TRACING AND COORDINATING CROSS-LAYER INFLUENCE FOR MULTIMODAL MODEL MERGING

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal model merging consolidates task experts into a single model while preserving their complementary capabilities. Most existing approaches follow the layer-wise design of unimodal merging and combine the parameter changes learned by experts within each layer. However, merging one layer alters downstream representations and affects how visual and textual information interact. These dependencies make it difficult to trace the influence of individual experts across depth and jointly select merging weights. To tackle these challenges, we propose `TAC-Merge`, which contains two modules, i.e., multimodal influence mapping (MIM) and coupled merge control (CMC). MIM builds graphs to trace expert influence across depth and distinguish responses to image and instruction changes. The shared objective combines Ricci curvature matching on these graphs with prediction agreement to retain expert capabilities. CMC then uses the measured responses to identify a compact subspace for optimizing this objective. Within this subspace, CMC fits a local quadratic model of weight interactions and jointly selects regional weights to reduce interference. Experiments across diverse multimodal tasks demonstrate the effectiveness of `TAC-Merge` and its generalization to unseen tasks.

1 INTRODUCTION

Large pretrained models are often adapted into experts for different tasks and domains (Hu et al., 2022; Zeng et al., 2025). Storing and deploying each expert separately incurs substantial overhead (Yadav et al., 2023). Model merging offers an efficient alternative by combining expert updates into one deployable model (Wortsman et al., 2022; Ilharco et al., 2023). This is particularly valuable for multimodal large language models (MLLMs), whose experts span question answering, visual reasoning, chart understanding, and object localization (Wei et al., 2026). The objective is to consolidate these complementary capabilities in one model while preserving the interactions through which visual evidence and textual instructions jointly support task predictions.

Most model merging methods adopt a layer-wise design. Their strategies include direct parameter-space aggregation (Wortsman et al., 2022; Ilharco et al., 2023), interference-aware update transformation (Yadav et al., 2023; Yu et al., 2024; Gargiulo et al., 2025; Stoica et al., 2025), and data-dependent merge estimation (Matena & Raffel, 2022; Jin et al., 2023; Tam et al., 2024; Yang et al., 2024). Multimodal merging methods largely retain this layer-wise formulation, adapting how updates are combined to the architectural components and task requirements of multimodal models (Chen et al., 2024; Du et al., 2025; Zeng et al., 2025; Qu et al., 2025; Wei et al., 2026).

Some recent methods move beyond this layer-wise view by modeling dependencies across model depth. Chain of Merges refreshes activation statistics as layers are merged sequentially (Buzzega et al., 2026), whereas RegMean++ introduces explicit correlations between layers (Nguyen et al., 2026). These designs reflect a basic property of deep models: changing an intermediate representation can alter downstream computation (Ghandeharioun et al., 2024). Their formulations, however, do not distinguish how these effects are distributed across modalities.

This perspective presents two challenges. Firstly, *how to characterize the multimodal influence of individual expert updates across depth* (**CH1**). An expert update can alter representations far beyond the region where it is applied, making local parameter changes insufficient to characterize its influence. At later depths, effects from different regions can overlap, obscuring the contribution

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.

of each update. Multimodal fusion adds ambiguity because visual evidence and textual instructions jointly shape hidden representations. For example, a change at an image-token position may reflect altered visual processing, a different response to the instruction, or both. Neither token labels nor final predictions alone reveal these distinctions. The challenge is therefore to trace effects across depth back to individual expert updates while distinguishing how they alter image- and instruction-dependent computation within shared multimodal representations.

Secondly, *how to jointly combine expert updates based on their multimodal influence* (**CH2**). Given the influence information, fusion requires determining how much of each expert update to apply in each region. The weights assigned to visual and language updates must be chosen together to form one model. A shared objective evaluates both multimodal influence structure and agreement with expert predictions. The preferred weight for one update depends on the weights assigned to the others. The measured influence must therefore support modeling complete combinations within a coefficient neighborhood to determine the regional weights of one shared model.

To address these challenges, we propose `TAC-Merge`, a framework that traces multimodal influence across depth and uses it to coordinate expert updates. It consists of two modules, i.e., multimodal influence mapping (MIM) and coupled merge control (CMC). MIM constructs influence graphs linking regional expert updates to hidden representations and conditional image–instruction responses at different depths and token positions. It evaluates candidate merges by measuring agreement with both the Ricci curvature of task-expert graphs and expert predictions. CMC uses the measured responses to identify a compact coefficient subspace and fit a local model of interactions among coefficient adjustments. It then jointly selects regional weights within a bounded neighborhood and combines the original expert updates into one static model shared across tasks.

We summarize our contributions as follows: (1) We identify a limitation of layer-wise multimodal merging, where local parameter changes alone cannot fully capture how expert updates influence downstream representations and image–instruction interactions. (2) We propose `TAC-Merge`, combining multimodal influence mapping with coupled merge control to trace expert influence and coordinate regional update weights. (3) We evaluate `TAC-Merge` on two multimodal backbones across eight seen and four unseen tasks. `TAC-Merge` achieves the best average performance among the compared merging methods on both task groups. Its gains on unseen tasks further demonstrate that the consolidated capabilities extend beyond the datasets used for expert training.

2 Related Work

Multimodal Model Merging. Model merging consolidates task-specialized models into one shared set of parameters. On single-modality vision and language tasks, existing approaches can be grouped by how they construct the merged parameters. Direct parameter-space aggregation averages model weights or adds scaled task-specific updates to a common initialization (Wortsman et al., 2022; Ilharco et al., 2023). Interference-aware update transformation prunes or rescales update entries, resolves conflicting signs, or aligns singular subspaces before combining updates (Yadav et al., 2023; Yu et al., 2024; Gargiulo et al., 2025; Stoica et al., 2025). Data-dependent merge estimation uses parameter-importance statistics, feature-matching objectives, or calibration losses to determine parameter values and mixing coefficients (Matena & Raffel, 2022; Jin et al., 2023; Tam et al., 2024; Yang et al., 2024). Multimodal merging extends these strategies to models with modality-specific encoders and shared language components, including compositions that reuse encoders while merging language parameters (Chen et al., 2024). Architectural mappings support interpolation between heterogeneous multimodal models (Du et al., 2025), while uncertainty estimated from visual and textual inputs guides expert selection and composition order (Qu et al., 2025). Other methods normalize low-rank updates to preserve direction robustness (Zeng et al., 2025), or denoise task vectors and optimize their interactions to combine task and modality capabilities (Wei et al., 2026).

Layer and Modality Dependencies. Dependencies across depth arise because parameter updates change the intermediate representations used by subsequent layers. Sequential merging accounts for this propagation by refreshing activation statistics after preceding layers are merged (Buzzega et al., 2026). Regression-based formulations incorporate correlations within and between layers when estimating merged parameters (Nguyen et al., 2026). Related output-space formulations optimize coefficients against calibration responses and extend the procedure sequentially across layers (Evans

et al., 2026). In multimodal models, depth-dependent computation also determines where visual and textual information interact. Attention interventions localize information transfer among visual tokens, textual tokens, and answer-prediction positions (Zhang et al., 2025). Extensions to free-form generation trace these interactions during image captioning and chain-of-thought reasoning, revealing variation in visual–textual integration across architectures and tasks (Saporita et al., 2026).

3 METHOD

3.1 PROBLEM STATEMENT

Let $f(\cdot, \boldsymbol{\theta}_0)$ be a pretrained multimodal model, and let $\{\boldsymbol{\theta}_k\}_{k=1}^K$ be K expert models adapted from the same base model for different multimodal tasks. Each expert contributes a task-specific update, such as the effective weight update of a LoRA adapter:

$$\Delta\boldsymbol{\theta}^k = \boldsymbol{\theta}_k - \boldsymbol{\theta}_0, \qquad k = 1, \ldots, K, \tag{1}$$

where $\Delta\boldsymbol{\theta}^k$ is expert k's full update. Following data-dependent merging methods (Du et al., 2025; Zhang et al., 2026a; Buzzega et al., 2026), we assume access to a small calibration set $\mathcal{C} = \bigcup_{t=1}^K \mathcal{C}_t$ of multimodal inputs. Each task-specific subset is $\mathcal{C}_t = \{\boldsymbol{x}_i^t\}_{i=1}^{N_t}$, where $\boldsymbol{x}_i^t = (\boldsymbol{v}_i^t, \boldsymbol{q}_i^t)$ contains a visual input $\boldsymbol{v}_i^t$ and its textual instruction $\boldsymbol{q}_i^t$. Our objective is to select a single model from the regional weighted family:

$$\boldsymbol{\theta}(\boldsymbol{\alpha}) = \boldsymbol{\theta}_0 + \mathrm{Concat}_{\ell=0}^{L-1}\left(\sum_{k=1}^K \alpha_{\ell k}\Delta\boldsymbol{\theta}_\ell^k\right), \tag{2}$$

where L is the number of parameter regions, $\Delta\boldsymbol{\theta}_\ell^k$ is expert k's update block in region ℓ, and $\alpha_{\ell k}$ is its weight. Concatenation follows the original parameter order. The vector $\boldsymbol{\alpha}$ collects these coefficients. Given $\mathcal{C}$, we seek coefficients that preserve the complementary multimodal capabilities represented by the experts in one model shared across tasks.

3.2 FRAMEWORK OVERVIEW

Figure 1 illustrates `TAC-Merge`, which combines multimodal influence mapping (MIM) with coupled merge control (CMC). MIM constructs influence graphs from original readouts and responses to image and instruction changes, tracing regional update effects across depth and token positions. Matching task-expert curvature and predictions defines the reference objective. CMC uses measured responses to identify a compact coefficient subspace and fits a quadratic model from individual and paired coefficient variations. It jointly optimizes these coefficients around a uniform merge and accepts the proposal only if it improves the measured objective. The selected weights combine regional expert updates into one shared model with fixed inference coefficients.

