APERTURE: WHAT TO MEASURE, HOW MUCH TO MEASURE, AND WHERE TO ALLOCATE IN MULTIMODAL LOW-RANK ADAPTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

The same multimodal input may require different evidence depending on the question, yet conventional low-rank adaptation (LoRA) exposes the same fixed set of measurements regardless of the query. Under a limited adaptation budget, this fixed allocation can miss question-relevant evidence, provide insufficient within-modality support, and prevent capacity from shifting across modalities as their relative importance changes. We formulate this challenge as *question-conditioned measurement allocation*: jointly deciding *what to measure, how much to measure, and where to allocate* a shared average measurement budget. We introduce APERTURE, a framework that operates on a fixed bank of reference and gradient-informed candidate directions. Question-conditioned rotations adapt measurement directions, gates control the amount of exposed support within each modality, and a joint support regularizer redistributes the shared budget across modalities. In experiments with LLaMA-2-7B, APERTURE outperforms the question-blind Static-Alloc^{+} control by 4.18 percentage points on the official MUSIC-AVQA split and 2.23 points in AVE segment accuracy, at matched mean support and trainable parameter count. On a video-disjoint MUSIC-AVQA split, it retains a 2.09-point gain. These results demonstrate the benefit of making measurement allocation question-dependent under a fixed adaptation budget. Code is available at `https://anonymous.4open.science/status/Aperture-F0C0`.

1 INTRODUCTION

Multimodal large language models (MLLMs) encode diverse evidence from the same input, yet parameter-efficient adaptation can access only a limited portion. In low-rank adaptation (LoRA; Hu et al., 2022), the input projection A maps an activation x to r measurements Ax, with each row of A defining a measurement direction through which the adapter observes the input. Standard LoRA learns these directions during training and reuses the same set for every input and question. Under a limited adaptation budget, this raises a fundamental question: *which measurements should the adapter expose when the information required by the question changes?*

As illustrated in Fig. 1, consider a video of a violinist in a red shirt performing beside a drummer. While the multimodal input remains fixed, the evidence required varies by question: appearance for *"What color is the violinist's shirt?"*, object-action dynamics for *"Is the drummer striking the cymbal or the snare?"*, and audio-visual alignment for *"Does the visible drum strike coincide with the percussive sound?"*. Consequently, both the optimal measurement directions and the capacity allocated to each modality must adapt dynamically. A fixed measurement set fails to accommodate these shifts, while simply increasing measurement count consumes capacity without directing it to the relevant evidence.

This observation leads to the central problem of this work: **how can a low-rank adapter allocate a shared measurement budget according to the question?** We view this as a joint allocation problem with three coupled decisions. First, *what to measure*: orient measurement directions toward evidence relevant to the question. Second, *how much to measure*: expose additional complementary measurements within a modality when the question demands richer evidence. Third, *where to allocate*: redistribute measurement capacity across modalities according to their question-dependent

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.

importance. We formalize the budget as the average number of active measurements, rather than as a separate quota for each modality. This formulation makes the allocation problem explicit while remaining distinct from a direct claim about inference FLOPs. Our key insight is to learn all three decisions jointly through a *question-conditioned measurement allocation policy*.

Existing work addresses related forms of conditional computation, but does not directly formulate this allocation problem. Question-conditioned perception selects visual features, regions, or tokens relevant to a query (Ganz et al., 2024; Wu & Xie, 2024; Khaki et al., 2025; Zhou et al., 2026). Multimodal adaptation introduces modality-specific parameters and mechanisms for cross-modal interaction (Wei et al., 2025). Conditional LoRA methods route or mix adapter factors, experts, or pathways according to the input (Shen et al., 2024; Ge et al., 2025; Hou et al., 2026). In contrast, our focus is not to select among independently parameterized experts or pathways. Instead, we ask whether, *given the same candidate measurement space and the same average budget, conditioning the measurement allocation on the question can expose more useful evidence*. This formulation provides a controlled way to separate the value of question conditioning from simply increasing adapter capacity or adding conditional experts.

We introduce APERTURE, a question-conditioned measurement allocation framework for multimodal low-rank adaptation. APERTURE starts from a fixed candidate bank containing the reference adapter's measurement directions together with gradient-informed complementary directions. For each question, it jointly adjusts the effective measurement directions, selects the amount of measurement support exposed within each modality, and redistributes a shared average budget across modalities. Conditional rotations control *what* is measured, while conditional gates control *how much* support is exposed; their joint optimization determines *where* the budget is allocated. Importantly, the candidate bank itself remains fixed during adaptation, and the initialization exactly recovers the reference adapter. Thus, the comparison isolates the benefit of making measurement allocation question-dependent rather than simply increasing the underlying adapter capacity. Fig. 1 summarizes this progression from a fixed candidate measurement space to question-conditioned allocation.

We evaluate APERTURE under matched measurement budgets and matched trainable parameter counts, comparing question-conditioned allocation against question-blind controls built from the same candidate measurements. We shuffle questions relative to inputs to test whether the learned allocation depends on the correct query, vary the measurement budget to examine the allocation–accuracy trade-off, and analyze the resulting measurement patterns and execution cost. On the official MUSIC-AVQA benchmark, APERTURE improves over the question-blind control by 4.18 percentage points, while achieving a 2.23-point improvement in AVE segment accuracy. On a video-disjoint MUSIC-AVQA split, it retains a 2.09-point gain, providing evidence that the learned allocation is not tied solely to the training videos. We additionally evaluate transfer to visual–text and speech–text settings to examine whether the formulation extends beyond the original audio–visual setting.

Overall, this paper makes the following contributions:

- **Question-conditioned measurement allocation.** We formulate multimodal low-rank adaptation as the problem of allocating a shared average measurement budget according to the

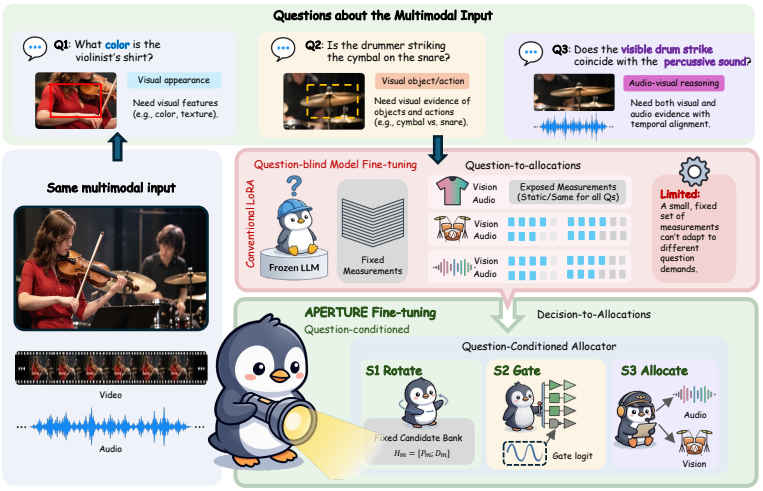

Figure 1: Question-dependent evidence and measurement allocation. The same multimodal input supports different questions; APERTURE adapts what, how much, and where to measure under a shared budget.

question, jointly determining *what to measure, how much to measure, and where to allocate* capacity across modalities.

