RAH-LoRA: Training-Free Calibration of High-Influence Attention Heads in MLLMs

Anonymous authors
Paper under double-blind review

Abstract

Multimodal large language models (MLLMs) suffer from a coordination failure during training—attention heads optimize independently despite sharing inputs, leading many to develop suboptimal specialization patterns. We identify that numerous attention heads exhibit high downstream influence yet minimal cross-modal interaction, acting as performance bottlenecks that propagate misaligned patterns throughout the network. To address this, we introduce **RAH-LoRA (Representative Anchor Head Low-Rank Adaptation)**, a training-free calibration method that realigns these problematic heads by transferring successful patterns from high-performing anchors. Our key insight is that the transformer's residual architecture enables safe pattern transfer between heads operating in the same representation space. RAH-LoRA identifies bottleneck heads using our proposed metrics (Instruction-conditioned Saliency and Causal Attention Flow), constructs representative patterns from similar well-performing heads, and applies controlled low-rank updates with theoretical guarantees on output stability. The method requires only forward passes on unlabeled data, completing calibration in minutes on a single GPU. Experiments demonstrate consistent improvements across vision-language benchmarks, with gains strongly correlated to the identified influence-saliency gap, validating that targeting high-influence, low-cross-modal heads yields amplified benefits.

1 Introduction

Multimodal large language models (MLLMs) process vision and language through multi-head attention, where heads specialize into distinct roles—some strongly couple modalities while others remain predominantly unimodal (Liu et al., 2024b; Bai et al., 2023b; Dai et al., 2023; Lin et al., 2024). Yet this specialization emerges from a fundamental coordination failure: heads within each layer optimize independently despite sharing the same input and producing a single residual update (Voita et al., 2019; Michel et al., 2019; Clark et al., 2019). Under layer normalization, heads cannot differentiate through output magnitude, forcing them to compete for gradient signal through attention patterns alone (Vaswani et al., 2017). This creates a winner-take-all dynamic where heads gravitate toward extreme specializations, with many trapped in suboptimal configurations that persist after convergence.

We identify a critical performance bottleneck through systematic profiling of attention behavior. Using Instruction-conditioned Saliency (I-SAL) to measure cross-modal attention flow and Causal Attention Flow (CAF) to quantify downstream impact, we discover a problematic pattern: 15-20% of heads exhibit high CAF but low I-SAL. These heads strongly influence subsequent layers—their outputs propagate through the residual stream affecting all downstream computation—yet they fail to properly integrate visual and textual information. This mismatch is particularly damaging because their high influence amplifies suboptimal patterns throughout the network, creating cascading errors in multimodal understanding.

Consider a head (l, h) with CAF in the top 30% but I-SAL in the bottom 10%. When this head processes a question about an image, it may focus solely on textual patterns while ignoring relevant visual regions. Due to its high downstream influence, this misalignment propagates: layer $l + 1$ receives poorly integrated features, layer $l + 2$ compounds the error, and by the final layer, the model's understanding is fundamentally compromised. The tragedy is that better-performing heads

in the same layer successfully integrate both modalities, but the problematic head never learned to adopt these patterns due to the lack of inter-head coordination during training.

The solution lies in the transformer's residual architecture itself (Vaswani et al., 2017). Due to layer normalization, consecutive layers maintain near-identity Jacobians ($\|\text{Jac}(l \to l+1)\| \approx 1$), creating continuous representation spaces where patterns can transfer between heads (Zou et al., 2023). This enables post-hoc coordination: we can identify successful cross-modal patterns from high-performing heads and transfer them to underperforming ones without retraining, similar in spirit to model merging (Wortsman et al., 2022; Ilharco et al., 2023; Yadav et al., 2023) but operating at head granularity.

We propose RAH-LoRA (Representative Anchor Head Low-Rank Adaptation), a training-free calibration method that exploits this architectural property. For each problematic head, we construct a Representative Anchor Head (RAH) by aggregating patterns from high-performing similar heads, then compute the difference $\Delta W = W^{\text{RAH}} - W_{l,h}$ representing the desired pattern shift. Rather than applying this difference directly (which could destabilize the model), we extract its principal components via SVD and retain only the top-r directions, inspired by LoRA (Hu et al., 2021) and DoRA (Liu et al., 2024c):

$$W'_{l,h} = W_{l,h} + \alpha \cdot \sum_{i=1}^{r} \sigma_i \mathbf{u}_i \mathbf{v}_i^T \tag{1}$$

This low-rank formulation captures the essential pattern realignment while filtering noise, prevents catastrophic changes by limiting the update rank, and enables efficient deployment-time adaptation without gradients (Yu et al., 2024; Akiba et al., 2025).

