives one inclusion, and $N_t F_t^\dagger = a_t K_t P_t$ gives the other. Thus taskwise responses can recover the required space without storing every full task activation second-moment matrix. The target support must refer to the effective represented update; an ideal adapter support or a numerically truncated basis cannot replace this condition without an error analysis. Keeping only $\sum_t N_t$ is insufficient: with $H = 2I$, the task pairs $\{(1, r)^\top, (1, -r)^\top\}$ and $\{(1, 0)^\top, (1, 0)^\top\}$ have the same sum but different joint spans for $r \neq 0$. This distinguishes cancellation in one merged solution from a space that preserves the whole response family.

Fixed statistics versus changing task weights. The preceding response-family result fixes the normal matrix. Reweighting tasks is a different intervention: for unweighted task moments K_t, M_t, write $H(a) = \lambda I + \sum_t a_t K_t$ and $N(a) = \sum_t a_t M_t$. Let $U_0 = H(a_0)^{-1} N(a_0)$, $\Delta H_v = \sum_t v_t K_t$ and $R_v = \sum_t v_t (M_t - K_t U_0)$. Whenever the perturbed system is invertible,

$$U(a_0 + \epsilon v) - U_0 = \epsilon (H(a_0) + \epsilon \Delta H_v)^{-1} R_v.$$

Subtracting $(H(a_0) + \epsilon \Delta H_v) U_0$ from the perturbed right-hand side proves the identity. Thus the local response is $H(a_0)^{-1} R_v$, not generally $H(a_0)^{-1} \sum_t v_t M_t$. An input space containing U_0 preserves every allowed first-order response exactly when it contains these derivative columns. This distinguishes response variation at fixed geometry from changing the geometry itself; the evaluated construction does not perform task-weight search. Cached task right-hand sides suffice for the preceding fixed-geometry argument, but need not determine $K_t U_0$ for this different question.

C.5 Uniform fidelity across expert responses

The instance criterion above also distinguishes an accidentally lossless expert combination from an input restriction that is justified throughout a family of expert responses. Fix $P, Q, \{K_t, a_t\}$ and $\lambda \geq 0$, with $a_t > 0$, $H \succ 0$ and at least one output coordinate. Allow the coefficient matrices C_t to vary independently while retaining these fixed statistics. Write $A_t = P^\top K_t P$, $B_t = R^\top K_t P$ and $\mathcal{T} = B^\top A^{-1}$, using the blocks defined above.

Proposition 5 (Uniform input-space fidelity)**.** *Fix expert input second moments K_t, positive task weights a_t, and a ridge coefficient $\lambda \geq 0$. Let P and R span retained and omitted input directions, and set $H = \sum_t a_t K_t + \lambda I \succ 0$, $A = P^\top H P$, and $B = P^\top H R$. For expert updates $D_t =$*

$QC_t^\top P^\top$ with a shared output basis Q, define each expert's retained second-moment block $A_t = P^\top K_t P$ and cross block $B_t = R^\top K_t P$, and let $\mathcal{T} = B^\top A^{-1}$ be the common cross-block map. The input restriction preserves the reconstruction optimum for every independent choice of expert coefficients C_t if and only if

$$B_t = \mathcal{T} A_t \quad \text{for every } t.$$

When $\lambda = 0$, this is equivalent to the existence of a common map from the retained second-moment block to each cross block. When $\lambda > 0$, it is equivalent to $B_t = 0$ for every task.

