APPENDIX

A ARCHITECTURAL GENERALIZATION

A.1 EXTENDED RESULTS ON DIFFERENT ARCHITECTURES

Table 6: RAH-LoRA performance across different MLLM architectures. Results show consistent improvements except for architectures with bottleneck designs.

Model	Size	VQAv2	TextVQA	GQA	POPE	Avg Δ
Standard Multi-Head Attention						
LLaVA-1.5	7B	+1.3	+1.4	+1.1	+1.2	+1.25
LLaVA-1.5	13B	+1.1	+1.2	+0.9	+1.3	+1.13
Qwen-VL-Chat	7B	+1.2	+1.3	+1.0	+1.1	+1.15
VILA	7B	+1.0	+1.1	+0.8	+1.0	+0.98

B DOMAIN MISMATCH ANALYSIS

B.1 CROSS-DOMAIN CALIBRATION

Table 7: Performance degradation with domain mismatch between calibration and evaluation data.

Calib \ Eval	VQA	TextVQA	GQA	SciQA
VQA	**1.35**	1.12	1.08	0.82
TextVQA	1.05	**1.42**	0.95	0.73
GQA	1.10	0.98	**1.31**	0.85
SciQA	0.75	0.68	0.71	**1.48**
CC3M	0.88	0.82	0.85	0.90

C SPECTRAL ANALYSIS OF WEIGHT UPDATES

C.1 SINGULAR VALUE ANALYSIS

Table 8: Singular value decay and variance captured across layers.

Layer Group	Decay Rate	Var @ r=4	Var @ r=8	Var @ r=16	Effective Rank
Early (0-7)	0.68	68%	82%	91%	5.2
Middle (8-15)	0.71	71%	85%	93%	6.1
Late (16-23)	0.73	73%	87%	94%	6.8
Deep (24-31)	0.75	75%	89%	95%	7.3

C.2 ANALYSIS

Exponential decay rates (0.68-0.75) confirm the low-rank structure of beneficial updates. Deeper layers show slightly higher effective rank, suggesting more complex cross-modal patterns. The 87% variance captured at r=8 (late layers) validates our default rank choice, balancing expressiveness and regularization.

Table 9: Concentration of improvements across calibrated heads.

Head Percentile	# Heads	Cum. Gain	Avg α	Avg $\|\Delta W\|_F$
Top 20%	8	65%	0.124	0.218
20-40%	8	82%	0.091	0.156
40-60%	8	91%	0.073	0.112
60-80%	8	97%	0.058	0.089
Bottom 20%	8	100%	0.042	0.065

D Per-Head Contribution Analysis

D.1 Contribution Statistics

D.2 Analysis

Pareto principle confirmed: top 20% of heads contribute 65% of gains with 3× larger updates ($\alpha = 0.124$ vs 0.042). These high-impact heads cluster in layers 14-20, corresponding to peak cross-modal pattern formation. The correlation between update magnitude and contribution (r=0.78) suggests our selection criteria effectively identify bottlenecks.

E Failure Case Analysis

E.1 Qualitative Failure Examples

Figure 4 shows representative success and failure cases. Success cases (left) demonstrate improved focus on question-relevant regions, particularly for OCR and spatial tasks. Failure cases (right) typically involve counting or pure language reasoning where dispersed attention may be beneficial.

F Implementation Details

F.1 Calibration Data Selection

- Random sampling from training split with fixed seed (42) - Balanced sampling across question types when available - Context length: maximum 256 tokens for questions - Image resolution: standard model preprocessing

Figure 4

F.2 Algorithm Pseudocode

Algorithm 1 RAH-LoRA: Representative Anchor Head Low-Rank Adaptation

Require: Model $\mathcal{M}$, unlabeled calibration data $\mathcal{D}_{\text{cal}}$
Ensure: Calibrated model $\mathcal{M}'$

// Step 1: Profile attention patterns
Compute I-SAL scores for all heads using $\mathcal{D}_{\text{cal}}$
Compute CAF scores via gradient-based importance
// Step 2: Select calibration targets
for each layer l **do**
  Identify heads with low I-SAL (bottom 10-15%)
  Filter by CAF to exclude critical auxiliary heads
  Add surviving heads to target set $\mathcal{TH}$
end for
// Step 3: Calibrate each target head
for each target head $(l, h) \in \mathcal{TH}$ **do**
  Find high-performing anchor heads in layer l
  Construct RAH via weighted aggregation of anchors
  Compute low-rank approximation of difference
  Apply update with trust-region bounded step size
end for
return $\mathcal{M}'$

G Additional Validation Experiments

H Impact of Layer Range Selection

H.1 Layer Range Analysis

Table 10: Calibration statistics and performance across different layer ranges (LLaVA-1.5-7B).

Layer Range	Layers	Targets	Calibrated	Avg Δ	Time
Early (0-7)	8	42	8	+0.65	1.2m
Middle (8-15)	8	65	18	+0.92	1.8m
Late (16-23)	8	71	23	+0.78	2.1m
Deep (24-31)	8	38	12	+0.41	1.5m
Optimal (0-15)	**16**	**107**	**26**	**+1.48**	**3.0m**
Default (12-23)	12	124	38	+1.35	3.2m
Full (0-31)	32	284	72	+1.51	8.5m

H.2 Key Findings

The optimal range (layers 0-15) achieves the best performance (+1.48%) by capturing both early visual feature extraction and initial cross-modal integration patterns. This range provides:

- **Early layers (0-7)**: Foundation visual processing that benefits from alignment
- **Middle layers (8-15)**: Critical cross-modal pattern formation where most coordination failures occur
- **Efficiency gain**: 98% of full-model performance with only 50% of layers

While the full model achieves marginally higher gains (+1.51%), the 0-15 range offers the best efficiency-performance trade-off with 2.8× faster calibration. For deployment, we recommend layers 0-15 as the optimal configuration.

H.3 Key Findings

The default range (layers 12-23) captures 73% of problematic heads while achieving 95% of full-model gains. This sweet spot emerges because: - Early layers (0-7): Primarily unimodal processing, few targets - Middle-late layers (12-23): Peak cross-modal pattern formation with coordination failures - Deep layers (24-31): Most heads already well-aligned

Full model calibration yields marginal gains (+0.07%) for 2.6× computation, confirming diminishing returns. For deployment, we recommend the default 12-23 range as optimal balance between coverage and efficiency.