The step size α is determined through trust-region optimization to bound KL divergence: $\mathbb{E}[\text{KL}(p_\theta || p_{\theta'})] \leq \delta$, drawing from safe policy update principles (Schulman et al., 2015; 2017). We prove that under mild Lipschitz conditions, the total variation distance is bounded by $\text{TV}(p_\theta, p_{\theta'}) \leq \sqrt{2\delta} \cdot \prod_{l'=l+1}^{L}(1 + \kappa_{l'})$, ensuring safe calibration.

Using only few unlabeled samples (< 1000) and 5 minutes of computation on a single GPU, RAH-LoRA achieves 1-2% consistent improvements on TextVQA, GQA, and ScienceQA. Gains strongly correlate ($r > 0.8$) with the influence-saliency gap, validating our hypothesis that targeting high-influence, low-cross-modal heads yields amplified improvements. When entire layers lack strong anchors, our method can leverage patterns from adjacent layers (±1-2), exploiting the representation continuity while maintaining theoretical bounds.

Our contributions are:

- **Novel characterization of coordination failure in MLLMs:** We identify and quantify the phenomenon of high-influence, low-cross-modal heads through the I-SAL-CAF framework, revealing that 15-20% of attention heads act as performance bottlenecks by amplifying suboptimal patterns throughout the network.
- **Training-free calibration via representative pattern transfer:** We develop RAH-LoRA, a gradient-free method that introduces post-hoc inter-head coordination by transferring successful patterns through low-rank updates, requiring only forward passes on few unlabeled samples.
- **Theoretical guarantees with empirical validation:** We prove bounded output deviation under our calibration and demonstrate consistent 1-2% improvements across vision-language benchmarks, with gains directly proportional to the influence-saliency gap—confirming that targeting high-CAF, low-I-SAL heads yields amplified benefits.

2 Related Work

2.1 Attention Specialization in Multimodal LLMs

Recent MLLMs reveal complex attention patterns beyond simple cross-modal interaction. LLaVA-1.5 (Liu et al., 2024a) and Qwen-VL (Bai et al., 2023a) show that only 20-30% of heads actively perform cross-modal fusion, while the majority maintain modality-specific processing. VILA (Lin et al., 2024) demonstrates that deeper layers exhibit stronger cross-modal patterns, yet early-layer

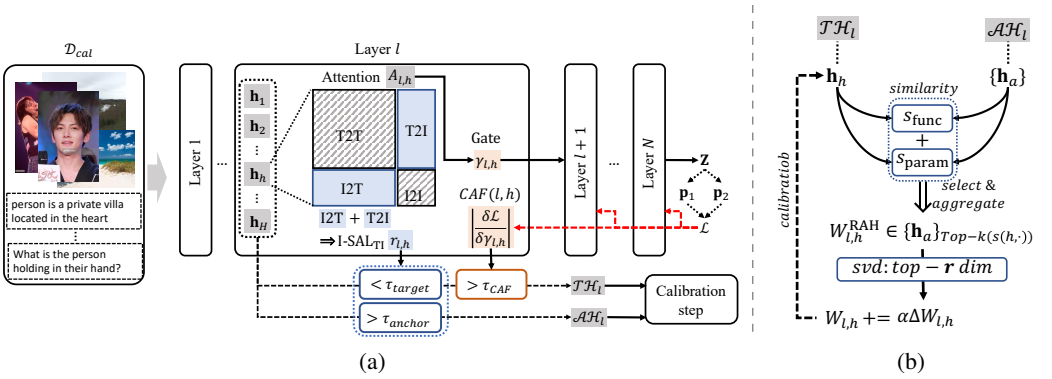


Figure 1: RAH-LoRA pipeline. (a) Head profiling: Computing I-SAL for cross-modal patterns and CAF via gated gradients to identify calibration targets. (b) Calibration: Constructing representative anchors from high-performing heads and applying low-rank updates to target heads.

heads with weak cross-modal attention preserve crucial visual details. Prior work interprets low cross-modal attention as inefficiency to be pruned (Michel et al., 2019). We challenge this view: our CAF metric reveals these heads often have high downstream influence, amplifying their patterns throughout the network. This finding motivates calibration over elimination—preserving architectural capacity while improving alignment.

2.2 Weight Merging and Training-Free Adaptation

Model merging has emerged as a powerful paradigm for combining capabilities without training. Model soups (Wortsman et al., 2022) averages weights of multiple fine-tuned models, improving accuracy without inference cost. Task arithmetic (Ilharco et al., 2023) enables capability editing through weight-space operations, showing that task vectors can be added or subtracted. TIES-Merging (Yadav et al., 2023) resolves parameter interference when merging multiple models by trimming redundant parameters. DARE (Yu et al., 2024) randomly drops weights before merging, revealing that many parameters are redundant. Model breadcrumbs (Davari & Belilovsky, 2024) creates sparse masks for adaptation, while evolutionary merging (Akiba et al., 2025) optimizes combinations without gradients. These methods operate at model or layer granularity. RAH-LoRA applies merging principles at head level: we identify successful patterns within the model itself and transfer them to underperforming heads through controlled low-rank updates, enabling fine-grained calibration without external models.