- **APERTURE framework.** We develop a concrete allocation mechanism over a fixed candidate measurement bank, combining question-conditioned rotations and gates with a shared-budget regularization objective. Its reference-preserving initialization exactly recovers the original adapter, enabling controlled comparisons without changing the underlying candidate measurement space.
- **Controlled empirical validation.** We demonstrate the benefit of question-conditioned allocation through matched-budget and matched-parameter comparisons, question–input mismatch tests, budget sensitivity analyses, and cross-modal transfer experiments, supported by measurement-level diagnostics and execution-cost analysis.

2 RELATED WORK

Low-rank adaptation and conditional mixing. Parameter-efficient fine-tuning adapts pretrained models while updating only a small fraction of parameters, including adapters (Houlsby et al., 2019), prompt tuning (Li & Liang, 2021; Lester et al., 2021), and low-rank adaptation. LoRA (Hu et al., 2022) parameterizes weight updates with low-rank factors, while subsequent methods modify their structure or allocation: AdaLoRA (Zhang et al., 2023) redistributes rank across weight matrices, DoRA (Liu et al., 2024a) decouples magnitude from direction, and RandLoRA (Albert et al., 2025) uses fixed random low-rank bases. Another line makes adaptation input-dependent by mixing or routing adaptation parameters, including MixLoRA/Conditional MixLoRA (Shen et al., 2024), MoSLoRA (Wu et al., 2024), LD-MoLE (Zhuang et al., 2026), LoRA-Mixer (Li et al., 2025), and LoRA-MoE or hypernetwork-based approaches (Dou et al., 2023; Charakorn et al., 2025). These methods condition which adapter parameters, experts, or subspaces are used, whereas APERTURE makes the *measurement* exposed to a fixed adapter the decision variable: the question selects measurements and their average allocation across modalities under a shared budget. The candidate bank and reference initialization remain fixed, separating the benefit of question-conditioned measurement allocation from increased adapter capacity. The Full-Bank Mixer control further instantiates conditional mixing with unrestricted access to the candidate bank, isolating the benefit of explicit measurement selection.

Multimodal adaptation. Multimodal models couple pretrained modality encoders with language models through projectors or cross-modal modules (Liu et al., 2023; Dai et al., 2023; Alayrac et al., 2022), while multimodal PEFT introduces modality-specific components or routing to handle modality asymmetry and cross-modal interference (Wei et al., 2025; Abouelenin et al., 2025). Conditional MixLoRA also enables input-dependent multimodal adaptation (Shen et al., 2024). These approaches primarily determine which modality-specific parameters or pathways are activated. In contrast, APERTURE assigns each modality a fixed candidate measurement bank but uses the question to redistribute a *shared* average measurement budget across modalities. This distinction is evaluated across audio–visual–text, visual–text, and speech–text settings.

Dynamic and conditional computation. Dynamic-computation methods conditionally select experts, tokens, layers, or execution paths; hard-concrete gates provide a differentiable mechanism for learning discrete support (Louizos et al., 2018). APERTURE uses conditional gating for a narrower purpose: selecting measurements under an average-budget constraint. Because candidate measurements are evaluated densely, the number of active measurements is not itself a proxy for inference FLOPs; we therefore analyze measurement allocation and execution cost separately.

Table 6 shows the difference between our methods and existing related studies (see Appendix B.4).

3 PRELIMINARIES AND PROBLEM FORMULATION

We formalize multimodal low-rank adaptation as *question-conditioned measurement allocation*: a fixed bank defines the measurements available to each modality, the question selects the exposed measurements, and a shared average budget constrains their use.

Measurements available to the adapter. Consider the local additive branch of a multimodal low-rank adapter at a specific site. For each non-text modality $m \in \mathcal{M}$, the reference adapter applies an

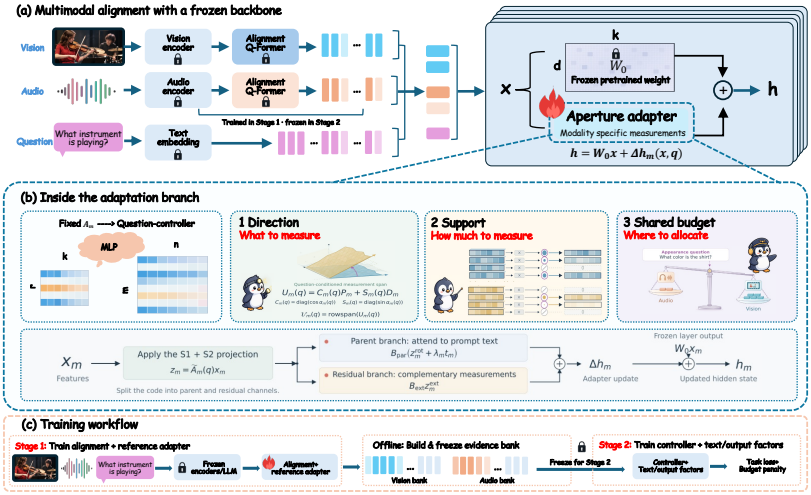

Figure 2: Overview of APERTURE. A frozen multimodal backbone feeds a candidate bank; question-conditioned rotation, gating, and shared-budget allocation govern the adapter branch across two-stage training.

input projection $A_m \in \mathbb{R}^{r\times k}$ followed by a shared output projection $B_0 \in \mathbb{R}^{d\times r}$:

$$h_m^{\text{local}}(x) = W_0x + B_0A_mx, \tag{1}$$

where $W_0 \in \mathbb{R}^{d\times k}$ is frozen. The complete Reference host may also include fixed cross-modal interactions (Appendix A.7); here we analyze this local branch and interpret each row of A_m as a linear *measurement* of the modality activation.