3.3 MULTIMODAL INFLUENCE MAPPING

Multimodal Influence Graph. Visual influence varies across depth and token positions (Zhang et al., 2025), motivating a source-specific account of how expert updates alter image- and instruction-dependent computation. The L disjoint parameter regions are ordered along the visual and language pathways, with boundaries and readout positions fixed throughout coefficient selection. The regional coefficients in Eq. (2) are shared across inputs and collected in $\boldsymbol{\alpha} \in \mathbb{R}^{LK}$. The uniform reference $\bar{\boldsymbol{\alpha}}$ has every entry equal to $1/K$. Let $\boldsymbol{h}_r(\boldsymbol{v}, \boldsymbol{q}, \boldsymbol{\alpha})$ denote a pooled hidden readout, where r specifies a region boundary and the group of token positions pooled there. The readouts cover visual representations and language image-token and instruction-token positions carrying fused information (Saporita et al., 2026). For each task t, a fixed seeded permutation forms disjoint calibration pairs (i, j) with $i \neq j$. Both pair orders are reused throughout coefficient selection, and any unpaired input still contributes to original readouts and output agreement. For a fixed pair (i, j), suppressing task superscripts, define the image-change and instruction-change responses separately:

$$\boldsymbol{h}_{ij,r}^v(\boldsymbol{\alpha}) = \big(\boldsymbol{h}_r(\boldsymbol{v}_i, \boldsymbol{q}_i, \boldsymbol{\alpha}) - \boldsymbol{h}_r(\boldsymbol{v}_j, \boldsymbol{q}_i, \boldsymbol{\alpha})\big), \quad \boldsymbol{h}_{ij,r}^q(\boldsymbol{\alpha}) = \big(\boldsymbol{h}_r(\boldsymbol{v}_i, \boldsymbol{q}_i, \boldsymbol{\alpha}) - \boldsymbol{h}_r(\boldsymbol{v}_i, \boldsymbol{q}_j, \boldsymbol{\alpha})\big), \tag{3}$$

Here $\boldsymbol{h}_{ij,r}^v$ varies the image at a fixed instruction, whereas $\boldsymbol{h}_{ij,r}^q$ varies the instruction at a fixed image. Both comparisons hold the readout location r and coefficients $\boldsymbol{\alpha}$ fixed. We retain the original

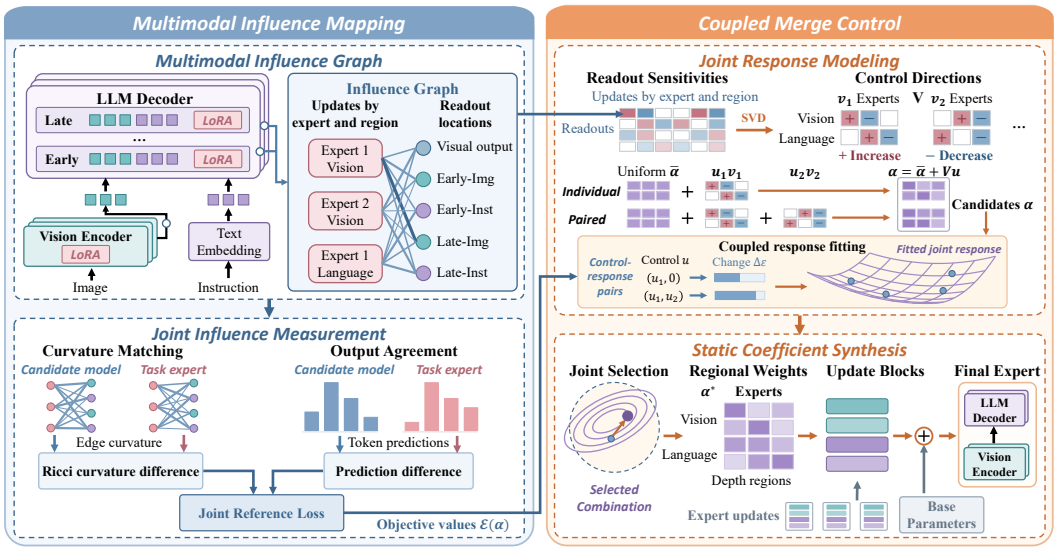

Figure 1: Overview of `TAC-Merge`. MIM maps multimodal expert influence, and CMC coordinates regional coefficients using the resulting reference objective to synthesize one shared model.

representation as $\boldsymbol{h}^0_{i,r}(\boldsymbol{\alpha}) = \boldsymbol{h}_r(\boldsymbol{v}_i, \boldsymbol{q}_i, \boldsymbol{\alpha})$ and index these three readout channels by $c \in \{0, v, q\}$. Channel 0 is always retained. Channel v or q is retained when at least one fixed pair changes that modality, with this choice fixed throughout coefficient selection. Identical instructions give $\boldsymbol{h}^q_{ij,r} = \mathbf{0}$ for every coefficient configuration, while the image-change response remains available. For each available channel, a bipartite graph links expert–region sources to observations labeled by depth and token-position group. An edge (ℓ, k, r) connects expert k in region ℓ to observation r reachable for channel c in the forward computation. The architecture and attention mask fix these connections, excluding observations structurally independent of the changed modality. Its task-specific influence measures the sensitivity of the corresponding readout channel:

$$w^{t,c}_{\ell k,r}(\boldsymbol{\alpha}) = \mathbb{E}\left[\left\|\frac{\partial \boldsymbol{h}^c_r(\boldsymbol{\alpha})}{\partial \alpha_{\ell k}}\right\|_2^2\right], \tag{4}$$

where sample indices are suppressed in $\boldsymbol{h}^c_r$, and $\|\cdot\|_2$ is the Euclidean norm. The expectation averages uniformly over all original inputs of task t for $c = 0$ and over its fixed ordered pairs for $c = v, q$. The separate graphs preserve the distinction between the three sensitivities. To characterize the allocation of influence around each node, incident weights are normalized into a probability distribution. For a source, neighbor masses describe relative influence across observations, while for an observation they describe relative sensitivity to expert–region updates. For nodes with positive total incident weight, suppressing task, channel, and coefficient arguments, we retain $m_{\ell k}(\ell, k) = m_r(r) = 1/2$ at the source and observation nodes themselves and distribute the remaining mass among their architectural neighbors:

$$m_{\ell k}(r) = \frac{w_{\ell k,r}}{2\sum_{r'} w_{\ell k,r'}}, \qquad m_r(\ell, k) = \frac{w_{\ell k,r}}{2\sum_{\ell',k'} w_{\ell' k',r}}. \tag{5}$$

The sums range over architecturally connected observations and sources, respectively. Other nodes receive zero probability mass. A node with zero incident influence places all its probability mass at itself. The equal self and neighbor masses retain endpoint identity and relative influence allocation for comparison across coefficient configurations.

Joint Influence Measurement. Joint updates change downstream computation (Buzzega et al., 2026), so selecting their coefficients requires characterizing their combined effect on multimodal influence. For edge (ℓ, k, r) in channel c, let $m_{\ell k}$ and m_r denote the endpoint neighborhood distributions defined above, with task, channel, and coefficient arguments suppressed. Distances use the fixed shortest-path metric on the channel's undirected architectural support with unit edge lengths.

Since an edge's endpoints are at unit distance, its Ollivier–Ricci curvature (Ollivier, 2009) is:

$$\kappa^{t,c}_{\ell k,r}(\boldsymbol{\alpha}) = 1 - W_1(m_{\ell k}, m_r), \tag{6}$$

where W_1 is the minimum cost of transporting one endpoint distribution to the other using this graph distance. For a fixed graph support, changes in curvature summarize changes in the relative influence patterns around each source–observation pair. The task-t expert corresponds to coefficients equal to one for expert t and zero for the other experts. We evaluate its channel graphs on $\mathcal{C}_t$ using the same fixed pairs and record $\kappa^{t,c,\star}_{\ell k,r}$ as the task-specific curvature target. For each original input, let $\boldsymbol{p}^t_i$ and $\boldsymbol{p}_i(\boldsymbol{\alpha})$ denote expert and merged next-token distributions on fixed continuation prefixes generated by expert t. Combining curvature matching with expert-output agreement gives:

$$\mathcal{E}(\boldsymbol{\alpha}) = \mathbb{E}_{t,i} D_{\mathrm{KL}}\left(\boldsymbol{p}^t_i \,\|\, \boldsymbol{p}_i(\boldsymbol{\alpha})\right) + \lambda\, \mathbb{E}_{t,c,\ell,k,r}\left[\left(\kappa^{t,c}_{\ell k,r}(\boldsymbol{\alpha}) - \kappa^{t,c,\star}_{\ell k,r}\right)^2\right], \tag{7}$$

where D_{KL} is Kullback–Leibler divergence averaged over continuation positions and $\lambda \geq 0$ weights the geometric reference. Both expectations assign equal weight to tasks. Within each task, the first weights original inputs equally, and the second weights available channels equally and architecturally valid edges equally within each channel. Crossed inputs supply response measurements without answer targets. At each queried coefficient configuration $\boldsymbol{\alpha}$, this module computes the readout derivatives and Eq. (7) on $\mathcal{C}$. Curvature matching constrains the relative organization of influence across regions and modalities, while output agreement penalizes deviations from the expert's next-token distributions on the original calibration inputs. Optimizing both terms links representation-level influence preservation to the predictive behavior needed for each task.

3.4 Coupled Merge Control

Joint Response Modeling. Selecting regional weights requires modeling how expert updates jointly affect the reference objective, since the effect of one update can depend on the others. Directly modeling these interactions over all LK coefficients is costly, so we construct a compact control subspace from measured multimodal responses and fit a local quadratic model within it. To select this subspace, we stack the readout derivatives returned by MIM at $\bar{\boldsymbol{\alpha}}$ into a response matrix, with one column per expert–region source and one row per scalar readout component across tasks, samples, channels, and readout locations. The squared stacking weights assign equal mass to tasks, then to available channels, and uniform mass to samples and readout locations within each channel. If the response matrix is zero, we retain the uniform merge. For a nonzero matrix, let $\sigma_1 \geq \sigma_2 \geq \cdots$ denote its singular values. We select the control dimension d based on the stable rank $\sum_a \sigma_a^2/\sigma_1^2$, which reflects how response energy is distributed across singular directions. The leading d right singular vectors form an orthonormal basis $\boldsymbol{V} \in \mathbb{R}^{LK\times d}$, with each column specifying a coordinated adjustment of expert weights across regions. The control vector $\boldsymbol{u} \in \mathbb{R}^d$ adjusts these directions within radius $R > 0$ of the uniform reference:

$$\boldsymbol{\alpha} = \bar{\boldsymbol{\alpha}} + \boldsymbol{V}\boldsymbol{u}, \qquad \|\boldsymbol{u}\|_2 \leq R. \tag{8}$$

Orthonormality makes R a bound on the aggregate coefficient displacement from $\bar{\boldsymbol{\alpha}}$. Within this neighborhood, a quadratic model approximates the objective change relative to $\bar{\boldsymbol{\alpha}}$:

$$\widehat{\mathcal{E}}(\boldsymbol{u}) = \boldsymbol{g}^\top \boldsymbol{u} + \tfrac{1}{2}\boldsymbol{u}^\top \boldsymbol{H}\boldsymbol{u}, \tag{9}$$

where $\boldsymbol{g} \in \mathbb{R}^d$ represents linear trends and the symmetric matrix $\boldsymbol{H} \in \mathbb{R}^{d\times d}$ represents quadratic effects within and between control directions. To fit these terms, we use MIM to evaluate Eq. (7) under a fixed set of symmetric coefficient variations. With $\boldsymbol{e}_a$ denoting a coordinate unit vector, we use signed individual variations $\boldsymbol{u} = \pm R\boldsymbol{e}_a$ for $1 \leq a \leq d$. We also use paired variations $\boldsymbol{u} = R(\pm\boldsymbol{e}_a \pm \boldsymbol{e}_b)/\sqrt{2}$ for $1 \leq a < b \leq d$, with all independent sign choices. Least-squares fitting then identifies the model coefficients:

$$(\boldsymbol{g}, \boldsymbol{H}) \in \underset{\boldsymbol{g},\, \boldsymbol{H}=\boldsymbol{H}^\top}{\arg\min}\ \mathbb{E}_{\boldsymbol{u}}\left[\left(\mathcal{E}(\bar{\boldsymbol{\alpha}} + \boldsymbol{V}\boldsymbol{u}) - \mathcal{E}(\bar{\boldsymbol{\alpha}}) - \boldsymbol{g}^\top\boldsymbol{u} - \tfrac{1}{2}\boldsymbol{u}^\top\boldsymbol{H}\boldsymbol{u}\right)^2\right], \tag{10}$$

where $\mathbb{E}_{\boldsymbol{u}}$ averages over the evaluated coefficient configurations. Individual variations identify linear trends and quadratic terms within each direction, while paired variations identify interactions between directions through the off-diagonal entries of $\boldsymbol{H}$. The interaction terms let the controller evaluate each adjustment in the context of the others, helping identify expert combinations that work together across regions to preserve multimodal influence.