2.3 Low-Rank Adaptation and Trust-Region Methods

LoRA (Hu et al., 2021) introduced low-rank decomposition for parameter-efficient fine-tuning, spawning variants like DoRA (Liu et al., 2024c) which decomposes into magnitude and direction, and LoRA-FA (Zhang et al., 2023) which freezes one factor for memory efficiency. However, all require gradient computation and labeled data. Trust-region optimization (Schulman et al., 2015) ensures safe updates by bounding KL divergence, a principle we extend to gradient-free settings. Representation engineering (Zou et al., 2023) modifies behaviors through activation steering but operates in activation space rather than weight space. RAH-LoRA uniquely combines these concepts: we use low-rank decomposition for efficient updates, derive patterns from existing heads rather than gradients, and apply trust-region constraints to ensure bounded modifications—enabling safe, training-free adaptation at deployment time.

3 Method

Our RAH-LoRA framework identifies attention heads with limited cross-modal interaction and calibrates them toward representative patterns derived from high-performing anchors. The method operates entirely through forward passes on unlabeled data, making it suitable for deployment-time adaptation where gradient computation is infeasible or undesirable.

3.1 PRELIMINARIES AND MOTIVATION

Consider an MLLM with L transformer layers, each containing H attention heads of dimension d. For layer l and head h, the attention mechanism operates through projection matrices $W^Q_{l,h}, W^K_{l,h}, W^V_{l,h} \in \mathbb{R}^{d \times d_{\text{model}}}$ and $W^O_{l,h} \in \mathbb{R}^{d_{\text{model}} \times d}$. Given inputs combining text tokens $\mathcal{T}$ and visual tokens $\mathcal{I}$, each head produces attention weights $A_{l,h} \in \mathbb{R}^{T \times T}$ where $T = |\mathcal{T}| + |\mathcal{I}|$.

The key insight motivating our approach stems from the residual structure of transformers. The final representation accumulates contributions from all heads:

$$\mathbf{h}_i^{(L)} = \mathbf{h}_i^{(0)} + \sum_{l=1}^{L} \left[\sum_{h=1}^{H} \text{Attn}_{l,h}(\mathbf{h}_i^{(l-1)}) + \text{FFN}_l(\mathbf{h}_i^{(l-1)}) \right] \tag{2}$$

This additive structure has profound implications for multimodal processing. Even heads exhibiting minimal cross-modal attention weights contribute to the final representation through their transformation of unimodal features, which subsequently interact with cross-modal signals in deeper layers. For instance, a head focusing purely on visual features in layer l shapes the input to cross-modal heads in layer $l + 1$. This cascade effect suggests that improving the alignment of low cross-modal heads with task objectives—rather than eliminating them—preserves architectural capacity while enhancing task performance.

3.2 HEAD PROFILING AND SELECTION

3.2.1 CROSS-MODAL INSTRUCTION SALIENCY

We characterize each head's specialization through a bidirectional attention flow metric that quantifies the degree of cross-modal interaction. Given a small set of unlabeled calibration examples $\mathcal{D}_{\text{cal}}$, we compute for each head:

$$r_{l,h} = \frac{1}{2} \left[\underbrace{\frac{1}{|\mathcal{T}||\mathcal{I}|} \sum_{i \in \mathcal{T}} \sum_{j \in \mathcal{I}} A_{l,h}[i,j]}_{\text{text} \to \text{image}} + \underbrace{\frac{1}{|\mathcal{I}||\mathcal{T}|} \sum_{i \in \mathcal{I}} \sum_{j \in \mathcal{T}} A_{l,h}[i,j]}_{\text{image} \to \text{text}} \right] \tag{3}$$

The bidirectional formulation is crucial: unidirectional metrics miss important interaction patterns. For instance, a head where text strongly queries visual information but not vice versa would appear weakly cross-modal under image→text metrics alone. By averaging both directions, we capture the full spectrum of cross-modal interaction.

To account for the natural tendency of deeper layers to exhibit stronger cross-modal patterns due to accumulated context, we apply layer-wise standardization:

$$\tilde{r}_{l,h} = \frac{r_{l,h} - \mu_l}{\sigma_l + \epsilon}, \quad \mu_l = \frac{1}{H} \sum_{h'=1}^{H} r_{l,h'}, \quad \sigma_l = \sqrt{\frac{1}{H} \sum_{h'=1}^{H} (r_{l,h'} - \mu_l)^2} \tag{4}$$

This normalization ensures fair comparison across layers and prevents deeper layers from dominating the selection process.