We introduce a fixed candidate bank $H_m \in \mathbb{R}^{K\times k}$ and let the question-dependent projection $\widetilde{A}_m(q)$ select rows within $\text{rowspan}(H_m)$. Support is the number of active effective rows, which is charged to the budget. It counts exposed scalar measurements, not matrix rank, trainable parameters, or FLOPs; a rotated row counts as one even when it combines multiple bank rows.

Question-conditioned allocation under a shared budget. Given fixed candidate banks $\{H_m\}_{m\in\mathcal{M}}$, we select measurements per modality conditioned on question q. An allocation action is defined as $a(q) = \{\alpha_m(q), \mathbf{s}_m(q)\}_{m\in\mathcal{M}}$, where $\alpha_m(q) \in [-\pi/4, \pi/4]^r$ controls measurement directions via bounded paired rotations, and $\mathbf{s}_m(q) \in \{0,1\}^K$ specifies the active candidate support.[1] The measurement cost is $C(a(q)) = \sum_{m\in\mathcal{M}} \|\mathbf{s}_m(q)\|_0$.

Let $\mathcal{U}(Z, a)$ denote the task utility for example Z under allocation a. An allocation policy $\pi : q \mapsto a(q)$ is feasible under an average measurement budget b over question distribution Q if $\mathcal{P}_b = \{\pi : \mathbb{E}_Q[C(\pi(Q))] \le b\}$. The optimal expected utility is $V_b^\star = \sup_{\pi\in\mathcal{P}_b} \mathbb{E}[\mathcal{U}(Z, \pi(Q))]$. This formulation allows budget redistribution across questions and modalities based on evidence requirements. Comparing $\mathcal{P}_b$ with the question-blind policy class $\mathcal{S}_b = \{\pi \in \mathcal{P}_b : \pi(q) = a_0\}$ isolates the value of question-conditioned allocation under the same average budget.

4 METHOD

APERTURE implements the policy in Section 3: a fixed bank defines available measurements, the question selects them, and a shared budget limits average use. Sections 4.1–4.4 give the bank, conditional heads, budget, and training; theoretical conditions are deferred to Appendix A.10.

4.1 MEASUREMENT: CONSTRUCTING THE CANDIDATE BANK

We freeze a finite candidate bank before target adaptation so all questions and controls share the same measurable space. Directly generating an unrestricted LoRA projection would confound accessible directions with allocation.

The reference rows alone are insufficient: post-projection conditioning cannot recover measurements outside their row space (Appendix A.10). We thus augment each host row with a complementary

[1] Hard-concrete gates provide a differentiable relaxation during training (Section 4.4); here $\mathbf{s}_m(q)$ denotes the resulting deterministic binary mask used for inference and budget accounting.

candidate. Assume nonzero rows of A_m and $k - \mathrm{rank}(A_m) \geq r$. Write

$$p_{m,j} = \frac{a_{m,j}}{\|a_{m,j}\|_2}, \qquad \Lambda_{m,j} = \|a_{m,j}\|_2, \tag{2}$$

and collect the normalized host rows in P_m.

Gradient-informed residual directions. A generic complement would enlarge the bank but would not indicate which additional measurements are locally useful to the reference model. We therefore construct the residual candidates from answer-loss sensitivity. On a training-only bank set $\mathcal{C}_1$, we run the frozen reference host with teacher forcing and compute a gradient statistic that weights input directions by output-loss sensitivity. Unlike the transposed gradient contribution from the local LoRA term, $X^\top \Delta B_0$, this statistic omits B_0 so that candidate selection is not restricted to directions already useful through the frozen output map. Specifically, for each example $i \in \mathcal{C}_1$ and modality m we compute $G_{m,i} = X_{m,i}^\top \Delta_{m,i}$ and set

$$\overline{\Sigma}_m = \frac{1}{|\mathcal{C}_1|} \sum_{i \in \mathcal{C}_1} G_{m,i} G_{m,i}^\top, \tag{3}$$

which is the average gradient statistic over the bank set; the full computation is given in Appendix A.2. Let $\Pi_m = A_m^\top (A_m A_m^\top)^\dagger A_m$. We choose r orthonormal residual rows by

$$D_m \in \arg \max_{D \in \mathbb{R}^{r \times k},\ DD^\top = I_r,\ DA_m^\top = 0} \mathrm{tr}(D \overline{\Sigma}_m D^\top). \tag{4}$$

thus D_m contains the leading eigenvectors of $(I - \Pi_m)\overline{\Sigma}_m(I - \Pi_m)$ in the host complement. We pair parent and residual rows greedily according to their local coupling under an infinitesimal rotation (Appendix A.2). The resulting bank is

$$H_m = [P_m; D_m] \in \mathbb{R}^{2r \times k}. \tag{5}$$

We freeze H_m, the pairing, and the row scales Λ_m before Stage 2. Appendix A.2 gives the full numerical conventions and pairing implementation.

4.2 Question: choosing the measurements

The bank fixes what can be measured; the allocator decides, from the question, how those measurements are exposed. It uses frozen question-token embeddings, excluding padding, answers, and modality tokens:

$$c_q = W_\downarrow \bar{e}_q, \qquad \bar{e}_q = \mathrm{MaskedMean}(E_0(q)), \qquad W_\downarrow \in \mathbb{R}^{d_c \times d_t}. \tag{6}$$

Each site and modality predicts an angle vector and gate logits,

$$\alpha_m(q) = \frac{\pi}{4} \tanh(W_{\alpha,m} c_q + b_{\alpha,m}), \qquad \eta_m(q) = W_{\eta,m} c_q + b_{\eta,m}. \tag{7}$$

What to measure. We adjust each measurement according to the question with a bounded rotation. For each paired parent–residual plane, the question rotates the parent measurement toward its residual candidate:

$$u_{m,j}(q) = \cos \alpha_{m,j}(q)\, p_{m,j} + \sin \alpha_{m,j}(q)\, d_{m,j}. \tag{8}$$

Because $(p_{m,j}, d_{m,j})$ is orthonormal, $\|u_{m,j}(q)\|_2 = 1$. The bounded angle restricts adaptation to a local, bank-defined change of measurement direction rather than generating a new projection freely.