Static Coefficient Synthesis. Because the fitted model is only a local approximation, a predicted improvement may not reduce the measured fusion objective. We therefore limit coefficient changes to the prescribed neighborhood and check the resulting proposal against the uniform reference. We first obtain a joint coefficient proposal by minimizing the fitted objective:

$$\boldsymbol{u}^\star \in \arg\min_{\|\boldsymbol{u}\|_2 \le R} \widehat{\mathcal{E}}(\boldsymbol{u}), \tag{11}$$

The reference $\boldsymbol{u} = \mathbf{0}$ is feasible and predicts zero objective change. We solve Eq. (11) using an eigendecomposition of $\boldsymbol{H}$ and the spectral trust-region conditions (Moré & Sorensen, 1983):

$$(\boldsymbol{H} + \mu \boldsymbol{I}_d)\boldsymbol{u}^\star = -\boldsymbol{g}, \qquad \boldsymbol{H} + \mu \boldsymbol{I}_d \succeq \mathbf{0}, \tag{12}$$

where $\mu \ge 0$ is a spectral shift satisfying $\mu(\|\boldsymbol{u}^\star\|_2 - R) = 0$ and $\succeq \mathbf{0}$ denotes positive semidefiniteness. An interior solution has $\mu = 0$, while a boundary solution uses a shift consistent with $\|\boldsymbol{u}^\star\|_2 = R$. In the eigenbasis, the boundary search reduces to a scalar norm condition, with a nullspace component included when required by a singular boundary solution. Together with feasibility, Eq. (12) characterizes a global minimizer of the fixed quadratic model over the prescribed ball. To verify the predicted improvement, we evaluate the proposal using Eq. (7) on the same calibration data and expert targets. The final coefficients are $\boldsymbol{\alpha}^\star = \bar{\boldsymbol{\alpha}} + \boldsymbol{V}\boldsymbol{u}^\star$ if $\mathcal{E}(\bar{\boldsymbol{\alpha}} + \boldsymbol{V}\boldsymbol{u}^\star) < \mathcal{E}(\bar{\boldsymbol{\alpha}}) - \tau$, and $\boldsymbol{\alpha}^\star = \bar{\boldsymbol{\alpha}}$ otherwise, where $\tau \ge 0$ is a fixed numerical comparison tolerance. This check requires no iterative refitting of the response model. The selected coefficients $\boldsymbol{\alpha}^\star$ synthesize the shared model:

$$\boldsymbol{\theta}^\star = \boldsymbol{\theta}_0 + \mathrm{Concat}_{\ell=0}^{L-1}\left(\sum_{k=1}^{K} \alpha_{\ell k}^\star \Delta\boldsymbol{\theta}_\ell^k\right), \tag{13}$$

The regional blocks are concatenated in the original parameter order to obtain $f(\cdot, \boldsymbol{\theta}^\star)$. By selecting regional coefficients jointly, CMC accounts for how updates in one part of the model affect the inputs seen by later updates. This coordination helps reduce interference across layers and modalities and retain the complementary capabilities of different experts.

3.5 Theoretical Analysis

We prove that coupled control can achieve gains unavailable to single-direction adjustments.

Theorem 1 (Coupling advantage)**.** *Under the conditions and notation of Appendix A.1, CMC attains a lower objective than any adjustment along one response basis vector within radius R, with*

$$\mathcal{E}(\bar{\boldsymbol{\alpha}}) - \mathcal{E}(\boldsymbol{\alpha}^\star) \ge \frac{\gamma R^2}{2} - \frac{2d + \sqrt{d} + 1}{6} M R^3. \tag{14}$$

The proof is provided in Appendix A.5.

4 Experiments

4.1 Experimental Setups

Datasets. Following RobustMerge (Zeng et al., 2025), we adopt MM-MergeBench with eight seen datasets: ScienceQA (Lu et al., 2022), ImageNet (Deng et al., 2009), VQAv2 (Goyal et al., 2017), REC-COCO (Kazemzadeh et al., 2014; Mao et al., 2016), OCR-VQA (Mishra et al., 2019), VizWiz-Caption (Gurari et al., 2020), Flickr30k (Plummer et al., 2017), and IconQA (Lu et al., 2021). These datasets cover visual question answering, referring expression comprehension, image classification, and image captioning. Generalization is evaluated on four unseen datasets: A-OKVQA (Schwenk et al., 2022), ImageNet-R (Hendrycks et al., 2021), Screen2Words (Wang et al., 2021), and TabMWP (Lu et al., 2023). The seen-task evaluation measures how well one merged model retains the capabilities of the eight adapted experts. Unseen tasks evaluate whether the same coefficients transfer to datasets outside expert adaptation, testing whether the consolidated capabilities remain useful under a change in task distribution.

Table 1: Performance comparison on Qwen3.5-4B.

	Seen tasks									Unseen tasks				
Method	SciQA	Image	VQA	REC	OCR	VizWiz	Flickr	IconQA	Avg.	AVQA	Image-R	S2W	TabMWP	Avg.
Zero-shot	67.13	38.81	72.00	76.15	70.58	23.74	25.65	25.86	49.99	78.69	87.27	5.35	64.39	58.93
Individual	97.92	97.47	74.85	91.53	71.51	72.80	60.44	99.15	83.21	–	–	–	–	–
TA	94.99	59.03	72.52	73.48	68.78	52.21	58.11	79.79	69.86	86.20	90.95	15.95	69.95	65.76
TIES	94.15	56.36	72.71	74.22	69.19	52.42	57.80	70.39	68.41	83.67	**90.97**	15.58	69.57	64.95
DC-Merge	95.29	59.03	72.52	74.55	68.46	52.42	57.95	79.79	70.00	86.20	90.10	15.59	69.95	65.46
CoM	95.29	63.21	72.54	74.62	68.46	52.59	58.04	79.25	70.50	84.72	90.10	15.59	71.82	65.56
RegMean++	95.84	85.86	69.92	75.55	67.41	53.16	57.95	94.85	75.07	87.25	85.83	15.19	70.78	64.76
RobustMerge	96.03	59.03	72.15	74.55	66.75	50.76	57.59	92.15	71.13	87.07	88.83	15.75	67.85	64.88
OptMerge	93.80	**88.46**	72.52	74.55	68.46	48.76	56.50	79.79	72.86	86.20	90.10	14.92	69.95	65.29
TAC-Merge	**97.12**	88.08	**73.86**	**83.42**	**69.93**	**58.68**	**59.84**	**97.07**	**78.50**	**88.47**	90.57	**18.72**	**74.64**	**68.10**

Table 2: Performance comparison on Qwen3-VL-8B.

	Seen tasks									Unseen tasks				
Method	SciQA	Image	VQA	REC	OCR	VizWiz	Flickr	IconQA	Avg.	AVQA	Image-R	S2W	TabMWP	Avg.
Zero-shot	94.40	35.03	72.34	77.32	69.61	27.26	28.66	44.20	56.10	79.65	90.48	5.87	60.40	59.10
Individual	95.79	94.69	74.45	89.33	74.32	69.68	59.54	90.94	81.09	–	–	–	–	–
TA	95.19	58.77	72.87	77.08	72.43	48.78	56.86	72.29	69.28	87.16	**90.67**	9.41	61.67	62.23
TIES	96.13	54.77	70.65	84.07	70.24	51.12	58.18	82.49	70.96	86.72	89.60	14.04	63.14	63.38
DC-Merge	**96.58**	72.08	70.52	86.55	73.25	52.76	57.67	86.94	74.54	87.34	89.95	14.41	66.56	64.57
CoM	95.98	46.85	72.67	74.13	72.59	51.38	57.89	83.40	69.36	87.51	90.07	13.03	64.10	63.68
RegMean++	96.43	77.49	73.03	83.92	73.27	52.33	57.77	86.45	75.09	87.95	90.33	12.35	67.55	64.55
RobustMerge	95.14	58.77	72.68	69.90	71.61	49.83	58.17	80.78	69.61	87.95	90.22	14.35	64.05	64.14
OptMerge	96.23	58.77	70.76	84.15	70.77	50.25	57.78	82.16	71.36	87.60	89.75	13.33	61.75	63.11
TAC-Merge	96.38	**82.63**	**73.58**	**87.42**	**73.46**	**56.83**	**59.37**	**90.33**	**77.50**	**88.65**	90.53	**16.84**	**69.98**	**66.50**

Baselines. To evaluate the effectiveness of TAC-Merge, we compare it with the following representative baselines. We first consider two reference settings: (1) Zero-shot, which directly evaluates the pretrained model, and (2) Individual, which evaluates each task expert on its corresponding task. For single-modality merging, we include (3) TA (Ilharco et al., 2023), which adds the summed expert updates to the base model with a shared scaling coefficient, (4) TIES (Yadav et al., 2023), which trims small updates, resolves sign conflicts, and averages the remaining updates with consistent signs, and (5) DC-Merge (Zhang et al., 2026b), which smooths singular values and aligns task vectors in a shared orthogonal subspace before aggregation. For cross-layer merging, we compare (6) Chain of Merges (Buzzega et al., 2026), which merges layers sequentially while refreshing activation statistics, and (7) RegMean++ (Nguyen et al., 2026), which incorporates intra-layer and cross-layer dependencies into regression-based merging. For multimodal merging, we include (8) RobustMerge (Zeng et al., 2025), which prunes and rescales low-rank parameters and normalizes task contributions to preserve direction robustness, and (9) OptMerge (Wei et al., 2026), which denoises task vectors and optimizes the merged update through an objective over their interactions.