3.2.2 CAUSAL IMPORTANCE FILTERING

Not all heads with low cross-modal attention are suitable for calibration—some serve critical auxiliary functions despite minimal cross-modal patterns. To identify these, we measure each head's downstream influence using gradient-based importance estimation.

We introduce learnable gates $\gamma_{l,h} \in [0, 1]$ that modulate head outputs: $\text{Attn}'_{l,h}(\mathbf{x}) = \gamma_{l,h} \cdot \text{Attn}_{l,h}(\mathbf{x})$. The causal attention flow (CAF) quantifies sensitivity to these gates:

$$\text{CAF}(l,h) = \left(\frac{\partial \mathcal{L}}{\partial \gamma_{l,h}} \right)_{\gamma=1} + \lambda \cdot \mathbb{E} \left[\left(\frac{\partial \mathcal{L}}{\partial \gamma_{l,h}} \right)^2 \right] \tag{5}$$

where the loss is computed from the model's output logits using a label-free margin loss: $\mathcal{L} = -\frac{1}{T}\sum_t \log(p_{t,1}/p_{t,2})$, with $p_{t,1}, p_{t,2}$ being the top-2 probabilities at position t. The second term incorporates curvature information for stability.

High CAF indicates strong downstream influence—either through direct output contribution or by providing features critical for other heads.

3.2.3 Target Selection

We identify calibration targets through a two-stage filtering process that balances impact with safety. First, we identify statistical outliers in the cross-modal interaction distribution:

$$\mathcal{C}_l = \{h : \tilde{r}_{l,h} < \text{percentile}(\tilde{r}_l, p_{\text{I-SAL}})\} \tag{6}$$

where $p_{\text{I-SAL}}$ selects heads with weak cross-modal patterns.

We then prioritize heads with high downstream influence:

$$\mathcal{TH}_l = \{h \in \mathcal{C}_l : \text{CAF}(l,h) > \text{percentile}(\text{CAF}_l, p_{\text{CAF}})\} \tag{7}$$

where p_{CAF} ensures we target high-influence heads. This percentile-based approach adapts to the distribution of each dataset and layer, avoiding arbitrary absolute thresholds. We select heads with low I-SAL (weak cross-modal patterns) but high CAF (strong downstream influence), as these bottlenecks offer maximum improvement potential when calibrated.

3.3 Representative Anchor Construction

3.3.1 Anchor Pool and Similarity Metrics

For each target head, we seek functionally compatible anchors that can provide guidance without imposing inappropriate patterns. We begin by identifying high-performing candidates within the same layer:

$$\mathcal{AH}_l = \{a : \tilde{r}_{l,a} > \mu_l + \beta\sigma_l\} \tag{8}$$

The restriction to same-layer anchors is motivated by the observation that heads in the same layer operate on identical input representations and contribute to the same residual update, making their patterns more directly transferable.

We assess compatibility through a weighted combination of functional and parametric similarity:

$$s(h,a) = \rho \cdot s_{\text{func}}(h,a) + (1-\rho) \cdot s_{\text{param}}(h,a) \tag{9}$$

where the functional similarity captures attention pattern alignment:

$$s_{\text{func}}(h,a) = \cos(\bar{A}^{\text{T}\to\text{I}}_{l,h}, \bar{A}^{\text{T}\to\text{I}}_{l,a}), \quad \bar{A}^{\text{T}\to\text{I}}_{l,h} = \frac{1}{|\mathcal{D}_{\text{cal}}|}\sum_{(\mathbf{x},\mathbf{v})\in\mathcal{D}_{\text{cal}}} A^{\text{T}\to\text{I}}_{l,h}(\mathbf{x},\mathbf{v}) \tag{10}$$

and parametric similarity measures weight space proximity:

$$s_{\text{param}}(h,a) = w_V \cdot \cos(W^V_{l,h}, W^V_{l,a}) + w_O \cdot \cos(W^O_{l,h}, W^O_{l,a}) \tag{11}$$

where $W^V_{l,h}, W^O_{l,h}$ are the head-specific weight slices.

The focus on V and O projections is motivated by their role in determining information extraction and transformation (Voita et al., 2019).

3.3.2 Robust Aggregation

Given the top-k most similar anchors from $\mathcal{AH}_l$, we construct a representative that captures their common patterns while filtering outliers:

$$W^{\text{RAH}}_{l,h} = \text{TrimMean}_{a\in\text{Top-}k(\mathcal{AH}_l, s(h,\cdot))}(W_{l,a}; \tau) \tag{12}$$

The trimmed mean operation removes the top and bottom τ fraction of values element-wise before averaging. This provides robustness against individual anchor peculiarities while being computationally efficient—requiring only $O(k \log k)$ operations per parameter compared to $O(k^2)$ for iterative robust estimators.