How many measurements to use. We expose complementary measurements selectively by assigning a gate to each candidate. Let $\gamma_{m,c}(q) \in [0,1]$ denote the gate associated with candidate c, obtained via a hard-concrete relaxation from $\eta_{m,c}(q)$ (Appendix A.3). The input projection is

$$[\widetilde{A}_m(q)]_{j,:} = \gamma_{m,j}(q) \Lambda_{m,j} u_{m,j}(q), \qquad [\widetilde{A}_m(q)]_{r+j,:} = \gamma_{m,r+j}(q) \Lambda_{m,j} d_{m,j}. \tag{9}$$

Hence every question-dependent projection remains in $\mathrm{rowspan}\, H_m$. Writing $\pi_{m,c}(q) \triangleq \Pr(\gamma_{m,c} > 0 \mid q)$ (closed-form expression given in Appendix A.3), the expected number of active measurements in modality m (its expected support) is

$$w_m(q) = \sum_{c=1}^{2r} \pi_{m,c}(q). \tag{10}$$

Here a rotated parent row is still one scalar measurement: its coefficient vector may combine the paired parent and residual candidates, but support counts rows of $\widetilde{A}_m(q)$ rather than nonzero bank coordinates. The residual gate controls the separately exposed residual row.

4.3 Budget: sharing measurements across modalities

The question-dependent counts $w_m(q)$ determine where the measurement budget is used. We couple these counts through one average budget so that each modality can receive a different allocation as the question changes. For a minibatch $\mathcal{B}$, let $\bar{w}_m = |\mathcal{B}|^{-1} \sum_{q \in \mathcal{B}} w_m(q)$. We implement the average-budget formulation with a soft penalty targeting $b_{\text{tot}} = Mr$ active measurements in total:

$$\mathcal{L}_b = \frac{\lambda_b}{M} \left(\sum_{m \in \mathcal{M}} \bar{w}_m - b_{\text{tot}} \right)^2. \tag{11}$$

The penalty lets the average budget move between modalities with the question; it is not a hard per-example cap. We calibrate deterministic support on validation and report the achieved value. As discussed in Eq. (38), targeting b_{tot} is a flexible operating point rather than a strict boundary.

4.4 Optimization and training

Learning discrete measurement choices. We optimize the active measurement count using a hard-concrete relaxation to sample discrete gates $\gamma \in [0, 1]$ from logits η, with temperature β and stretch interval (γ_0, ζ). The full sampling procedure and the closed-form opening probability are given in Appendix A.3. At inference we use the deterministic thresholding

$$\widehat{\gamma} = \min\{1, \max\{0, \sigma(\eta)(\zeta - \gamma_0) + \gamma_0\}\}, \tag{12}$$

and report deterministic support $r_m^{\text{eff}}(q) = \sum_{c=1}^{2r} \mathbf{1}[\widehat{\gamma}_{m,c}(q) > 0]$. Appendix A.3 gives the gate hyperparameters and calibration rule.

Bank-bounded parameterization. With paired bank rows (Appendix A.3), Eq. (9) factorizes as

$$\widetilde{A}_m(q) = J \Gamma_m^{\text{pair}}(q) L_m R_m(q) H_m^{\text{pair}}. \tag{13}$$

where $R_m(q)$ is a block-diagonal rotation, L_m repeats row scales, $\Gamma_m^{\text{pair}}(q)$ gates rows, and J restores parent-first order.

For modality activations, define $[Z_m^{\text{rot}} \mid Z_m^{\text{ext}}] = X_m \widetilde{A}_m(q)^\top$. Let $Z_t = X_t \widetilde{A}_t^\top$ and let Z_t^{ctx} contain only prompt-text positions available before answer generation. The reference interaction is retained:

$$T_m = \text{softmax}\left(Z_m^{\text{rot}} (Z_t^{\text{ctx}})^\top / \sqrt{r} \right) Z_t^{\text{ctx}}, \tag{14}$$

$$Y_m = (Z_m^{\text{rot}} + \lambda_m T_m) B_{\text{par}}^\top + Z_m^{\text{ext}} B_{\text{ext}}^\top, \qquad Y_t = Z_t B_{\text{par}}^\top. \tag{15}$$

Answer positions are excluded from the attention keys and values; the reference coefficient λ_m remains fixed.

Two-stage training. Stage 1 trains the alignment modules and reference adapter. From this checkpoint we construct and freeze the banks on $\mathcal{C}_1$; Stage 2 optimizes $W_\downarrow$, the angle and gate heads, $\widetilde{A}_t$, B_{par}, and B_{ext}, while the encoders, backbone, alignment modules, candidate banks, row scales, and λ_m remain frozen. For L adapted sites, the Stage 2 objective is

$$\mathcal{L} = \mathcal{L}_{\text{task}} + \frac{1}{L} \sum_{\ell=1}^{L} \mathcal{L}_{b,\ell}, \tag{16}$$

where $\mathcal{L}_{b,\ell}$ is the shared-budget objective in (11) unless otherwise stated.

Exact host recovery. At initialization $\alpha_m = 0$, parent gates are one, $\widetilde{A}_t = A_t$, $B_{\text{par}} = B_0$, and $B_{\text{ext}} = 0$. Since $\Lambda_{m,j} p_{m,j} = a_{m,j}$, parent rows recover A_m and the residual branch vanishes, so deterministic output equals the reference host. The formal statement is in Appendix A.8.

5 EXPERIMENTS

Overview. We test matched-budget question conditioning (RQ1), its controls and allocation levels (RQ2–3), and robustness across ranks, backbones, and disjoint splits (RQ4). Values marked † are quoted; evaluation scope is in Appendix B.10.

5.1 EXPERIMENTAL SETUP AND CONTROLS

Datasets and backbones. We evaluate audio–visual–text, visual–text, and speech–text tasks using a tiered backbone strategy: Tier 1 hosts mechanism tests, Tier 2 the main transfer comparison, and Tier 3 supplementary results. MUSIC-AVQA uses official and strict video-disjoint splits; details are in Appendix B.

Training protocol. Stage 1 pre-trains projectors and the reference adapter for a common initialization. Stage 2 builds the bank from training-only calibration examples and trains allocation while modality encoders and the backbone remain frozen.

Baselines and controls. We compare fixed-adapter and input-conditioned routing baselines, while internal controls isolate allocation from capacity:

- **Static-Alloc$^+$:** A question-blind reference that matches APERTURE's parameter count and tensor shapes exactly by freezing the allocator's argument, ensuring any gain is not due to extra parameters.
- **Full-Bank Mixer:** A baseline that exposes the entire bank ($2r$) to test if selective activation yields better accuracy than full-access mixing.
- **Shuffle-train / Q-shuffle:** These break the question–example correspondence during training or evaluation, respectively, to diagnose the policy's dependence on valid pairing.