Proof. The omitted normal-equation forcing is $b = \sum_t a_t(B_t - \mathcal{T}A_t)C_t$. Since the C_t vary independently, b vanishes identically precisely when $B_t - \mathcal{T}A_t = 0$ for every t. Summing $B_t = \mathcal{T}A_t$ with weights a_t gives $B^\top = \mathcal{T}(A - \lambda I)$, whereas the definition of $\mathcal{T}$ gives $B^\top = \mathcal{T}A$. Thus $\lambda\mathcal{T} = 0$. For positive ridge, $\mathcal{T} = 0$ and all $B_t = 0$; this condition is also sufficient. For zero ridge, any common map satisfying the equations equals $B^\top A^{-1}$, proving the other equivalence without requiring each A_t to be invertible. □

The family-level condition complements the instance-specific gap. With a common map and positive ridge the forcing is $b = \lambda\mathcal{T}V$, so the excess loss is of order λ^2 when S^{-1} and V remain bounded.

An attained worst-case cost. Let $\mathcal{C}$ stack all C_t and define

$$\mathcal{L} = [a_1(B_1 - \mathcal{T}A_1) \ \cdots \ a_T(B_T - \mathcal{T}A_T)].$$

The exact instance gap and its worst case under a coefficient-norm budget are

$$G(\mathcal{C}) = \tfrac{1}{2}\|S^{-1/2}\mathcal{L}\mathcal{C}\|_F^2, \qquad \sup_{\|\mathcal{C}\|_F \le 1} G(\mathcal{C}) = \tfrac{1}{2}\|S^{-1/2}\mathcal{L}\|_{\mathrm{op}}^2.$$

A leading right singular vector, placed in one output column, attains the bound, giving an attained stress case under a unit coefficient-norm budget. This spectral characterization of the response-reconstruction restriction complements the reduced-rank constructions discussed in Mukherjee & Zhu (2011).

C.6 Residual-guided relaxation of an input restriction

The loss characterization also gives a way to judge whether a restricted construction is sufficient without first explicitly computing the full solution. Let $U_P = P(P^\top HP)^{-1}P^\top N$ and $\mathcal{R}_P = N - HU_P = H(U_* - U_P)$. Thus $\mathcal{R}_P$ is the normal-equation residual, whereas $E = U_* - U_P$ in the visual-module construction is the solution difference. For valid bounds $mI \preceq H \preceq MI$, $m > 0$,

$$\frac{\|\mathcal{R}_P\|_F^2}{2M} \le J(U_P) - J(U_*) = \tfrac{1}{2}\operatorname{tr}(\mathcal{R}_P^\top H^{-1}\mathcal{R}_P) \le \frac{\|\mathcal{R}_P\|_F^2}{2m}.$$

These follow by bounding the eigenvalues of H^{-1}. A positive ridge provides $m = \lambda$ when the activation second-moment matrices are positive semidefinite in the represented problem; rounding errors require their own numerical verification rather than assuming this is an interval certificate.

Proposition 6 (Exact gain from additional input directions)**.** *Let $H \succ 0$ be the reconstruction normal matrix, N its right-hand side, and $J(U) = \frac{1}{2}\operatorname{tr}(U^\top HU) - \operatorname{tr}(U^\top N)$ the reduced objective. For a retained orthonormal input basis P, write $U_P = P(P^\top HP)^{-1}P^\top N$ for the restricted solution and $\mathcal{R}_P = N - HU_P$ for its normal-equation residual. Let Z have independent columns orthogonal to P, and define*

$$S_Z = Z^\top HZ - Z^\top HP(P^\top HP)^{-1}P^\top HZ.$$

Reoptimizing over the extended input space yields

$$J(U_P) - J(U_{[P,Z]}) = \tfrac{1}{2}\operatorname{tr}\big[(Z^\top\mathcal{R}_P)^\top S_Z^{-1}(Z^\top\mathcal{R}_P)\big].$$

Proof. Eliminate the old coordinates in the block normal equations on $[P, Z]$. The remaining right-hand side is $Z^\top(N - HU_P)$ and its Hessian is $S_Z \succ 0$. Completing the square gives the reduction in the optimum. □

The equations motivate a residual-driven construction option: check the remaining-loss bound, add a direction with nonzero projection on $\mathcal{R}_P$, and reoptimize. Such an addition strictly decreases the fixed quadratic in exact arithmetic; adding independent directions eventually recovers the full input space. This residual subspace-enrichment option complements the evaluated full-input solve and the auxiliary decoder's full-solution spectral completion in Section 3.3 by specifying a stopping and extension criterion without requiring that full solution in advance. Actual floating-point acceptance, generated answers and the nonlinear visual component remain separate checks.