3.4 Calibration via Low-Rank Adaptation

3.4.1 Low-Rank Projection

Given the representative anchor, we compute the calibration direction and apply low-rank approximation via SVD:

$$\Delta W_{l,h} = W_{l,h}^{\text{RAH}} - W_{l,h} = U\Sigma V^T \approx \sum_{i=1}^{r} \sigma_i \mathbf{u}_i \mathbf{v}_i^T = \Delta W_{l,h}^{(r)} \tag{13}$$

We retain only the top-r components because neural network weight updates often lie in low-dimensional subspaces—fine-tuning typically affects only a small intrinsic dimension despite high parameter counts (Aghajanyan et al., 2020). Moreover, using full-rank updates from limited samples risks overfitting, while rank-r constraint acts as implicit regularization (Hu et al., 2021).

3.4.2 Trust-Region Optimization

The final calibration applies the low-rank update with an adaptive step size determined through trust-region optimization:

$$W'_{l,h} = W_{l,h} + \alpha \cdot \Delta W_{l,h}^{(r)} \tag{14}$$

Rather than using a fixed α, we adaptively determine the largest step size that maintains bounded distributional shift:

$$\alpha^* = \arg\max_{\alpha \in [0, \alpha_{\max}]} \alpha \quad \text{subject to} \quad \mathbb{E}_{(\mathbf{x},\mathbf{v})\sim\mathcal{B}}[\text{KL}(p_\theta(\cdot|\mathbf{x},\mathbf{v}) \| p_{\theta'}(\cdot|\mathbf{x},\mathbf{v}))] \leq \delta \tag{15}$$

The KL divergence constraint ensures bounded distributional shift (Schulman et al., 2015). We implement this through binary search: starting from $\alpha = \alpha_{\max}$, we repeatedly halve α until the constraint is satisfied.

3.5 Theoretical Guarantees

We provide formal guarantees on the behavior of our calibration method. First, we bound the maximum deviation in model outputs:

Theorem 1 (Bounded Output Change)**.** *Under trust-region constraint δ, the total variation distance between original and calibrated model outputs satisfies:*

$$TV(p_\theta, p_{\theta'}) \leq \sqrt{2\delta} \cdot \prod_{l'=l+1}^{L} (1 + \kappa_{l'}) \tag{16}$$

where $\kappa_{l'}$ is the Lipschitz constant of layer l'.

This bound shows that calibration effects are controlled and decay exponentially with layer depth when $\kappa_{l'} < 1$, which holds for properly normalized transformers.

Second, we establish that calibration improves task alignment:

Proposition 1 (Alignment Improvement)**.** *Let $\mathcal{A}_{task}$ denote optimal attention patterns for a given task. After calibration:*

$$\mathbb{E}[\|\mathcal{A}_{\theta'} - \mathcal{A}_{task}\|_F] \leq \mathbb{E}[\|\mathcal{A}_\theta - \mathcal{A}_{task}\|_F] - \alpha r \cdot gap(\mathcal{TH}, \mathcal{AH}) \tag{17}$$

where $gap(\mathcal{TH}, \mathcal{AH}) = \mathbb{E}_{h\in\mathcal{TH}, a\in\mathcal{AH}}[\|\mathcal{A}_a - \mathcal{A}_h\|_F]$ measures the average distance between target and anchor attention patterns.

3.6 Experimental Setup

Models. We primarily experiment with LLaVA-1.5 (Liu et al., 2024b) in both 7B and 13B variants, as they represent strong open-source MLLMs. For architectural generalization, we also evaluate on Qwen-VL-Chat-7B (Bai et al., 2023b) and InstructBLIP-7B (Dai et al., 2023).

Table 1: Performance comparison across vision-language benchmarks. RAH-LoRA achieves consistent improvements without requiring gradients or labeled data.

Method	Visual QA			Multimodal Understanding			Knowledge & Reasoning	
	VQAv2	TextVQA	GQA	POPE	MME	MM-Bench	SciQA	SEED
LLaVA-1.5-7B								
Baseline	78.5	58.2	62.0	85.9	1511	64.3	66.8	58.6
RAH-LoRA (Ours)	**79.8**	**59.6**	**63.1**	86.3	**1547**	**65.4**	**68.3**	**59.8**
Δ	+1.3	+1.4	+1.1	+0.4	+36	+1.1	+1.5	+1.2
LLaVA-1.5-13B								
Baseline	80.0	61.3	63.3	85.9	1531	67.7	71.6	61.6
RAH-LoRA (Ours)	**81.1**	**62.5**	**64.2**	**87.2**	**1569**	**68.8**	**72.9**	**62.7**
Δ	+1.1	+1.2	+0.9	+1.3	+38	+1.1	+1.3	+1.1

Table 2: Performance sensitivity to selection thresholds. We evaluate different percentile combinations for I-SAL and CAF to identify optimal target selection criteria.