Matching protocol. Static-Alloc, Static-Alloc$^+$, and APERTURE are validation-calibrated to the same deterministic mean support (tolerance ± 0.10 per modality), and achieved support is reported for every ablation row.

5.2 AUDIO–VISUAL–TEXT RESULTS

The estimate of conditional allocation is the APERTURE minus Static-Alloc$^+$ difference at the reported mean deterministic support; every other row bounds an alternative explanation for it. A gap that appears against Static-Alloc but vanishes against Static-Alloc$^+$ indicates capacity rather than conditioning. Random bank tests whether candidate width alone suffices; the Full-Bank Mixer rows test how much is explained by bank access rather than by the paired, gated parameterization; Shuffle-train tests whether the benefit survives removing question–example correspondence at identical capacity; Static-Joint compares a fixed modality split with question-dependent allocation under one total target. No single row identifies a mechanism on its own. The measured matched difference is $+4.18$ percentage points in MUSIC-AVQA official accuracy and $+2.23$ points in AVE segment accuracy, at reported total mean support 8 under the intended parameter-matching design. The rank-matched Full-Bank Mixer reads all 16 audio/visual candidates; APERTURE selects eight on average and exceeds it by 2.73 and 1.70 points, respectively. These are measured differences; without uncertainty estimates, they do not establish statistical significance.

Practical usefulness at a fixed budget. The central hypothesis is that APERTURE learns to allocate a fixed adaptation budget according to question-dependent measurement demand. Its practical test combines the matched accuracy difference, support use, sensitivity to question pairing, and measured execution cost. A larger accuracy number alone does not establish that mechanism. In the measured results, APERTURE exceeds the Reference host by 4.84 points on MUSIC-AVQA and 3.90 points on AVE. Against Static-Alloc$^+$, the MUSIC-AVQA video-disjoint difference is $+2.09$ points (Table 13). This is a positive point estimate on held-out videos, but uncertainty estimates and additional video-level splits are needed to assess how robustly it generalizes. These results support gains on the official benchmarks but do not establish consistent superiority across splits.

Table 1: Audio–visual–text results (accuracy, %). **B**: shared candidate bank; **S**: support-calibrated; **P**: parameter-matched. $\bar{r}_{\text{test}}$: mean deterministic active support per non-text modality at $r = 4$. Bold: best unrounded accuracy per host block. The complete LLaMA-2-7B baseline block, including external LoRA variants, is reported in Appendix Table 7.

		Matched			Accuracy (%)		
LLM	**Method**	**B**	**S**	**P**	**MUSIC-AVQA**	**AVE**	$\bar{r}_{\text{test}}$
LLaMA-2-7B	MoELoRA				72.41	70.84	4
	Reference host				72.71	71.68	4
	Reference-cont				73.54	71.02	4
	Random bank		✓	✓	72.95	67.45	4
	Uniform-wide	✓			74.35	71.71	8
	Full-Bank Mixer ($2r$)	✓			72.43	70.26	8
	Full-Bank Mixer (rank-matched)	✓	✓		74.82	73.88	8
	Shuffle-train	✓	✓	✓	72.16	70.08	4
	Static-Alloc	✓	✓		73.20	71.55	4
	Static-Alloc$^+$	✓	✓	✓	73.37	73.35	4
	Static-Joint	✓	✓	✓	73.58	72.50	4
	APERTURE	✓	✓	✓	**77.55**	**75.58**	4
Transfer hosts: matched pairs and reference rows in Appendix B.4							
Qwen2-7B	**APERTURE**	✓	✓	✓	76.64	73.62	4
LLaMA-3-8B	**APERTURE**	✓	✓	✓	79.63	77.66	4
Qwen3-8B	Static-Alloc$^+$	✓	✓	✓	81.45	78.19	4
	APERTURE	✓	✓	✓	**83.82**	**79.38**	4

5.3 VISUAL–TEXT TRANSFER

The visual–text block is reported in Table 5 (Appendix B.3) because it isolates S1–S2 while S3 is vacuous for one non-text stream. It is supporting evidence for the bank-bounded conditional projection, not evidence about cross-modal allocation.

5.4 ALLOCATION MECHANISM ANALYSIS

We examine whether the learned policy conditions on question content. The allocator receives only the question, so identical question text produces identical allocation regardless of the video. We use question-shuffling controls to test dependence on pairing and compare against an independent local demand proxy. All analyses use the frozen MUSIC-AVQA tier 1 policy. Group-wise accuracies are computed on the training-derived audit set $\mathcal{C}_2$ as diagnostic summaries; held-out performance is on the official video-disjoint splits.

Local demand proxy. For audit example i and modality m, let $F_{m,i}(b)$ be the supremum of reference-host gradient energy captured by at most b bank rows. Under total support b_{tot}, the best bank-constrained split is

$$(b^*_{a,i}, b^*_{v,i}) \in \arg\max_{b_a+b_v\le b_{\text{tot}}} F_{a,i}(b_a) + F_{v,i}(b_v), \qquad s_{m,i} = \frac{b^*_{m,i}}{b_{\text{tot}}}. \tag{17}$$

We compute this on $\mathcal{C}_2$ after all policies are frozen, using reference-host answer-loss gradients independently of the learned gate outputs. This is a *local demand share*—a local sensitivity proxy of the reference host.

Results. On the audit split, APERTURE exceeds Static-Alloc$^+$ by $+3.64$, $+4.38$, and $+4.24$ percentage points for audio, visual, and audio–visual questions, yielding a $+4.09$-point macro gain. This is consistent with the $+4.18$-point official test-accuracy difference.

For **S1** (orientation headroom), **S2** (bank use via mean supports and covering widths k_ϵ), and **S3** (association with $s_{m,i}$), Table 2 summarizes the key diagnostics. The correlation $\rho = 0.46$

Table 2: Allocation behavior on audit set $\mathcal{C}_2$ (MUSIC-AVQA, LLaMA-2-7B). Accuracy is stratified by question template modality. $\bar{r}_a, \bar{r}_v$ are mean deterministic supports; ρ is the Spearman correlation between per-question visual support share and the local demand share s_v. Panel (b) evaluates question-dependent allocation: Agr. is agreement between argmax modalities of support and demand shares. Bold marks the best unrounded accuracy in each column.