Checking the measured gap under approximate solves. For computed coefficients $\widehat{U}$ and $P\widehat{V}$, define $r_f = H\widehat{U} - N$ and $r_r = (P^\top HP)\widehat{V} - P^\top N$. Completing the square in each feasible space gives

$$J(PV_*) - J(U_*) = J(P\widehat{V}) - J(\widehat{U}) - \tfrac{1}{2}\langle r_r, (P^\top HP)^{-1} r_r\rangle + \tfrac{1}{2}\langle r_f, H^{-1} r_f\rangle.$$

In particular, the difference energy $\|P\widehat{V} - \widehat{U}\|_H^2/2$ also requires the linear correction $\langle P\widehat{V} - \widehat{U}, r_f\rangle$ to equal the computed objective difference. A source-statistic audit of all seven predefined block-zero modules finds positive corrected gaps (0.13–11.95). The sum of the absolute linear correction and the two solve-error energies is at most 2.40×10^{-4} of the reference gap; an independent higher-accuracy quadratic solve agrees to relative error below 7.42×10^{-13}. These checks quantify numerical agreement for the represented quadratic.

Stability of the restriction-loss calculation. The exact gap can also be separated from perturbations in the represented statistics. Let $\widehat{S} \succ 0$ and $\widehat{b}$ approximate the Schur complement and omitted forcing above. Suppose verified bounds satisfy

$$\|\widehat{S}^{-1/2}(S - \widehat{S})\widehat{S}^{-1/2}\|_{\mathrm{op}} \le \epsilon < 1, \qquad \|\widehat{S}^{-1/2}(b - \widehat{b})\|_F \le \eta.$$

Writing $r = \|\widehat{S}^{-1/2}\widehat{b}\|_F$, the restriction gap obeys

$$\frac{\max(r - \eta, 0)^2}{2(1+\epsilon)} \le G \le \frac{(r+\eta)^2}{2(1-\epsilon)}.$$

Indeed, the first bound implies $(1-\epsilon)\widehat{S} \preceq S \preceq (1+\epsilon)\widehat{S}$; invert this order and apply the triangle and reverse-triangle inequalities to the forcing. Thus $r > \eta$ preserves a strictly positive gap under the stated perturbations. Applying this stability result to a different calibration population requires corresponding bounds on ϵ and η.

Boundary of fixed left-weighted expert combinations. Let P have orthonormal columns spanning the joint row space of the fixed expert updates, so $D_t(I - PP^\top) = 0$. For arbitrary fixed left multipliers A_t, every update $D = \sum_t A_t D_t$ satisfies $D(I - PP^\top) = \sum_t A_t D_t (I - PP^\top) = 0$. The realizable coefficient family in Equation 24 of Evans et al. (2026), $D = \sum_{t,p} d_{tp} q_p q_p^\top D_t$, is an instance, irrespective of whether its coefficients are optimized jointly. If the q_p also lie in the output space Q, this family is contained in $\{QVP^\top\}$, itself contained in $\{QU^\top\}$. Under the same positive-definite layer-reconstruction quadratic used above, its best loss therefore exceeds the unrestricted minimum by at least the input-restriction gap. The comparison evaluates this fixed coefficient family under the shared layer-reconstruction objective.