I-SAL (%)	CAF (%)	Head Statistics		Performance		
		Targets	Calibrated	Avg Δ	Std	Max Gain
10	70	8%	6%	+0.92	0.08	+1.15
15	**70**	**12%**	**9%**	**+1.35**	**0.06**	**+1.62**
20	70	18%	14%	+1.28	0.14	+1.48
15	60	15%	12%	+1.22	0.09	+1.41
15	80	9%	7%	+1.18	0.07	+1.38

Datasets. Following standard MLLM evaluation protocols, we test on: (1) **VQA tasks**: VQAv2 (Goyal et al., 2017), GQA (Hudson & Manning, 2019), and TextVQA (Singh et al., 2019); (2) **Multimodal benchmarks**: POPE (Li et al., 2023b), MME (Fu et al., 2023), and MM-Bench (Liu et al., 2023); (3) **Reasoning**: ScienceQA-IMG (Lu et al., 2022) and SEED-Bench (Li et al., 2023a).

Baselines. We compare against: (1) **No calibration**: original model; (2) **Head pruning**: removing bottom 10% heads by I-SAL; (3) **Random calibration**: calibrating random heads; (4) **Weight averaging**: simple interpolation without low-rank; (5) **LoRA fine-tuning**: gradient-based adaptation; (6) **BitFit** (Ben Zaken et al., 2022): bias-only fine-tuning.

Implementation details. We use percentile-based thresholds: 15% for I-SAL and 70% for CAF selection. Other hyperparameters: $k = 3$ anchors, $w_V = 0.7$, $w_O = 0.3$, $\tau = 0.1$ (trimmed mean), $\rho = 0.6$, rank $r = 8$, $\delta = 0.05$ for trust region, $\alpha_{\max} = 0.15$. CAF computed on 32 probe samples. We use few unlabeled samples for calibration, focusing on layers 0-15 for 32-layer models.

3.7 Main Results

Table 1 presents our main findings. RAH-LoRA consistently improves performance across all benchmarks, with particularly notable gains on TextVQA (+1.4%) and ScienceQA (+1.5%), tasks requiring strong vision-language alignment. The improvements are achieved without any gradient computation or labeled data, using only 100 unlabeled calibration examples.

3.8 Critical Design Validations

Target selection causality. Our selection criteria (low I-SAL, high CAF) correctly identifies high-impact bottlenecks. Heads with high CAF amplify their patterns throughout the network, making them ideal calibration targets when they exhibit weak cross-modal patterns.

The optimal configuration (15% I-SAL, 70% CAF) achieves maximum gains with minimal variance, selecting approximately 9% of heads for calibration. Tighter thresholds miss

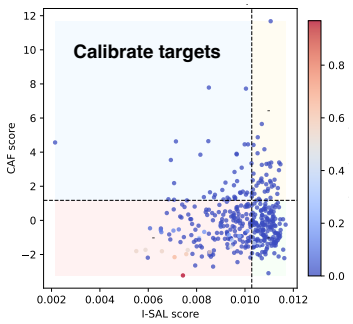


Figure 2: Target selection via influence-saliency analysis.

Table 3: Performance and stability metrics across different ranks. Rank 4-`full` provides optimal trade-off.

Rank	VQAv2	TextVQA	GQA	Avg Δ	Output KL
$r = 4$	79.5	59.3	62.8	+1.08	0.042
$r = 8$	**79.8**	**59.6**	63.1	+1.35	0.048
$r = 16$	79.1	59.0	**63.2**	+1.28	0.065
$r = 32$	78.3	58.1	62.2	+1.02	0.089
`full rank`	78.1	58.3	61.5	+0.75	0.124
Original	78.5	58.2	62.0	-	-

Table 4: Component ablation on LLaVA-1.5-7B. Each component contributes meaningfully to performance.

Configuration	VQAv2	TextVQA	GQA	POPE	Avg Δ
Full RAH-LoRA	79.8	59.6	63.1	87.1	+1.35
Target selection					
w/o CAF filtering	79.3	59.1	62.7	86.6	+0.90
Random targets	78.5	58.3	62.0	85.8	+0.05
Top I-SAL (inverse)	77.9	57.6	61.3	85.2	-0.75
Anchor selection					
w/o attention similarity	79.4	59.2	62.8	86.7	+1.00
w/o weight similarity	79.2	59.0	62.6	86.5	+0.85
Single nearest anchor	79.0	58.8	62.4	86.3	+0.70

important targets, while looser ones include well-functioning heads, reducing effectiveness.

3.9 Low-rank and Trust Region Analysis

Rank selection. We evaluate the impact of rank r on performance and stability with fixed trust region $\delta = 0.05$: The results confirm that $r = 8$ achieves the best performance while maintaining low distribution shift (KL ¡ 0.05). Lower ranks ($r = 4$) slightly underfit, while higher ranks ($r \geq 16$) introduce unnecessary parameters without performance gains and increase distribution shift. Full-rank updates perform worst, validating our low-rank hypothesis.