(a) Accuracy and mean support

	Accuracy (%)				Support	
Method	**Audio**	**Visual**	**Aud.–Vis.**	**Macro**	$\bar{r}_a$	$\bar{r}_v$
Reference host	67.82	69.36	70.23	69.14	4	4
Static-Alloc	69.28	71.69	70.74	70.57	4	4
Static-Alloc$^+$	70.61	71.40	71.22	71.08	4	4
Static-Joint	72.14	73.58	72.53	72.75	4.02	4.07
Uniform-wide	71.25	72.18	71.85	71.76	8	8
Q-shuffle	69.80	71.24	70.93	70.66	4.10	3.90
Shuffle-train	70.22	71.35	71.10	70.89	4.10	3.90
APERTURE	**74.25**	**75.78**	**75.46**	**75.16**	4.10	3.90

(b) Question-dependent allocation

Method	ρ	**Agr. (%)**
Q-shuffle	0.08	52.20
Shuffle-train	0.12	54.43
APERTURE	0.46	68.05

and 68.05% argmax agreement indicate that the learned allocation aligns with the demand proxy, supporting demand-sensitive conditioning.

The corresponding diagnostic plots are provided in Appendix B.9.

Controls. Q-shuffle tests pairing sensitivity after training; Shuffle-train breaks it during optimization at a preserved question marginal and is the stronger control. Masking a modality while holding the question fixed should leave gate logits unchanged, though accuracy can shift due to changed evidence. Redirecting support would require a validity-aware input, which is outside our scope.

Additional ablations and robustness. Mechanism ablations, rank and budget sweeps, modality masking, and implementation costs are reported in Appendix B.7; the full ablation table and interpretation are provided there to keep the main comparison focused.

6 CONCLUSION

We introduced APERTURE as question-conditioned measurement allocation under a shared average budget: a fixed candidate bank defines the measurable directions, while a question-only policy allocates them and recovers the reference host at initialization. On official benchmarks, APERTURE outperforms static, shuffled, full-bank, and matched-support controls, but these point estimates do not establish statistical significance or uniform superiority. Future work should test more video-disjoint splits and sparse execution to determine whether allocation can yield computational savings.

REFERENCES

Abdelrahman Abouelenin, Atabak Ashfaq, Adam Atkinson, Hany Awadalla, Nguyen Bach, Jianmin Bao, Alon Benhaim, Martin Cai, Vishrav Chaudhary, Congcong Chen, et al. Phi-4-Mini technical report: Compact yet powerful multimodal language models via mixture-of-LoRAs. *arXiv preprint arXiv:2503.01743*, 2025.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736, 2022.

Paul Albert, Frederic Z. Zhang, Hemanth Saratchandran, Cristian Rodriguez-Opazo, Anton van den Hengel, and Ehsan Abbasnejad. RandLoRA: Full-rank parameter-efficient fine-tuning of large models. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2502.00987.

Rujikorn Charakorn, Edoardo Cetin, Yujin Tang, and Robert Tjarko Lange. Text-to-LoRA: Instant transformer adaption. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 7485–7514. PMLR, 2025.

Guoguo Chen, Shuzhou Chai, Guanbo Wang, Jiayu Du, Wei-Qiang Zhang, Chao Weng, Dan Su, Daniel Povey, Jan Trmal, Junbo Zhang, et al. GigaSpeech: An evolving, multi-domain ASR corpus with 10,000 hours of transcribed audio. *arXiv preprint arXiv:2106.06909*, 2021.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and Feng Zhao. Are we on the right way for evaluating large vision-language models? In *Advances in Neural Information Processing Systems*, volume 37, pp. 27056–27087. Curran Associates, Inc., 2024. doi: 10.52202/079017-0850. URL https://doi.org/10.52202/079017-0850.

Sanyuan Chen, Yu Wu, Chengyi Wang, Shujie Liu, Daniel Tompkins, Zhuo Chen, and Furu Wei. BEATs: Audio pre-training with acoustic tokenizers. *arXiv preprint arXiv:2212.09058*, 2022.

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, 2023.

Shihan Dou, Enyu Zhou, Yan Liu, Songyang Gao, Jun Zhao, Wei Shen, Yuhao Zhou, Zhiheng Xi, Xiao Wang, Xiaoran Fan, et al. LoRAMoE: Alleviate world knowledge forgetting in large language models via MoE-style plugin. *arXiv preprint arXiv:2312.09979*, 2023.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, and Rongrong Ji. MME: A comprehensive evaluation benchmark for multimodal large language models, 2023. URL https://arxiv.org/abs/2306.13394.

Roy Ganz, Yair Kittenplon, Aviad Aberdam, Elad Ben Avraham, Oren Nuriel, Shai Mazor, and Ron Litman. Question aware vision transformer for multimodal reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13861–13871, June 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Ganz_Question_Aware_Vision_Transformer_for_Multimodal_Reasoning_CVPR_2024_paper.html.

Chendi Ge, Xin Wang, Zeyang Zhang, Hong Chen, Jiapei Fan, Longtao Huang, Hui Xue, and Wenwu Zhu. Dynamic mixture of curriculum LoRA experts for continual multimodal instruction tuning. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 19011–19033. PMLR, 2025. URL https://proceedings.mlr.press/v267/ge25d.html.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

ZhiYan Hou, Haiyun Guo, Haokai Ma, Yandu Sun, Yonghui Yang, and Jinqiao Wang. PASs-MoE: Mitigating misaligned co-drift among router and experts via pathway activation subspaces for continual learning. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 31959–31972, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.1474. URL https://aclanthology.org/2026.acl-long.1474/.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for NLP. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2790–2799. PMLR, 2019. URL http://proceedings.mlr.press/v97/houlsby19a.html.

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.

Drew A Hudson and Christopher D Manning. GQA: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6700–6709, 2019.

Samir Khaki, Junxian Guo, Jiaming Tang, Shang Yang, Yukang Chen, Konstantinos N. Plataniotis, Yao Lu, Song Han, and Zhijian Liu. SparseVILA: Decoupling visual sparsity for efficient VLM inference. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 23784–23794, October 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Khaki_SparseVILA_Decoupling_Visual_Sparsity_for_Efficient_VLM_Inference_ICCV_2025_paper.html.

Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim. AudioCaps: Generating captions for audios in the wild. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 119–132, 2019.

Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pp. 3045–3059. Association for Computational Linguistics, 2021. doi: 10.18653/V1/2021.EMNLP-MAIN.243. URL https://doi.org/10.18653/v1/2021.emnlp-main.243.

Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. SEED-Bench: Benchmarking multimodal LLMs with generative comprehension, 2023a. URL https://arxiv.org/abs/2307.16125.