For a strict boundary example with one output coordinate, take $D_1 = (1, 0)$, $D_2 = (-1, 0)$, equal task weights, $K_1 = \begin{bmatrix} 1 & 1/2 \\ 1/2 & 1 \end{bmatrix}$, $K_2 = \begin{bmatrix} 1 & -1/2 \\ -1/2 & 1 \end{bmatrix}$, and ridge $\lambda > 0$. Then $H = (1+\lambda)I$ and $N = (0, 1/2)^\top$. The unrestricted solution is $(0, 1/[2(1+\lambda)])^\top$, while the best solution in the expert input span is zero; the excess loss is $1/[8(1+\lambda)]$. With identical task activation second moments the opposite targets cancel and this example has zero gap. Right-side remapping or recomputing free update matrices need not preserve the input span and is outside the fixed-left-multiplier assertion. The comparison therefore isolates the input support preserved by fixed-left aggregation under the stated quadratic.

C.7 Accounting for space, solution and weight export

The earlier propositions concern an exact, unscaled quadratic. Here we identify the object changed by numerical solution and deployment gain. Write $Z = D^\top$, $N_{\mathrm{dense}} = \sum_t a_t K_t D_t^\top$, and $Z_* =$

$H^{-1}N_{\text{dense}}$. Up to a constant, the dense objective is $\mathcal{F}(Z) = \frac{1}{2}\operatorname{tr}(Z^\top HZ) - \operatorname{tr}(Z^\top N_{\text{dense}})$; hence $\mathcal{F}(Z) - \mathcal{F}(Z_*) = \frac{1}{2}\|H^{1/2}(Z - Z_*)\|_F^2$. For any orthonormal output basis Q, put $U_* = Z_*Q$. For an orthonormal input basis P, let

$$U_P = P(P^\top HP)^{-1}P^\top N_{\text{dense}}Q, \qquad \widehat{U} = U_P + E, \quad E \in \operatorname{range}(P).$$

The last condition means every column of the in-space solve error E belongs to the declared input space. The method's full-input solve uses $P = I$; the restricted construction is an analysis control.

Proposition 7 (Exact fixed-statistic error accounting)**.** *Let $\mathcal{F}$ be the fixed-statistic ridge objective with positive-definite normal matrix H and dense optimum Z_*. For an orthonormal output basis Q, set $U_* = Z_*Q$; for a retained orthonormal input basis P, let U_P be its restricted optimum and $\widehat{U} = U_P + E$ its computed solution, where every column of the in-space error E lies in* $\operatorname{range}(P)$. *For deployment gain γ, write $Z_\gamma = \gamma\widehat{U}Q^\top$. Then*

$$\begin{aligned}\mathcal{F}(Z_\gamma) - \mathcal{F}(Z_*) = \tfrac{1}{2}\|H^{1/2}Z_*(I - QQ^\top)\|_F^2 + \tfrac{1}{2}\|H^{1/2}(U_* - U_P)\|_F^2 \\ + \tfrac{1}{2}\|H^{1/2}[(\gamma - 1)U_P + \gamma E]\|_F^2. \end{aligned} \tag{9}$$

The last term jointly accounts for in-space solve error and gain; these two effects need not be independently additive.

Proof. The output-orthogonal part of $Z_\gamma - Z_*$ is $-Z_*(I - QQ^\top)$. Its remaining part is $[\gamma(U_P + E) - U_*]Q^\top$, orthogonal to the first in the H-weighted Frobenius product. The restricted normal equation gives $P^\top H(U_* - U_P) = 0$. Consequently $U_* - U_P$ is H-orthogonal to both U_P and E, proving the second separation and the identity. □

If Q exactly contains every expert update, $P = I$ and $E = 0$, then

$$\mathcal{F}(\gamma Z_*) - \mathcal{F}(Z_*) = (\gamma - 1)^2[\mathcal{F}(0) - \mathcal{F}(Z_*)].$$

For an actual reloaded update $Z_{\text{pub}} = Z_\gamma + R_{\text{pub}}$, the triangle and reverse-triangle inequalities bound the square root of twice its objective gap between

$$\max\{0, e_\gamma - r_{\text{pub}}\} \quad \text{and} \quad e_\gamma + r_{\text{pub}}, \quad e_\gamma = \|H^{1/2}(Z_\gamma - Z_*)\|_F, \quad r_{\text{pub}} = \|H^{1/2}R_{\text{pub}}\|_F.$$

These statements quantify reconstruction error under the same fixed statistics and objective at every stage.