Implementation Details. We use Qwen3.5-4B (Qwen Team, 2026) and Qwen3-VL-8B-Instruct (Bai et al., 2025) as the base models and train one LoRA expert for each of the eight seen tasks using Unsloth (Han et al., 2023). LoRA adapters are inserted into the attention and MLP modules of both the vision encoder and the language backbone, and the two parts are optimized jointly on each task's image–instruction examples while the pretrained weights remain frozen. The LoRA rank and scaling factor are both 16. Experts train for one epoch with AdamW, learning rate 2×10^{-4}, linear decay, warmup ratio 0.03, weight decay 0.001, and a 2,048-token sequence limit. Further implementation details are provided in Appendix C.

Table 3: Ablation study.

	Qwen3.5-4B		Qwen3-VL-8B	
Variant	Seen	Unseen	Seen	Unseen
w/o MIM	76.12 ± 0.39	67.04 ± 0.27	76.03 ± 0.43	65.88 ± 0.30
MIM: w/o modality contrasts	77.46 ± 0.26	67.87 ± 0.21	77.09 ± 0.29	66.29 ± 0.24
w/o CMC	77.03 ± 0.46	67.58 ± 0.32	76.64 ± 0.48	66.26 ± 0.34
CMC: w/o directional coupling	77.79 ± 0.30	67.71 ± 0.20	77.02 ± 0.36	66.38 ± 0.19
`TAC-Merge` (Full)	$\mathbf{78.50 \pm 0.24}$	$\mathbf{68.10 \pm 0.18}$	$\mathbf{77.50 \pm 0.31}$	$\mathbf{66.50 \pm 0.22}$

4.2 Performance Evaluation

Main Results. Tables 1 and 2 compare performance on seen and unseen tasks using Qwen3.5-4B and Qwen3-VL-8B, respectively. The results yield four main observations. (1) Traditional single-modality merging methods struggle to balance heterogeneous multimodal capabilities. Task Arithmetic, TIES, and DC-Merge determine update compatibility through parameter aggregation, sign consistency, or subspace alignment, leaving the role of each update in visual–language computation implicit. Parameter-compatible updates can still alter the representations through which visual and textual information interact. The resulting trade-offs are evident on the 8B model, where TIES gains 6.99 points on REC-COCO over Task Arithmetic while losing 4.00 points on ImageNet and 2.22 points on VQAv2. (2) Existing multimodal merging methods remain limited in preserving the propagation of expert capabilities across depth. RobustMerge stabilizes low-rank update directions, and OptMerge optimizes interactions among task vectors within individual layers. These parameter-space objectives provide indirect control over how a visual update changes the input encountered by later language updates. Their lower aggregate performance is consistent with this gap between local update compatibility and preservation of multimodal influence throughout the model. (3) Cross-layer methods improve knowledge consolidation, with remaining limitations in multimodal transfer. Chain of Merges refreshes activation statistics sequentially, and RegMean++ models correlations across layers, accounting for representation changes induced by preceding updates. Their objectives emphasize activation agreement and cross-layer correlations, leaving source-specific image–instruction effects implicit. On the 4B model, RegMean++ improves the seen average over TIES by 6.66 points, while its unseen average remains 1.00 point below Task Arithmetic. This uneven transfer supports extending cross-layer dependency modeling with explicit multimodal influence measurements when jointly selecting expert coefficients. (4) `TAC-Merge` achieves the best average performance on both seen and unseen tasks across the two backbones. On Qwen3.5-4B, it exceeds the strongest merging baselines by 3.43 points on seen tasks and 2.34 points on unseen tasks. The corresponding gains on Qwen3-VL-8B are 2.41 and 1.93 points, respectively. It also leads the merging methods on seven of eight seen tasks and three of four unseen tasks for each backbone. These results support jointly selecting regional coefficients based on multimodal influence to preserve complementary expert capabilities in a shared model.

Ablation Study. Table 3 reports module ablations as mean $\pm$ standard deviation. All variants share expert checkpoints, calibration inputs, coefficient radius, and objective-evaluation budget, with matched control dimensions whenever a basis is used. (1) Effect of MIM. The variant without MIM uses output KL alone and a fixed random orthonormal basis for CMC, omitting influence graphs and readout responses. Removing modality contrasts retains the original readouts $\boldsymbol{h}^0$ and their influence graphs but excludes image-change and instruction-change channels from curvature matching and response-based subspace selection. Removing MIM reduces seen averages by 2.38 and 1.47 points on 4B and 8B, compared with 1.04 and 0.41 points when removing only the contrasts. These results support modality-dependent responses for influence matching and coefficient-direction selection. (2) Effect of CMC. Removing CMC retains the full reference objective and replaces subspace selection and quadratic control with fixed-seed search over regional coefficients sampled uniformly within the same radius, selecting the best candidate or uniform reference. Removing directional coupling instead zeros only the off-diagonal entries of $\boldsymbol{H}$, preserving the response basis, fitted linear and diagonal terms, calibration evaluations, and proposal check. The

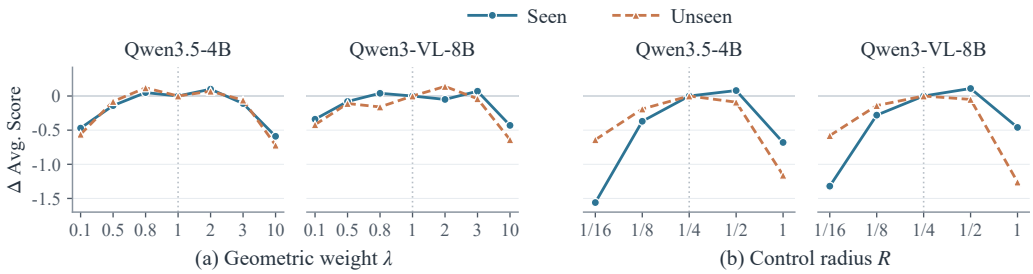

Figure 2: Sensitivity to (a) λ and (b) R. Δ is the average-score change from $(\lambda, R) = (1, 0.25)$ in score points. Settings are equally spaced, and dotted lines mark the default values.

unseen averages decrease by 0.52 and 0.24 points after removing CMC, compared with 0.39 and 0.12 points after removing only coupling. Each control direction changes multiple regional weights, so the mixed terms capture how coordinated adjustments affect one another. These results support retaining such interactions for transfer to unseen tasks. (3) Overall comparison. The full configuration achieves the highest mean in all four evaluation settings. Removing CMC also lowers seen averages by 1.47 and 0.86 points on 4B and 8B, supporting coefficient control under a fixed evaluation budget. The modules serve complementary roles: MIM specifies the multimodal influence to preserve, while CMC selects coordinated regional weights using that objective. Further component analyses appear in Appendices D.1 and D.2.

Hyperparameter Sensitivity. To examine the sensitivity of `TAC-Merge`, Figure 2 compares changes in seen and unseen average scores relative to the shared default $(\lambda, R) = (1, 0.25)$. The scans vary $\lambda \in \{0.1, 0.5, 0.8, 1, 2, 3, 10\}$ and $R \in \{1/16, 1/8, 1/4, 1/2, 1\}$ separately, holding the other parameter fixed. Each curve uses its corresponding default mean from the main tables as zero, allowing seen and unseen sensitivity to be compared on the same scale. The sensitivity curves show a broad intermediate plateau for λ, with small local fluctuations. The shared default lies within this relatively stable range for both backbones. Smaller weights place greater relative emphasis on expert-output agreement, while larger weights emphasize the relative influence patterns of the experts. The response subspace remains unchanged throughout the λ scan. Since the geometric targets are constructed from seen tasks, preserving their influence patterns and transferring to unseen inputs can favor different weights. For R, smaller neighborhoods restrict coordinated coefficient adjustments, whereas larger neighborhoods expand both the feasible displacement and the configurations used to fit the response model. The large-radius reference points represent accepted proposals with a greater decline on unseen tasks, illustrating a possible difference between calibration-objective improvement and transfer to new task distributions. The acceptance check can also produce plateaus when multiple proposals return the uniform merge.

5 Conclusion

This work identifies a central challenge in multimodal model merging: the effects of expert updates propagate across depth and alter how visual evidence and textual instructions interact. Local parameter compatibility alone therefore cannot fully characterize how updates from different experts should be combined across the visual and language pathways. We propose `TAC-Merge` to address this challenge through multimodal influence mapping and coupled merge control. The former traces regional expert influence through representations and responses to image and instruction changes, using influence-graph curvature and expert prediction agreement to guide merging. The latter models interactions among coefficient adjustments in a compact response subspace and jointly selects regional weights for a single shared model. Evaluations on two backbones across eight seen and four unseen tasks show that `TAC-Merge` achieves the best average performance in both task groups among the compared merging methods, while retaining the base inference architecture. Future work will focus on reducing the cost of influence estimation and extending coordinated merging to larger collections of task experts.

AI USE STATEMENT

We used generative AI tools for code development, writing refinement, and table and figure formatting.

ETHICS STATEMENT

We do not foresee specific ethical concerns arising from this work. The research conducted in this paper conforms to the ICLR Code of Ethics.

REPRODUCIBILITY STATEMENT

The experimental settings and implementation details are provided in Section 4.1 and Appendix C. We will release the code upon acceptance to facilitate reproducibility.

REFERENCES

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025. URL https://arxiv.org/abs/2511.21631.

Pietro Buzzega, Riccardo Salami, Angelo Porrello, and Simone Calderara. Rethinking layer-wise model merging through chain of merges, 2026. URL https://arxiv.org/abs/2508.21421.

Chi Chen, Yiyang Du, Zheng Fang, Ziyue Wang, Fuwen Luo, Peng Li, Ming Yan, Ji Zhang, Fei Huang, Maosong Sun, and Yang Liu. Model composition for multimodal large language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 11246–11262, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.606. URL https://aclanthology.org/2024.acl-long.606/.

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

Yiyang Du, Xiaochen Wang, Chi Chen, Jiabo Ye, Yiru Wang, Peng Li, Ming Yan, Ji Zhang, Fei Huang, Zhifang Sui, Maosong Sun, and Yang Liu. Adamms: Model merging for heterogeneous multimodal large language models with unsupervised coefficient optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9413–9422, June 2025.

Bethan Evans, Benjamin Etheridge, Stephen Roberts, and Jared Tanner. Model merging by output-space projection. arXiv preprint arXiv:2605.29101, 2026. URL https://arxiv.org/abs/2605.29101.

Antonio Andrea Gargiulo, Donato Crisostomi, Maria Sofia Bucarelli, Simone Scardapane, Fabrizio Silvestri, and Emanuele Rodolà. Task singular vectors: Reducing task interference in model merging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18695–18705, June 2025.

Asma Ghandeharioun, Avi Caciularu, Adam Pearce, Lucas Dixon, and Mor Geva. Patchscopes: A unifying framework for inspecting hidden representations of language models. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 15466–15490. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/ghandeharioun24a.html.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6904–6913, 2017. URL https://arxiv.org/abs/1612.00837.

Danna Gurari, Yinan Zhao, Meng Zhang, and Nilavra Bhattacharya. Captioning images taken by people who are blind. In *European Conference on Computer Vision (ECCV)*, 2020. URL https://arxiv.org/abs/2002.08565.