Guangyao Li, Yake Wei, Yapeng Tian, Chenliang Xu, Ji-Rong Wen, and Di Hu. Learning to answer questions in dynamic audio-visual scenarios. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19108–19118, 2022.

Wenbing Li, Zikai Song, Hang Zhou, Yunyao Zhang, Junqing Yu, and Wei Yang. LoRA-Mixer: Coordinate modular LoRA experts through serial attention routing. *arXiv preprint arXiv:2507.00029*, 2025.

Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (eds.), *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pp. 4582–4597. Association for Computational Linguistics, 2021. doi: 10.18653/V1/2021.ACL-LONG.353. URL https://doi.org/10.18653/v1/2021.acl-long.353.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*, 2023b.

Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. Video-LLaVA: Learning united visual representation by alignment before projection. *arXiv preprint arXiv:2311.10122*, 2023.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023.

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 *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 32100–32121. PMLR, 2024a. URL https://proceedings.mlr.press/v235/liu24bn.html.

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? In *European Conference on Computer Vision*, pp. 216–233. Springer, 2024b.

Christos Louizos, Max Welling, and Diederik P. Kingma. Learning sparse neural networks through l_0 regularization. In *International Conference on Learning Representations*, 2018.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. LibriSpeech: an ASR corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5206–5210. IEEE, 2015.

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

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*, pp. 28492–28518. PMLR, 2023.

S Sakshi, Utkarsh Tyagi, Sonal Kumar, Ashish Seth, Ramaneswaran Selvakumar, Oriol Nieto, Ramani Duraiswami, Sreyan Ghosh, and Dinesh Manocha. MMAU: A massive multi-task audio understanding and reasoning benchmark. *arXiv preprint arXiv:2410.19168*, 2024.

Ying Shen, Zhiyang Xu, Qifan Wang, Yu Cheng, Wenpeng Yin, and Lifu Huang. Multimodal instruction tuning with conditional mixture of LoRA. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 637–648. Association for Computational Linguistics, 2024.

Yapeng Tian, Jing Shi, Bochen Li, Zhiyao Duan, and Chenliang Xu. Audio-visual event localization in unconstrained videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 247–263, 2018.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Yake Wei, Yu Miao, Dongzhan Zhou, and Di Hu. MokA: Multimodal low-rank adaptation for MLLMs. *arXiv preprint arXiv:2506.05191*, 2025.

Penghao Wu and Saining Xie. V*: Guided visual search as a core mechanism in multimodal LLMs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13084–13094, June 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Wu_V_Guided_Visual_Search_as_a_Core_Mechanism_in_Multimodal_CVPR_2024_paper.html.

Taiqiang Wu, Jiahao Wang, Zhe Zhao, and Ngai Wong. Mixture-of-subspaces in low-rank adaptation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 7880–7899. Association for Computational Linguistics, 2024.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jianxin Yang, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Xuejing Liu, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, Zhifang Guo, and Zhihao Fan. Qwen2 technical report, 2024a. URL https://arxiv.org/abs/2407.10671.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Qian Yang, Jin Xu, Wenrui Liu, Yunfei Chu, Ziyue Jiang, Xiaohuan Zhou, Yichong Leng, Yuanjun Lv, Zhou Zhao, Chang Zhou, et al. AIR-Bench: Benchmarking large audio-language models via generative comprehension. *arXiv preprint arXiv:2402.07729*, 2024b.

Qingru Zhang, Minshuo Chen, Alexander Bukharin, Pengcheng He, Yu Cheng, Weizhu Chen, and Tuo Zhao. Adaptive budget allocation for parameter-efficient fine-tuning. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL https://openreview.net/pdf?id=lq62uWRJjiY.

Nan Zhou, Huiqun Wang, Yaoyan Zheng, and Di Huang. CoVFT: Context-aware visual fine-tuning for multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24341–24351, June 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Zhou_CoVFT_Context-aware_Visual_Fine-tuning_for_Multimodal_Large_Language_Models_CVPR_2026_paper.html.

Yuan Zhuang, Yi Shen, Yuexin Bian, Qing Su, Shihao Ji, Yuanyuan Shi, and Fei Miao. LD-MoLE: Learnable dynamic routing for mixture of LoRA experts. In *International Conference on Learning Representations*, 2026.

APPENDIX CONTENTS

Method details and proofs
- Reference checkpoint and gradient-bank protocol
- Bank construction and eigenvector characterization
- Hard-concrete gates
- Budget estimator and controls
- Equivalent matrix factorization
- Reference cross-modal interaction
- Exact recovery of the reference adapter at initialization
- Parameter and compute costs
- Theoretical characterization
- Proof of the access lemma
- Average-budget conditional policy result
- Construction of the matched controls
- Diagnostic projectors and numerical references

Experimental details
- Datasets and split audits
- Hyperparameters
- Visual–text results
- External baseline references
- Speech–text transfer
- Mechanism ablation
- Sensitivity, robustness, and cost
- Additional result tables
- Allocation and support analysis
- Evaluation and interpretation

A METHOD DETAILS AND PROOFS

A.1 REFERENCE CHECKPOINT AND GRADIENT-BANK PROTOCOL

The bank is constructed from the Stage 1 reference checkpoint before Stage 2 updates. We run the frozen host with dropout disabled and teacher forcing, and define the bank loss on answer tokens only:

$$\ell_i = -\frac{1}{|\mathcal{A}_i|} \sum_{t \in \mathcal{A}_i} \log p_{\text{ref}}(y_{i,t} \mid X_i, q_i, y_{i,<t}). \tag{18}$$

Here $\mathcal{A}_i$ is the set of answer-token positions, $y_{i,t}$ denotes the target answer token at position t, and $y_{i,<t}$ denotes the preceding target tokens. Prompt tokens are excluded from this loss. $\Delta_{m,i}$ is the derivative of this loss with respect to the additive site output, and $X_{m,i}$ is the valid input activation at that site. Per-example gradients are detached before forming second moments; no bank parameter is trained through the eigensolver. The primary $G = X^\top \Delta$ is an output-gradient-weighted input statistic. For the local LoRA term, the transposed input-factor gradient contribution is $X^\top \Delta B_0$; we omit B_0 because the extension branch has its own learned output map and the bank is intended to expose useful input directions before choosing that map. The token-count and unit-normalized variants below are alternative constructions, not evaluated accuracy comparisons.