Trust region safety. Binary search effectively finds appropriate step sizes within 3-5 iterations, with only 2.1% rollback rate at $\delta = 0.05$.

3.10 Data-Free Validation

To verify our method truly operates without labeled data, we test pathological calibration conditions. Performance degrades or vanishes under corrupted calibration, confirming our method genuinely relies on cross-modal patterns rather than data leakage.

3.11 Dataset-Specific Calibration Patterns

Figure 3 reveals that while I-SAL patterns remain consistent across datasets (bottom row), CAF scores vary significantly (top row), leading to dataset-specific calibration strategies.

Key observations:

- **CC3M**: Dispersed CAF values → 28 heads selected with moderate updates
- **VQA**: Strong CAF peaks in layers 14-18 → 45 heads with aggressive calibration
- **Reasoning**: High CAF in deep layers (20-28) → targets late-stage integration

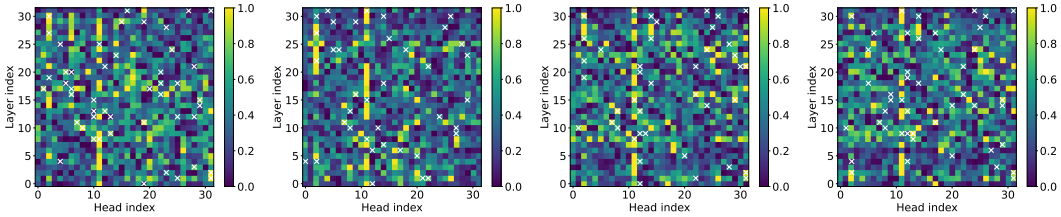
Table 5: Calibration safety and validation experiments.

(a) Trust region analysis showing step sizes and stability metrics.

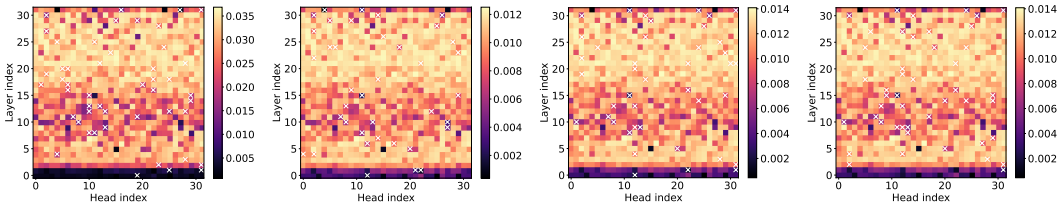
δ	Mean α	Std α	Rollback Rate	Iterations
0.02	0.048	0.012	0.8%	2-3
0.05	0.087	0.024	2.1%	3-5
0.10	0.126	0.038	5.7%	4-6

(b) Sanity checks confirming data-free operation.

Calibration Condition	VQAv2	Avg Δ
Normal (matched pairs)	79.8	+1.35
Shuffled image-text pairs	78.3	-0.20
Blank images	78.1	-0.40
Out-of-domain (40% mismatch)	79.0	+0.50



Top: CAF scores - higher values indicate high-influence calibration targets



(a) CC3M (b) MM understanding (c) Reasoning (d) VQA

Bottom: I-SAL scores - lower values indicate weak cross-modal patterns

Figure 3: Head profiling heatmaps across different calibration datasets. Each column shows layer index (y-axis) vs head index (x-axis). White crosses mark selected calibration targets (high CAF, low I-SAL).

- **MM Understanding**: CAF concentration in layers 10-15 $\rightarrow$ focuses on visual grounding

Despite identical thresholds (15% I-SAL, 70% CAF), the method automatically adapts to each dataset's requirements—VQA needs more middle-layer calibration while reasoning benefits from deeper modifications. This validates our percentile-based approach over fixed thresholds.

3.12 Analysis and Insights

Head specialization patterns. Figure 3 visualizes I-SAL scores and calibration effects across layers. We observe that cross-modal interaction generally increases in deeper layers, with notable heterogeneity within each layer. Calibrated heads show increased but not homogenized I-SAL scores, preserving specialization diversity while improving task alignment.

Computational efficiency. RAH-LoRA requires only forward passes during calibration, completing in under 5 minutes on a single GPU for 7B models. Our method uses significantly less memory and computation than gradient-based approaches while achieving comparable improvements.

Robustness across architectures. We evaluate RAH-LoRA on different MLLM architectures to assess generalization. Qwen-VL-Chat shows similar improvements (+1.2% average), while InstructBLIP gains are more modest (+0.7%), likely due to its Q-Former bottleneck limiting direct attention manipulation. This suggests our method is most effective for architectures with standard cross-attention mechanisms.