Daniel Han, Michael Han, and Unsloth team. Unsloth, 2023. URL https://github.com/unslothai/unsloth.

Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, Dawn Song, Jacob Steinhardt, and Justin Gilmer. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8340–8349, 2021. URL https://arxiv.org/abs/2006.16241.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. Editing models with task arithmetic. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=6t0Kwf8-jrj.

Xisen Jin, Xiang Ren, Daniel Preotiuc-Pietro, and Pengxiang Cheng. Dataless knowledge fusion by merging weights of language models. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=FCnohuR6AnM.

Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. ReferItGame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 787–798. Association for Computational Linguistics, 2014. doi: 10.3115/v1/D14-1086. URL https://aclanthology.org/D14-1086/.

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. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, 2021. URL https://iconqa.github.io/.

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *Advances in Neural Information Processing Systems*, volume 35, pp. 2507–2521, 2022. URL https://arxiv.org/abs/2209.09513.

Pan Lu, Liang Qiu, Kai-Wei Chang, Ying Nian Wu, Song-Chun Zhu, Tanmay Rajpurohit, Peter Clark, and Ashwin Kalyan. Dynamic prompt learning via policy gradient for semi-structured mathematical reasoning. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2209.14610.

Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L. Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11–20, 2016. URL https://arxiv.org/abs/1511.02283.

Michael S Matena and Colin Raffel. Merging models with fisher-weighted averaging. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 17703–17716. Curran Associates, Inc., 2022. doi: 10.52202/068431-1287. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/70c26937fbf3d4600b69a129031b66ec-Paper-Conference.pdf.

Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. OCR-VQA: Visual question answering by reading text in images. In *International Conference on Document Analysis and Recognition (ICDAR)*, 2019. URL https://ocr-vqa.github.io/.

Jorge J. Moré and D. C. Sorensen. Computing a trust region step. *SIAM Journal on Scientific and Statistical Computing*, 4(3):553–572, 1983. doi: 10.1137/0904038. URL https://doi.org/10.1137/0904038.

The-Hai Nguyen, Dang Huu-Tien, Takeshi Suzuki, and Le-Minh Nguyen. Regmean++: Enhancing effectiveness and generalization of regression mean for model merging. *Transactions on Machine Learning Research*, 2026. ISSN 2835-8856. URL https://openreview.net/forum?id=H5lDsSCS9i.

Yann Ollivier. Ricci curvature of Markov chains on metric spaces. *Journal of Functional Analysis*, 256(3):810–864, 2009. doi: 10.1016/j.jfa.2008.11.001. URL https://doi.org/10.1016/j.jfa.2008.11.001.

Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. *International Journal of Computer Vision*, 123(1):74–93, 2017. doi: 10.1007/s11263-016-0965-7.

Huaizhi Qu, Xinyu Zhao, Jie Peng, Kwonjoon Lee, Behzad Dariush, and Tianlong Chen. UQ-Merge: Uncertainty guided multimodal large language model merging. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 1401–1417, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.73. URL https://aclanthology.org/2025.findings-acl.73/.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL https://qwen.ai/blog?id=qwen3.5.

Alessia Saporita, Vittorio Pipoli, Federico Bolelli, Lorenzo Baraldi, Andrea Acquaviva, and Elisa Ficarra. Fg-tracer: Tracing information flow in multimodal large language models in free-form generation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 7903–7912, March 2026.

Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. A-OKVQA: A benchmark for visual question answering using world knowledge. In *European Conference on Computer Vision (ECCV)*, 2022. URL https://arxiv.org/abs/2206.01718.

George Stoica, Pratik Ramesh, Boglarka Ecsedi, Leshem Choshen, and Judy Hoffman. Model merging with svd to tie the knots. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 4501–4519, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/0d4f8a5109c5083b5307fcd0bddae7a7-Paper-Conference.pdf.

Derek Tam, Mohit Bansal, and Colin Raffel. Merging by matching models in task parameter subspaces. *Transactions on Machine Learning Research*, 2024. URL https://arxiv.org/abs/2312.04339.

Bryan Wang, Gang Li, Xin Zhou, Zhourong Chen, Tovi Grossman, and Yang Li. Screen2Words: Automatic mobile UI summarization with multimodal learning. In *Proceedings of the 34th Annual ACM Symposium on User Interface Software and Technology*, pp. 498–510, 2021. doi: 10.1145/3472749.3474765. URL https://arxiv.org/abs/2108.03353.

Yongxian Wei, Runxi Cheng, Weike Jin, Enneng Yang, Li Shen, LU HOU, SiNan Du, Chun Yuan, Xiaochun Cao, and Dacheng Tao. Optmerge: Unifying multimodal llm capabilities and modalities via model merging. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 61571–61595, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/64741c1e80cc3bbed205b1c2b040dfa8-Paper-Conference.pdf.

Mitchell Wortsman, Gabriel Ilharco, Samir Ya Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, and Ludwig Schmidt. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 23965–23998. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/wortsman22a.html.

Prateek Yadav, Derek Tam, Leshem Choshen, Colin Raffel, and Mohit Bansal. Ties-merging: Resolving interference when merging models. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 7093–7115. Curran Associates, Inc., 2023. doi: 10.52202/075280-0310. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/1644c9af28ab7916874f6fd6228a9bcf-Paper-Conference.pdf.

Enneng Yang, Zhenyi Wang, Li Shen, Shiwei Liu, Guibing Guo, Xingwei Wang, and Dacheng Tao. Adamerging: Adaptive model merging for multi-task learning. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 22743–22763, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/62868cc2fc1eb5cdf321d05b4b88510c-Paper-Conference.pdf.

Le Yu, Bowen Yu, Haiyang Yu, Fei Huang, and Yongbin Li. Language models are super mario: Absorbing abilities from homologous models as a free lunch. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 57755–57775. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/yu24p.html.

Fanhu Zeng, Haiyang Guo, Fei Zhu, Li Shen, and Hao Tang. Robustmerge: Parameter-efficient model merging for mllms with direction robustness. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 71071–71095. Curran Associates, Inc., 2025. doi: 10.52202/085713-2390. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/67101f97dc23fcc10346091181fff6cb-Paper-Conference.pdf.

Dengming Zhang, Xiaowen Ma, Zhenliang Ni, Zhenkai Wu, Han Shu, Xin Jiang, and Xinghao Chen. Expert merging: Model merging with unsupervised expert alignment and importance-guided layer chunking. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 21780–21803, 2026a. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/24dbe8babeca9f372712bbf19dd3584e-Paper-Conference.pdf.

Han-Chen Zhang, Zi-Hao Zhou, Mao-Lin Luo, Shimin Di, Min-Ling Zhang, and Tong Wei. DC-Merge: Improving model merging with directional consistency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22248–22258, June 2026b. URL https://openaccess.thecvf.com/content/CVPR2026/html/Zhang_DC-Merge_Improving_Model_Merging_with_Directional_Consistency_CVPR_2026_paper.html.

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

A Theoretical Analysis

A.1 Notation and Conditions

Fix the calibration inputs, pairs, graph supports, expert targets, and response basis $\boldsymbol{V}$. Write $F(\boldsymbol{u}) = \mathcal{E}(\bar{\boldsymbol{\alpha}} + \boldsymbol{V}\boldsymbol{u}) - \mathcal{E}(\bar{\boldsymbol{\alpha}})$ and $\boldsymbol{B} = \nabla^2 F(\mathbf{0})$. Assume $d \geq 2$, $R > 0$, $\beta = \min_a B_{aa} > 0$, and $\gamma = -\lambda_{\min}(\boldsymbol{B}) > 0$. On a neighborhood of the ball $\|\boldsymbol{u}\|_2 \leq R$, let $F \in C^2$ with

$$\|\nabla^2 F(\boldsymbol{u}) - \nabla^2 F(\boldsymbol{u}')\|_2 \leq M\|\boldsymbol{u} - \boldsymbol{u}'\|_2. \tag{15}$$

This holds for all $\boldsymbol{u}, \boldsymbol{u}'$ in the ball, with $M \geq 0$ and matrix $\|\cdot\|_2$ denoting the operator norm. Use exact probe evaluations in Eq. (10) and a global minimizer of Eq. (11). Define $\Delta_R = \gamma R^2/2 - (2d + \sqrt{d} + 1)MR^3/6$, and assume $MR < 3\beta$ and

$$\Delta_R > \max\left\{\tau, \frac{\|\nabla F(\mathbf{0})\|_\infty^2}{2(\beta - MR/3)}\right\}. \tag{16}$$

Connection to the merging objective. For the curvature penalty, suppressing graph indices and evaluating at $\boldsymbol{u} = \mathbf{0}$,

$$\nabla_{\boldsymbol{u}}^2\left[\lambda\mathbb{E}(\kappa - \kappa^\star)^2\right] = 2\lambda\mathbb{E}\left[(\nabla_{\boldsymbol{u}}\kappa)(\nabla_{\boldsymbol{u}}\kappa)^\top + (\kappa - \kappa^\star)\nabla_{\boldsymbol{u}}^2\kappa\right]. \tag{17}$$

The residual-weighted term can be indefinite. Positive diagonal entries with $B_{ab}^2 > B_{aa}B_{bb}$ give a negative determinant for the corresponding principal submatrix, hence $\lambda_{\min}(\boldsymbol{B}) < 0$. The response SVD alone does not guarantee this condition. The result concerns the fixed calibration objective within the selected response subspace.

A.2 Regularity of the Reference Objective

Proposition A.1 (Local regularity)**.** *On an open neighborhood of the coefficient ball, suppose the readouts are C^4, the prediction logits are C^3, and each node's total incident influence is identically zero or bounded away from zero. If each edgewise transport problem admits a fixed optimal Kantorovich potential, then $F \in C^3$ and Eq. (15) holds for some finite M.*

Proof. Equation (4) gives $w_{\ell k,r} \in C^3$. The nonvanishing denominators in Eq. (5) give $m_{\ell k}, m_r \in C^3$, while zero-influence nodes contribute constant point masses. For a fixed optimal potential $\varphi^\star$, transport duality (Ollivier, 2009) gives $W_1(m_{\ell k}, m_r) = \sum_z \varphi^\star(z)(m_{\ell k}(z) - m_r(z))$, so the curvature penalty is C^3. The C^3 logits and positive softmax probabilities likewise make the prediction loss C^3, hence $F \in C^3$. Compactness gives $M = \sup_{\|\boldsymbol{u}\|_2 \leq R} \|\nabla^3 F(\boldsymbol{u})\|_{\mathrm{op}} < \infty$, and

$$\begin{aligned}\|\nabla^2 F(\boldsymbol{u}) - \nabla^2 F(\boldsymbol{u}')\|_2 &= \left\| \int_0^1 \nabla^3 F\big(\boldsymbol{u}' + t(\boldsymbol{u} - \boldsymbol{u}')\big)[\boldsymbol{u} - \boldsymbol{u}']\, dt \right\|_2 \\ &\leq \int_0^1 M\|\boldsymbol{u} - \boldsymbol{u}'\|_2\, dt = M\|\boldsymbol{u} - \boldsymbol{u}'\|_2.\end{aligned} \tag{18}$$

□

For $\lambda = 0$, logit regularity suffices. Changes of optimal transport dual solution or zero-influence status require separate verification of Eq. (15).