C.8 The auxiliary rank budget and the visual transfer boundary

Write the coefficient component omitted by the input restriction as $E_P = U_* - U_P$, and its second-moment-weighted form as $M = H^{1/2}E_P$. For a rank-k correction Δ of U_P, completing the square gives

$$\min_{\operatorname{rank}(\Delta)\le k}[J(U_P + \Delta) - J(U_*)] = \tfrac{1}{2}\sum_{j>k}\sigma_j(M)^2.$$

Indeed multiplication by invertible $H^{1/2}$ preserves rank, so this is the best rank-k approximation to M. If V_k contains its leading right singular vectors, $\Delta_k = E_P V_k V_k^\top$ attains the minimum. Singular-value ties can yield several minimizers. This characterizes the auxiliary rank-16 correction under its specified quadratic.

To state the latter boundary, fix an image and answer prefix. Let v_t be the frozen projector-output tensors, p the prior simplex weights, $w = p + \delta p$ the affine training weights, and $q = \Pi_\Delta(w)$ the deployed weights. Write $V(c) = \sum_t c_t v_t$ and $D_V = \max_{j,k}\|v_j - v_k\|_F$. Since $q - w$ sums to zero, its positive and negative masses each equal $\|q - w\|_1/2$. Both normalized parts form convex combinations of the v_t, so

$$\|V(q) - V(w)\|_F \le \tfrac{1}{2}D_V\|q - w\|_1. \tag{10}$$

Let $L_F(c)$ and $L_A(c)$ be final- and auxiliary-decoder answer losses on this same prefix. If $|L_F(c) - L_A(c)| \le \delta_{\text{dec}}$ for $c \in \{p, w\}$ and L_F is L_V-Lipschitz in visual tensors along $V(w)$ to $V(q)$, inserting the intermediate losses yields

$$L_F(q) - L_F(p) \le L_A(w) - L_A(p) + 2\delta_{\text{dec}} + \tfrac{1}{2}L_V D_V\|q - w\|_1.$$

The two discrepancies describe decoder transfer and affine-to-simplex deployment separately. Neither constant is measured here, and minimizing answer CE plus reference KL does not by itself guarantee a lower CE term.

D Evaluation and Numerical Protocol

Formal UCIT evaluation uses 3,000 ordered 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 complete 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. The selected LLaVA construction uses FP32 for its core linear solves. The construction comparisons below specify the precision used for statistics, core solves, stored updates and inference separately.

Numerical conditions for the recovery comparison. The recovery family in Figure 4 has its own frozen input-restricted and full-input endpoints, 69.54 and 76.04 FAA, and uses FP32 source collection, construction and inference. Its 6.50-point contrast compares the two constructions under the same FP32 protocol. The 76.20 complete-method endpoint additionally uses the fitted visual coefficients transferred from the auxiliary decoder. Refinement coefficients are exported as absolute FP32 weights.

Full-input construction and coordinate provenance. The source-statistics collection for full-input reconstruction uses 256 examples per task and every expanded nonpadding token. It divides each task's token sum by its example count, not by the number of tokens in each example; thus variable-length examples have different total weight. Teacher forcing uses legal source labels with the frozen template. The solve is full-input-activation regression constrained to the expert output-feature column space (at most 96 dimensions), rather than projection of language-backbone input activations onto that output-feature space. The LLaVA normal system and core solve use FP32 with a fixed source-scaled ridge. The support statistics and expert basis are formed in FP64 before conversion, so this protocol uses mixed construction precision. The evaluated model is identified together with its deployment gain; scores are not transferred to models with different weights or gains.