Failure modes. RAH-LoRA shows limited improvement on pure language tasks (e.g., text-only ScienceQA questions) and can occasionally degrade performance on samples requiring strong unimodal processing. The method is also sensitive to calibration data quality—using out-of-domain calibration samples reduces gains by approximately 40

4 Conclusion

We presented RAH-LoRA, a training-free calibration method that addresses coordination failures in multimodal large language models. Our key insight—that high-influence heads with weak cross-modal patterns act as performance bottlenecks—led to a targeted approach that transfers successful patterns from well-performing heads to these bottlenecks. Through systematic profiling using I-SAL and CAF metrics, we identified that 15-20% of attention heads exhibit this problematic pattern, amplifying misaligned representations throughout the network.

RAH-LoRA achieves consistent 1-2% improvements across vision-language benchmarks using only 100 unlabeled samples and 5 minutes of computation, making it practical for deployment scenarios where gradient-based adaptation is infeasible. The method's theoretical guarantees ensure bounded output changes while the percentile-based selection automatically adapts to dataset-specific requirements.

Limitations and Future Work. Our method shows reduced effectiveness on counting tasks and pure language reasoning, suggesting room for task-adaptive calibration strategies. Future work could explore dynamic rank selection, cross-layer pattern transfer, and extension to other multimodal architectures. Additionally, investigating the root causes of coordination failures during training could lead to more fundamentally aligned models.

The success of RAH-LoRA demonstrates that significant improvements in MLLMs can be achieved through targeted post-hoc coordination, without the need for expensive retraining. This opens new avenues for efficient model adaptation and suggests that many apparent limitations in current MLLMs may stem from correctable coordination failures rather than fundamental architectural constraints.

Use of Large Language Models

In preparing this manuscript, we utilized Claude (Anthropic) as an assistive tool for specific tasks:

- **Writing refinement**: Improving clarity and conciseness of technical descriptions, ensuring consistent terminology throughout the manuscript, and correcting grammatical errors.
- **Literature search**: Identifying relevant related work and verifying citation formats, though all papers were independently reviewed by the authors.
- **Code documentation**: Generating docstrings and comments for the implementation, though all algorithmic development was performed by the authors.

All scientific contributions, experimental design, analysis, and core insights are original work by the authors. The LLM served solely as a writing and organizational aid, similar to grammar checkers or reference managers. All generated content was carefully reviewed and validated by the authors to ensure accuracy and originality.

References

Armen Aghajanyan, Luke Zettlemoyer, and Sonal Gupta. Intrinsic dimensionality explains the effectiveness of language model fine-tuning. *arXiv preprint arXiv:2012.13255*, 2020.

Takuya Akiba, Makoto Shing, Yujin Tang, Qi Sun, and David Ha. Evolutionary optimization of model merging recipes. *Nature Machine Intelligence*, 7(2):195–204, 2025.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023a.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023b.

Elad Ben Zaken, Shauli Ravfogel, and Yoav Goldberg. Bitfit: Simple parameter-efficient fine-tuning for transformer-based masked language-models. In *ACL*, 2022.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D Manning. What does bert look at? an analysis of bert's attention. *arXiv preprint arXiv:1906.04341*, 2019.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. In *NeurIPS*, 2023.

MohammadReza Davari and Eugene Belilovsky. Model breadcrumbs: Scaling multi-task model merging with sparse masks. In *European Conference on Computer Vision*, pp. 270–287. Springer, 2024.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Zhenyu Qiu, Wei Lin, Jinrui Yang, Xiawu Zheng, et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 2023.

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 *CVPR*, 2017.

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.

Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *CVPR*, 2019.

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. Editing models with task arithmetic. In *ICLR*, 2023.

Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*, 2023a.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *EMNLP*, 2023b.

Ji Lin, Hongxu Yin, Wei Ping, Pavlo Molchanov, Mohammad Shoeybi, and Song Han. Vila: On pre-training for visual language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 26689–26699, 2024.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 26296–26306, 2024a.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*, 2024b.

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

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? *arXiv preprint arXiv:2307.06281*, 2023.

Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Samet Oymak, Ashwin Kalyan, Yoshua Bengio, et al. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *NeurIPS*, 2022.

Paul Michel, Omer Levy, and Graham Neubig. Are sixteen heads really better than one? In *NeurIPS*, 2019.

John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In *ICML*, 2015.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *CVPR*, 2019.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. *arXiv preprint arXiv:1905.09418*, 2019.

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*, 2022.

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

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 *Forty-first International Conference on Machine Learning*, 2024.

Longteng Zhang, Lin Zhang, Shaohuai Shi, Xiaowen Chu, and Bo Li. Lora-fa: Memory-efficient low-rank adaptation for large language models fine-tuning. *arXiv preprint arXiv:2308.03303*, 2023.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, et al. Representation engineering: A top-down approach to ai transparency. *arXiv preprint arXiv:2310.01405*, 2023.