A.3 Accuracy of the Symmetric Probe Fit

The uniformly weighted probe set is

$$\mathcal{S}_R = \{\pm R\boldsymbol{e}_a : 1 \leq a \leq d\} \cup \left\{\frac{R}{\sqrt{2}}(s\boldsymbol{e}_a + t\boldsymbol{e}_b) : a < b,\ s, t \in \{-1, 1\}\right\}. \tag{19}$$

It contains $2d^2$ probes, and the constant term is fixed by $F(\mathbf{0}) = 0$.

Proposition A.2 (Identification and approximation)**.** *For $d \geq 2$, the fit in Eq. (10) is unique. Under Eq. (15), it satisfies $\|\boldsymbol{g} - \nabla F(\mathbf{0})\|_2 \leq \sqrt{d}\, MR^2/6$ and $\|\boldsymbol{H} - \boldsymbol{B}\|_2 \leq dMR/3$. The recovery is exact for quadratic F.*

Proof. **(Identification.)** Sign symmetry gives $\sum_{\boldsymbol{u}\in\mathcal{S}_R} \boldsymbol{u}\boldsymbol{u}^\top = 2dR^2\boldsymbol{I}_d$ and makes the linear and quadratic evaluation spaces orthogonal. For symmetric $\boldsymbol{A}$,

$$\begin{aligned}\sum_{\boldsymbol{u}\in\mathcal{S}_R} (\tfrac{1}{2}\boldsymbol{u}^\top\boldsymbol{A}\boldsymbol{u})^2 &= \frac{R^4}{2}\sum_a A_{aa}^2 + \frac{R^4}{4}\sum_{a<b}(A_{aa}+A_{bb})^2 + R^4\sum_{a<b}A_{ab}^2 \\ &= \frac{R^4}{4}\left[d\sum_a A_{aa}^2 + (\operatorname{tr}\boldsymbol{A})^2 + 4\sum_{a<b}A_{ab}^2\right] \\ &\geq \frac{R^4}{2}\|\boldsymbol{A}\|_{\mathrm{F}}^2.\end{aligned} \tag{20}$$

Thus the zero-intercept design has full column rank.

(Approximation.) Set $\rho(\boldsymbol{u}) = F(\boldsymbol{u}) - \nabla F(\mathbf{0})^\top\boldsymbol{u} - \frac{1}{2}\boldsymbol{u}^\top\boldsymbol{B}\boldsymbol{u}$. Integral Taylor expansion gives

$$\begin{aligned}|\rho(\boldsymbol{u})| &= \left|\int_0^1 (1-t)\boldsymbol{u}^\top\big(\nabla^2 F(t\boldsymbol{u}) - \boldsymbol{B}\big)\boldsymbol{u}\, dt\right| \\ &\leq M\|\boldsymbol{u}\|_2^3\int_0^1 t(1-t)\, dt = \frac{M}{6}\|\boldsymbol{u}\|_2^3.\end{aligned} \tag{21}$$

Every probe has norm R. Orthogonal least-squares projection therefore yields

$$\begin{aligned}2dR^2\|\boldsymbol{g} - \nabla F(\mathbf{0})\|_2^2 &= \sum_{\boldsymbol{u}\in\mathcal{S}_R}\big((\boldsymbol{g} - \nabla F(\mathbf{0}))^\top\boldsymbol{u}\big)^2 \\ &\leq \sum_{\boldsymbol{u}\in\mathcal{S}_R}\rho(\boldsymbol{u})^2 \leq 2d^2\left(\frac{MR^3}{6}\right)^2 = \frac{d^2M^2R^6}{18}.\end{aligned} \tag{22}$$

$$\begin{aligned}\frac{R^4}{2}\|\boldsymbol{H} - \boldsymbol{B}\|_{\mathrm{F}}^2 &\leq \sum_{\boldsymbol{u}\in\mathcal{S}_R}\big(\tfrac{1}{2}\boldsymbol{u}^\top(\boldsymbol{H} - \boldsymbol{B})\boldsymbol{u}\big)^2 \\ &\leq \sum_{\boldsymbol{u}\in\mathcal{S}_R}\rho(\boldsymbol{u})^2 \leq \frac{d^2M^2R^6}{18}.\end{aligned} \tag{23}$$

Taking square roots and using $\|\boldsymbol{H} - \boldsymbol{B}\|_2 \leq \|\boldsymbol{H} - \boldsymbol{B}\|_{\mathrm{F}}$ proves the coefficient bounds. For $\|\boldsymbol{u}\|_2 \leq R$,

$$\begin{aligned}|F(\boldsymbol{u}) - \widehat{\mathcal{E}}(\boldsymbol{u})| &\leq \|\boldsymbol{u}\|_2\|\boldsymbol{g} - \nabla F(\mathbf{0})\|_2 + \tfrac{1}{2}\|\boldsymbol{u}\|_2^2\|\boldsymbol{H} - \boldsymbol{B}\|_2 + |\rho(\boldsymbol{u})| \\ &\leq \frac{\sqrt{d}}{6}MR^3 + \frac{d}{6}MR^3 + \frac{1}{6}MR^3 = \frac{d+\sqrt{d}+1}{6}MR^3.\end{aligned} \tag{24}$$

Exact recovery for quadratic F follows from $\rho \equiv 0$. □

Identification of interactions. For $a < b$, sign symmetry reduces the off-diagonal normal equation to

$$\begin{aligned}H_{ab} &= \frac{1}{R^4}\sum_{\boldsymbol{u}\in\mathcal{S}_R} u_a u_b F(\boldsymbol{u}) \\ &= \frac{1}{2R^2}\sum_{s,t\in\{-1,1\}} st\, F\left(\frac{R}{\sqrt{2}}(s\boldsymbol{e}_a + t\boldsymbol{e}_b)\right).\end{aligned} \tag{25}$$

Axis probes instead give $\widehat{\mathcal{E}}(s\boldsymbol{e}_a) = sg_a + s^2H_{aa}/2$, which is independent of H_{ab} for $a \neq b$.

A.4 Comparison with Diagonal Control

The diagonal ablation uses the same fitted $\boldsymbol{g}$ and H_{aa}:

$$\boldsymbol{u}_{\mathrm{diag}} \in \operatorname*{arg\,min}_{\|\boldsymbol{u}\|_2 \le R} \left(\boldsymbol{g}^\top \boldsymbol{u} + \tfrac{1}{2} \sum_a H_{aa} u_a^2 \right). \tag{26}$$

Proposition A.3 (Order of the diagonal change)**.** *Under the assumptions of Proposition A.2, at a stationary reference with $\beta > 0$, the diagonal controller satisfies $\|\boldsymbol{u}_{\mathrm{diag}}\|_2 = O(R^2)$ and $|F(\boldsymbol{u}_{\mathrm{diag}})| = O(R^4)$ as $R \downarrow 0$, with the local constants fixed.*

Proof. For sufficiently small R, $dMR \le 3\beta/2$ and $H_{aa} \ge \beta - dMR/3 \ge \beta/2$. Writing $\mathrm{Diag}(\boldsymbol{H})$ for the diagonal part, optimality gives $(\mathrm{Diag}(\boldsymbol{H}) + \mu \boldsymbol{I}_d)\boldsymbol{u}_{\mathrm{diag}} = -\boldsymbol{g}$ for some $\mu \ge 0$. The coefficient bounds imply

$$\begin{aligned} \|\boldsymbol{u}_{\mathrm{diag}}\|_2 &\le \|(\mathrm{Diag}(\boldsymbol{H}) + \mu \boldsymbol{I}_d)^{-1}\|_2 \|\boldsymbol{g}\|_2 \\ &= \frac{\|\boldsymbol{g}\|_2}{\min_a H_{aa} + \mu} \le \frac{2}{\beta} \|\boldsymbol{g}\|_2 \le \frac{\sqrt{d}\, M}{3\beta} R^2. \end{aligned} \tag{27}$$

Using $\nabla F(\mathbf{0}) = \mathbf{0}$ and Eq. (21),

$$\begin{aligned} |F(\boldsymbol{u}_{\mathrm{diag}})| &\le \tfrac{1}{2} \|\boldsymbol{B}\|_2 \|\boldsymbol{u}_{\mathrm{diag}}\|_2^2 + \tfrac{M}{6} \|\boldsymbol{u}_{\mathrm{diag}}\|_2^3 \\ &\le \left(\frac{\|\boldsymbol{B}\|_2}{2} + \frac{MR}{6} \right) \|\boldsymbol{u}_{\mathrm{diag}}\|_2^2 \\ &\le \frac{dM^2}{18\beta^2} \left(\|\boldsymbol{B}\|_2 + \frac{MR}{3} \right) R^4. \end{aligned} \tag{28}$$

□

For an exact quadratic at the stationary reference, $M = 0$ gives $\boldsymbol{g} = \mathbf{0}$, $\boldsymbol{H} = \boldsymbol{B}$, and $\boldsymbol{u}_{\mathrm{diag}} = \mathbf{0}$.

A.5 Proof of Theorem 1

Proof. **(Coupled descent.)** Choose a unit vector $\boldsymbol{z}$ satisfying $\boldsymbol{B}\boldsymbol{z} = -\gamma \boldsymbol{z}$, with its sign fixed so that $\boldsymbol{g}^\top \boldsymbol{z} \le 0$. Global surrogate minimization, Proposition A.2, and Eq. (24) give

$$\begin{aligned} F(\boldsymbol{u}^\star) &\le \widehat{\mathcal{E}}(\boldsymbol{u}^\star) + \frac{d + \sqrt{d} + 1}{6} MR^3 \\ &\le R \boldsymbol{g}^\top \boldsymbol{z} + \frac{R^2}{2} \boldsymbol{z}^\top \boldsymbol{H} \boldsymbol{z} + \frac{d + \sqrt{d} + 1}{6} MR^3 \\ &\le -\frac{\gamma R^2}{2} + \frac{R^2}{2} \|\boldsymbol{H} - \boldsymbol{B}\|_2 + \frac{d + \sqrt{d} + 1}{6} MR^3 \\ &\le -\frac{\gamma R^2}{2} + \frac{2d + \sqrt{d} + 1}{6} MR^3 = -\Delta_R. \end{aligned} \tag{29}$$

(Single-direction lower bound.) For $|s| \le R$, completing the square gives

$$\begin{aligned} F(s\boldsymbol{e}_a) &\ge s\, \partial_a F(\mathbf{0}) + \tfrac{1}{2} B_{aa} s^2 - \tfrac{M}{6} |s|^3 \\ &\ge s\, \partial_a F(\mathbf{0}) + \tfrac{1}{2} (\beta - MR/3) s^2 \\ &= \frac{\beta - MR/3}{2} \left(s + \frac{\partial_a F(\mathbf{0})}{\beta - MR/3} \right)^2 - \frac{|\partial_a F(\mathbf{0})|^2}{2(\beta - MR/3)} \\ &\ge -\frac{|\partial_a F(\mathbf{0})|^2}{2(\beta - MR/3)} \ge -\frac{\|\nabla F(\mathbf{0})\|_\infty^2}{2(\beta - MR/3)}. \end{aligned} \tag{30}$$

At $\nabla F(\mathbf{0}) = \mathbf{0}$, this gives $F(s\boldsymbol{e}_a) \ge (\beta - MR/3)s^2/2 > 0$ whenever $0 < |s| \le R$.