InternVL numerical execution. The UCIT–InternVL construction uses higher-precision Gram accumulation and solves for numerically sensitive modules, then stores their factors in FP32. The reconstruction and auxiliary correction share the resulting Gram while retaining the same experts, source examples and objective. This is a mixed-precision construction and does not use evaluation labels.

Selected UCIT–LLaVA visual fit. The complete 76.20 endpoint fits 5,125 zero-initialized visual coefficients with AdamW. The batch size is six, with one example from each task per update. The fixed budget comprises 128 updates over 128 fitting examples per task, with 32 disjoint examples per task monitored separately without checkpoint selection. The final state transfers without refitting and is projected onto the simplex at deployment. Language-statistics collection uses a separate budget of 256 examples per task.

E Additional Visual Calibration Evidence

Visual calibration of a frozen reference configuration. Holding the reference language weights fixed, a calibrator driven by the frozen image encoder's classification-token (CLS) representation improves FAA on the full UCIT evaluation from 72.15 to 72.48, as shown in Table 6. This is a controlled visual-component comparison, with increases on three tasks, one tie and decreases on two tasks. Both evaluations use the same ordered 3,000 examples per task and independently checked scoring. The two rows isolate the effect of adding visual calibration to this frozen reference configuration. Paired QA predictions contain 113 corrections and 57 negative flips, yielding 56 net corrections across the four QA tasks. These account for 0.31 of the 0.32 FAA-point increase; the remaining 0.01 points come from the two corpus-level caption scores. The aggregate gain is therefore accounted for by paired QA correctness changes and corpus-level caption scores.

The source-budget arm of Figure 5 reports the four prespecified full-input fits. Their six-task final average scores range from 75.93 to 76.04, with one fitted state evaluated at each budget.

Table 6: **Frozen visual calibration improves an earlier language configuration.** These UCIT–LLaVA rows keep the earlier language weights, ordered examples, decoding, and scoring unchanged across six tasks with 3,000 examples each. IN-R, Arxiv, VizWiz, Icon, CLEVR, and Flickr denote ImageNet-R, ArxivQA, VizWiz-caption, IconQA, CLEVR-Math, and Flickr30k. The final average score (FAA) rises by 0.32 points after frozen classification-token calibration, with three task improvements, one tie, and two regressions. This is an earlier visual calibrator, not the final FerVA state.

Visual interface	IN-R	Arxiv	VizWiz	Icon	CLEVR	Flickr	FAA
Prototype-guided routing (Gao et al., 2026)	89.50	91.97	58.79	67.67	68.13	56.86	72.15
Frozen classification-token calibration	90.20	91.97	58.94	68.93	68.03	56.79	72.48

Table 7: **Direct transfer to the DCL test domains.** The selected FerVA UCIT–LLaVA model remains frozen. Each domain uses its complete native test population and original question text. Accuracy uses the native domain scorer; greedy decoding retains the source-model limit of 256 generated tokens. These are transfer accuracies, not a continual-learning trajectory or backward-transfer measure.

Test domain	Examples	Accuracy ↑
Remote sensing	10,000	30.60
Medical	9,790	35.27
Autonomous driving	10,000	13.77
Science	7,687	34.98
Finance	9,992	32.88

F Cross-Domain Transfer Without Target-Domain Training

Cross-domain transfer of the frozen model. Table 7 evaluates the UCIT-constructed model on the five DCL test domains without target-domain training or model selection. This transfer experiment is separate from the native continual-learning protocol in Table 3.

G Metric Aggregation

The final average score (FAA) gives equal weight to each final-stage task. The continual average (CAA) first averages the tasks available at each stage, then weights the stages equally. Backward transfer (BWT) averages each earlier task's change from its acquisition stage to the final stage, using the $T - 1$ earlier tasks as the denominator. Stage-wise metrics are computed from complete score matrices, not inferred from rounded final-task endpoints. Zero-shot and Individual means are reference summaries, not sequential acquisition-stage scores; Individual is additionally marked as a multi-model reference.