(Acceptance and separation.) Equation (16) ensures $F(\boldsymbol{u}^\star) \le -\Delta_R < -\tau$, so $\boldsymbol{\alpha}^\star = \bar{\boldsymbol{\alpha}} + \boldsymbol{V}\boldsymbol{u}^\star$ is accepted. Combining the two bounds,

$$\begin{aligned}\mathcal{E}(\boldsymbol{\alpha}^\star) &= \mathcal{E}(\bar{\boldsymbol{\alpha}}) + F(\boldsymbol{u}^\star) \\ &\le \mathcal{E}(\bar{\boldsymbol{\alpha}}) - \Delta_R \\ &< \mathcal{E}(\bar{\boldsymbol{\alpha}}) - \frac{\|\nabla F(\mathbf{0})\|_\infty^2}{2(\beta - MR/3)} \\ &\le \min_{\substack{1\le a\le d \\ |s|\le R}} \mathcal{E}(\bar{\boldsymbol{\alpha}} + s\boldsymbol{V}\boldsymbol{e}_a).\end{aligned} \tag{31}$$

The second line and $\Delta_R > \tau$ give Eq. (14). At a stationary reference, Proposition A.3 gives $|F(\boldsymbol{u}_{\mathrm{diag}})| = o(R^2)$, whereas Eq. (29) gives $-F(\boldsymbol{u}^\star) \ge \gamma R^2/2 - O(R^3)$. For an exact quadratic, the latter simplifies to $F(\boldsymbol{u}^\star) \le -\gamma R^2/2$. □

B Complexity

Let $P = KL$ be the number of regional coefficients, N the calibration size, D the pooled readout dimension, and d the control dimension. Each coefficient evaluation uses at most $2N$ original and crossed inputs. Computing all P directional derivatives and accumulating influence weights costs $O(NP(C_J + D))$, where C_J is the cost of one directional derivative traversal. Let C_G denote the total graph-transport cost and C_{KL} the output-agreement cost per evaluation. Then

$$C_{\mathrm{eval}} = O(NP(C_J + D) + C_G) + C_{\mathrm{KL}}. \tag{32}$$

Influence weights are averaged before transport, so each task–channel graph is processed once per coefficient evaluation. With E total edges across these graphs and Q readout nodes per graph, nested depth neighborhoods admit an exact structured solver with $C_G = O(E(P + Q + L))$. General supports retain the exact transport solver. Batching derivative directions improves execution efficiency while retaining all P directions.

The response basis is obtained from the streamed $P \times P$ Gram matrix, requiring $O(NDP^2 + P^3)$ additional work. In the d-dimensional subspace, the symmetric design uses $2d^2$ probes, followed by one acceptance evaluation. The structured quadratic fit costs $O(d^2)$, and the spectral control solve costs $O(d^3)$ at fixed numerical tolerance. Including the reference evaluation and one-time expert-target construction C_{target} gives

$$C_{\mathrm{select}} = C_{\mathrm{target}} + O\big(NDP^2 + P^3 + (d^2+1)C_{\mathrm{eval}} + d^3\big). \tag{33}$$

Thus the leading model-derivative term is $O(d^2NKLC_J)$. If each expert contributes n calibration examples, $N = Kn$ and this term becomes $O(d^2K^2LnC_J)$. The low-dimensional controller reduces objective queries, while every query still measures all KL coefficient directions. Pair-wise streaming requires $O(DP + P^2)$ auxiliary storage for the readout derivatives and Gram matrix, in addition to model weights, expert factors, graph statistics, targets, and batched activation workspace.

The computational budget is controllable through calibration size N, region count L, and control dimension d. The control Hessian is only $d \times d$, and the method does not form a model-parameter Hessian. For rank-r LoRA updates to matrices of shape $m_j \times n_j$, final folding costs $O(Kr\sum_j m_j n_j)$ and can use layer-wise temporary storage. After expert training, the entire merge procedure is performed once for each fixed expert set and merge configuration. The backbone and expert updates remain frozen during coefficient selection, which requires no multi-epoch fine-tuning or repeated rounds of model training. The resulting single dense model has unchanged parameter shapes and inference complexity, with no influence measurement or coefficient search during deployment.

C Experimental Details

Datasets. We use the released MM-MergeBench (Zeng et al., 2025), where a training record may contain multiple question–answer turns. The training/evaluation record counts for the eight seen tasks are as follows: ScienceQA, 12,726/4,241, ImageNet, 129,833/5,050, VQAv2, 82,783/10,000, REC-COCO, 55,885/6,000, OCR-VQA, 165,348/10,000, VizWiz-Caption,

Table 4: Ablation of the influence-matching objective.

	Qwen3.5-4B		Qwen3-VL-8B	
Objective	Seen	Unseen	Seen	Unseen
KL only	76.58 ± 0.26	66.52 ± 0.19	75.87 ± 0.33	64.99 ± 0.24
Curvature only	74.84 ± 0.38	64.91 ± 0.29	74.38 ± 0.50	63.98 ± 0.35
KL + normalized-edge MSE	77.04 ± 0.25	66.79 ± 0.19	76.16 ± 0.33	65.23 ± 0.23
KL + curvature	$\mathbf{78.50 \pm 0.24}$	$\mathbf{68.10 \pm 0.18}$	$\mathbf{77.50 \pm 0.31}$	$\mathbf{66.50 \pm 0.22}$

117,155/7,750, Flickr30k, 127,130/6,357, and IconQA, 29,859/9,954. Together, these datasets contain 720,719 training records and 59,352 seen-task evaluation records before filtering. Retaining only ScienceQA records that contain images reduces its training/evaluation counts to 6,218/2,017. The four unseen evaluation datasets contain 1,145 records for A-OKVQA, 5,999 for ImageNet-R, 4,310 for Screen2Words, and 7,686 for TabMWP, totaling 19,140 records.

Implementation details. `TAC-Merge`uses 32 unlabeled image–instruction inputs per seen task, for a total of 256 calibration inputs, with calibration seeds $\{0, 1, 2\}$ and fixed expert checkpoints. We partition the vision encoder and language backbone into four contiguous depth groups each, yielding $L = 8$ groups in total. Token-group readouts are mean-pooled and scaled by the inverse square root of the hidden width. Expert targets are generated using greedy decoding with a maximum of 128 continuation tokens. We compute sensitivities using FP32 Jacobian–vector products, solve graph transport using unregularized linear programming, and fit the quadratic surrogate in FP64. We set $d = 4$, $\lambda = 1$, $R = 0.25$, and the reference coefficients $\bar{\alpha}_{\ell k} = 1/8$. The selected coefficients are used to merge the expert updates into a single model shared across tasks.

D Additional Ablation Studies

D.1 Ablation of Multimodal Influence Mapping

We ablate two components of MIM: the influence-matching objective and the construction of multimodal influence graphs.

Influence-matching objective. Table 4 compares four influence-matching objectives. (1) KL only matches output probabilities. (2) Curvature only matches influence structure. (3) KL + normalized-edge MSE combines output matching with squared differences between influence weights normalized within each task and channel. (4) KL + curvature combines output matching with expert-curvature matching in the same response subspace. KL + curvature achieves the highest seen- and unseen-task scores for both models. Compared with KL only, it improves seen/unseen scores by 1.92/1.58 points for Qwen3.5-4B and 1.63/1.51 points for Qwen3-VL-8B. Curvature only underperforms KL only on both task groups, while combining the two objectives improves performance, suggesting that influence-structure matching complements direct output matching. KL + normalized-edge MSE also improves over KL only but remains below KL + curvature. This comparison suggests that matching the organization of neighboring influences through the architectural graph distance is more effective than matching individual normalized edge weights.

Modality contrasts and influence structure. Table 5 examines the effects of modality contrasts, observation boundaries, and source granularity. (1) Original only uses original readouts without modality contrasts and corresponds to MIM: w/o modality contrasts in Table 3. (2) Original + image additionally includes responses to image changes. (3) Original + instruction additionally includes responses to instruction changes. These channel choices apply to both the matching objective and the response basis. (4) Immediate boundary only restricts each source to its nearest reachable observation boundary. (5) Region-pooled sources aggregates expert influences within each region into a single source vertex. Variants (4) and (5) modify both the candidate and target graphs while retaining the full response basis. (6) Full MIM includes both modality contrasts, all reachable observation boundaries, and separate expert sources. Full MIM improves seen/unseen scores over

Table 5: Ablation of modality contrasts and influence structure.

	Qwen3.5-4B		Qwen3-VL-8B	
MIM variant	Seen	Unseen	Seen	Unseen
Original only	77.46 ± 0.26	67.87 ± 0.21	77.09 ± 0.29	66.29 ± 0.24
Original + image	78.08 ± 0.25	67.82 ± 0.19	77.18 ± 0.33	66.27 ± 0.23
Original + instruction	77.70 ± 0.26	67.66 ± 0.19	76.98 ± 0.33	66.12 ± 0.24
Immediate boundary only	77.42 ± 0.28	67.39 ± 0.21	76.67 ± 0.35	65.91 ± 0.25
Region-pooled sources	77.55 ± 0.28	67.46 ± 0.21	76.81 ± 0.35	65.99 ± 0.25
Full MIM	$\mathbf{78.50 \pm 0.24}$	$\mathbf{68.10 \pm 0.18}$	$\mathbf{77.50 \pm 0.31}$	$\mathbf{66.50 \pm 0.22}$

Original only by 1.04/0.23 points for Qwen3.5-4B and 0.41/0.21 points for Qwen3-VL-8B. Image contrasts alone yield higher scores than instruction contrasts alone, but neither single-contrast variant improves mean unseen-task performance over Original only. Instruction contrasts capture responses to changes in task specifications while holding the image fixed. The higher scores achieved by Full MIM suggest that using both contrasts provides useful information beyond either contrast alone. Restricting observations to the immediate boundary excludes influence measurements at later boundaries and reduces seen-task scores by 1.08 and 0.83 points for the two models, respectively. Pooling sources removes distinctions between expert contributions within a region and reduces seen-task scores by 0.95 and 0.69 points, respectively. Together, these ablations support retaining both modality contrasts, downstream observations, and individual expert identities in the influence graphs.

D.2 Ablation of Coupled Merge Control

Table 6 examines the choice of coefficient directions, surrogate order, and interactions between directions under the full MIM objective. (1) Regional random search corresponds to w/o CMC in Table 3 and selects the best candidate from a set containing the reference coefficients $\bar{\alpha}$ and coefficient vectors sampled uniformly within the prescribed ball. (2) Random basis + coupled fits a full quadratic controller in a fixed random orthonormal basis. (3) Response basis + linear uses the leading right singular vectors of the MIM response matrix as coefficient directions and retains only the linear term $g^\top u$. (4) Response basis + diagonal corresponds to CMC: w/o directional coupling in Table 3 and retains $g^\top u + \frac{1}{2}\sum_a H_{aa}u_a^2$. (5) Response basis + coupled uses the full quadratic model $g^\top u + \frac{1}{2}u^\top Hu$.

Response-informed coefficient directions. Variants (2) and (5) isolate the effect of basis choice while using the same coupled controller and objective-evaluation budget. The response basis improves seen/unseen scores over the random basis by 1.68/0.91 points for Qwen3.5-4B and 1.19/0.70 points for Qwen3-VL-8B. The response basis selects coefficient combinations that induce strong measured multimodal responses, whereas the random basis does not use this information. The observed gains support using calibration responses to identify effective control directions. Regional random search provides an additional comparison for coefficient selection without a fitted controller.

Surrogate order and directional coupling. Variants (3)–(5) use the same response basis and share the fitted surrogate coefficients for the terms they retain. The linear variant removes all quadratic terms, while the diagonal variant removes only cross-direction terms. Adding diagonal quadratic terms improves seen/unseen scores by 0.67/0.52 points for Qwen3.5-4B and 0.59/0.60 points for Qwen3-VL-8B. Retaining cross-direction terms yields further gains of 0.71/0.39 and 0.48/0.12 points, respectively. A nonzero linear objective over a radius-constrained ball attains its minimum on the boundary, whereas quadratic terms allow marginal effects to vary with the adjustment and can yield an interior optimum. Cross-direction terms capture how an adjustment along one direction changes the effect of an adjustment along another. Orthogonality in coefficient space does not imply that these directions act independently in the model. A quadratic surrogate is the lowest-degree polynomial that can represent such pairwise interactions. Coefficient selection is performed once per merge configuration, after which the selected updates are folded into a single model with unchanged parameter shapes and computational structure at inference.

Table 6: Ablation of coefficient directions and quadratic coupling.

	Qwen3.5-4B		Qwen3-VL-8B	
Control variant	Seen	Unseen	Seen	Unseen
Regional random search	77.03 ± 0.46	67.58 ± 0.32	76.64 ± 0.48	66.26 ± 0.34
Random basis + coupled	76.82 ± 0.31	67.19 ± 0.23	76.31 ± 0.40	65.80 ± 0.29
Response basis + linear	77.12 ± 0.30	67.19 ± 0.22	76.43 ± 0.38	65.78 ± 0.27
Response basis + diagonal	77.79 ± 0.30	67.71 ± 0.20	77.02 ± 0.36	66.38 ± 0.19
Response basis + coupled	$\mathbf{78.50 \pm 0.24}$	$\mathbf{68.10 \pm 0.18}$	$\mathbf{77.50 \pm 0.31}$	$\mathbf{66.50 \pm 0.22}$

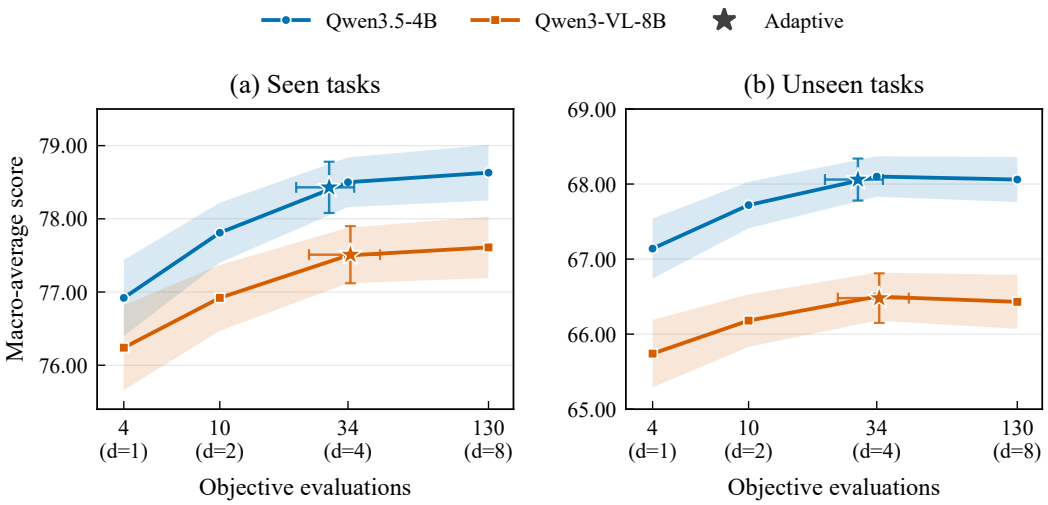

Figure 3: Performance as a function of control dimension and evaluation cost. Curves show results for fixed dimensions, and stars indicate adaptive selection. Bands and error bars indicate standard deviations.

Prediction at unseen coefficient configurations. We evaluate the predictive fidelity of the linear, diagonal quadratic, and coupled quadratic surrogates, corresponding to variants (3)–(5) above. The evaluation uses coefficient vectors sampled uniformly from the radius-R ball and excluded from surrogate fitting. These configurations are used only for diagnostics, with the reference objective evaluated on the same calibration inputs used for fitting. Table 7 compares predicted and reference objective changes, both measured relative to the objective value at $\bar{\alpha}$. NRMSE is the root mean square prediction error divided by the root mean square reference change. Spearman correlation measures agreement between the predicted and reference rankings of coefficient configurations. Replacing the linear surrogate with the coupled quadratic surrogate reduces NRMSE from 0.62 to 0.21 for Qwen3.5-4B and from 0.66 to 0.24 for Qwen3-VL-8B. Spearman correlation increases from 0.48 to 0.89 and from 0.44 to 0.86, respectively. The diagonal quadratic surrogate falls between the linear and coupled quadratic surrogates on both metrics. These results indicate that cross-direction terms improve both prediction accuracy and the ranking of candidate configurations. To account for residual surrogate error, the controller evaluates each proposal using the reference objective before accepting it.

Control dimension and adaptive selection. Fixed-dimension control uses the leading $d \in \{1, 2, 4, 8\}$ response directions. Adaptive control selects $d = \min\{8, \lceil \mathrm{srank}(J) \rceil\}$ using the nonzero response matrix J evaluated at $\bar{\alpha}$. Figure 3 compares task performance and the number of objective evaluations. Increasing d from 1 to 4 improves seen/unseen scores by 1.58/0.96 points for Qwen3.5-4B and 1.26/0.76 points for Qwen3-VL-8B. Increasing d from 4 to 8 yields only 0.13 and 0.11 additional seen-task points, respectively, with no improvement in mean unseen-task scores. Meanwhile, the number of objective evaluations increases from 34 to 130. Adaptive selection achieves

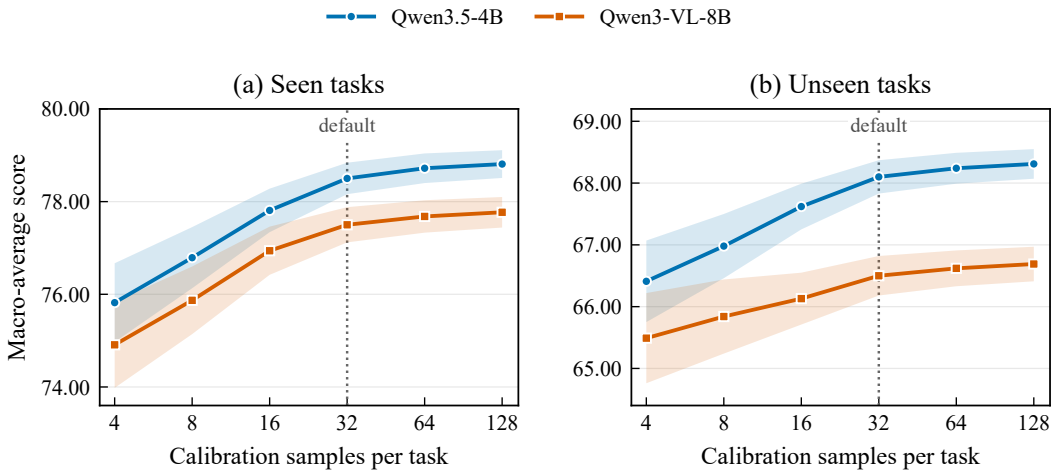

Figure 4: Performance as a function of calibration sample size. Bands indicate standard deviations. Dotted lines mark the default sample size.

Table 7: Local surrogate fidelity at coefficient configurations excluded from fitting ($d = 4$). Entries report mean $\pm$ standard deviation.

	Qwen3.5-4B		Qwen3-VL-8B	
Local surrogate	NRMSE ↓	Spearman ↑	NRMSE ↓	Spearman ↑
Linear	0.62 ± 0.06	0.48 ± 0.07	0.66 ± 0.07	0.44 ± 0.08
Diagonal quadratic	0.40 ± 0.05	0.72 ± 0.05	0.44 ± 0.05	0.69 ± 0.06
Coupled quadratic	$\mathbf{0.21 \pm 0.03}$	$\mathbf{0.89 \pm 0.03}$	$\mathbf{0.24 \pm 0.04}$	$\mathbf{0.86 \pm 0.04}$

scores and evaluation counts close to those of $d = 4$, using only the calibration response spectrum to choose the dimension. Evaluation counts follow $Q = 2d^2 + 2$ and exclude expert-target construction, so they quantify coefficient-selection cost rather than total merging time. The results for $d = 4$ match those reported in the preceding tables.

D.3 Calibration Data Efficiency

Calibration sample size. We vary the number of calibration examples per seen task over $N_t \in \{4, 8, 16, 32, 64, 128\}$. The calibration subsets are balanced, nested, and disjoint from the test data, with smaller subsets retaining complete pairs from the largest subset. Expert targets and the response basis are recomputed for each subset. Figure 4 shows how task performance varies with calibration sample size. The largest gains occur below 32 examples per task, while smaller subsets exhibit greater variability. Increasing the sample count from 16 to 32 improves seen/unseen scores by 0.69/0.48 points for Qwen3.5-4B and 0.56/0.37 points for Qwen3-VL-8B. Increasing it from 32 to 128 yields only 0.31/0.21 and 0.27/0.19 additional points, respectively, while quadrupling the number of examples processed per objective evaluation. We therefore use 32 examples per task by default to balance calibration cost and performance.