Equation (20) averages one positive-semidefinite matrix per example, so every example receives equal weight, although longer sequences can have larger $\|X^\top \Delta\|_F$. The main protocol intentionally preserves the total gradient signal associated with each example. Sensitivity to sequence length can be assessed by constructing banks from

$$\Sigma_{m,i}^{\text{token}} = \frac{G_{m,i} G_{m,i}^\top}{N_{m,i}}, \qquad \Sigma_{m,i}^{\text{unit}} = \frac{G_{m,i} G_{m,i}^\top}{\|G_{m,i}\|_F^2 + \varepsilon}. \tag{19}$$

The first provides an approximate per-token normalization; the second normalizes each nonzero example up to the stabilizer ε. These variants use the training-derived $\mathcal{C}_1$ for construction and the validation split for model selection. The training-derived $\mathcal{C}_2$ supports the diagnostic analysis described in Appendix A.17.

A.2 BANK CONSTRUCTION AND EIGENVECTOR CHARACTERIZATION

Gradient statistic computation. On a training-only bank set $\mathcal{C}_1$, we run the frozen reference host with teacher forcing. For each example $i \in \mathcal{C}_1$ and modality m, we collect the valid modality activations $X_{m,i} \in \mathbb{R}^{N_{m,i} \times k}$ and the gradients $\Delta_{m,i} \in \mathbb{R}^{N_{m,i} \times d}$ with respect to the additive site output, and compute

$$G_{m,i} = X_{m,i}^\top \Delta_{m,i}, \qquad \Sigma_{m,i} = G_{m,i} G_{m,i}^\top, \qquad \overline{\Sigma}_m = \frac{1}{|\mathcal{C}_1|} \sum_i \Sigma_{m,i}. \tag{20}$$

Let $\mathcal{S}_m = \text{rowspan}(A_m)$, $P_\perp = I_k - \Pi_m$, and $K_m = P_\perp \overline{\Sigma}_m P_\perp$. For any matrix D satisfying $DD^\top = I_r$ and $DP_m^\top = 0$, its rows lie in $\mathcal{S}_m^\perp$, and hence

$$\text{tr}(D \overline{\Sigma}_m D^\top) = \text{tr}(D K_m D^\top). \tag{21}$$

By the Ky Fan maximum principle, this trace is maximized by the top-r eigenvectors of K_m restricted to $\mathcal{S}_m^\perp$, proving Eq. (4). If fewer than r eigenvalues are numerically positive, we complete the bank with zero-eigenvalue vectors chosen inside $\mathcal{S}_m^\perp$ and record their zero-energy status. For repeated eigenvalues, we project the standard basis vectors onto each degenerate eigenspace and orthonormalize the linearly independent projections in coordinate order before applying the sign rule. This fixes the basis within each degenerate eigenspace.

The parent–residual pairing uses a deterministic greedy rule. Starting with all rows unmatched, we choose the remaining pair with the largest $|p_{m,j} \overline{\Sigma}_m d_{m,\ell}^\top|$, record it, and remove its parent and residual rows. We orient each eigenvector so that its largest-magnitude coordinate is positive (breaking coordinate ties by index), and break pairing ties lexicographically by row index. This procedure does not claim to optimize a joint assignment objective; it places locally coupled rotation directions in the same plane.

Numerical conventions. Let $\mathcal{S}_m = \mathrm{rowspan}(A_m)$, $P_\perp = I_k - \Pi_m$, and $K_m = P_\perp \overline{\Sigma}_m P_\perp$. For any matrix D satisfying $DD^\top = I_r$ and $DP_m^\top = 0$, its rows lie in $\mathcal{S}_m^\perp$, and hence

$$\mathrm{tr}(D\overline{\Sigma}_m D^\top) = \mathrm{tr}(DK_m D^\top). \tag{22}$$

By the Ky Fan maximum principle, this trace is maximized by the top-r eigenvectors of K_m restricted to $\mathcal{S}_m^\perp$, proving Eq. (4). If fewer than r eigenvalues are numerically positive, we complete the bank with zero-eigenvalue vectors chosen inside $\mathcal{S}_m^\perp$ and record their zero-energy status. For repeated eigenvalues, we project the standard basis vectors onto each degenerate eigenspace and orthonormalize the linearly independent projections in coordinate order before applying the sign rule. This fixes the basis within each degenerate eigenspace.

The parent–residual pairing uses a deterministic greedy rule. Starting with all rows unmatched, we choose the remaining pair with the largest $|p_{m,j}\overline{\Sigma}_m d_{m,\ell}^\top|$, record it, and remove its parent and residual rows. We orient each eigenvector so that its largest-magnitude coordinate is positive (breaking coordinate ties by index), and break pairing ties lexicographically by row index. This procedure does not claim to optimize a joint assignment objective; it places locally coupled rotation directions in the same plane.

A.3 HARD-CONCRETE GATES

For gate log-location η, temperature $\beta > 0$, and stretch interval (γ_0, ζ) with $\gamma_0 < 0 < 1 < \zeta$, we sample

$$\begin{aligned} u &\sim \mathrm{Uniform}(0,1), \\ s &= \sigma\left(\frac{\log u - \log(1-u) + \eta}{\beta}\right), \\ \gamma &= \min\{1, \max\{0, s(\zeta - \gamma_0) + \gamma_0\}\}. \end{aligned} \tag{23}$$

The analytic probability of an open gate is

$$\Pr(\gamma > 0) = \sigma\left(\eta - \beta \log \frac{-\gamma_0}{\zeta}\right). \tag{24}$$

At inference, we use the deterministic gate

$$\widehat{\gamma} = \min\{1, \max\{0, \sigma(\eta)(\zeta - \gamma_0) + \gamma_0\}\}. \tag{25}$$

Following the standard parameterization (Louizos et al., 2018), we use $(\gamma_0, \zeta) = (-0.1, 1.1)$ and $\beta = 2/3$, as listed in Table 4. We initialize angle heads and gate-head weights to zero and set parent-gate biases to 3.0, so their deterministic gates clamp to one. For extension gates, the initialization condition $\eta_{\mathrm{ext}} < \log(-\gamma_0/\zeta)$ makes the deterministic value in Eq. (25) exactly zero while retaining a nonzero analytic open probability in Eq. (24) for finite biases. Because $B_{\mathrm{ext}} = 0$ makes the initial task gradient with respect to extension gates zero, we linearly warm up the budget coefficient.

Bank-bounded parameterization: construction details. This paragraph gives the explicit construction of the matrices in the factorization $\widetilde{A}_m(q) = J\Gamma_m^{\mathrm{pair}}(q) L_m R_m(q) H_m^{\mathrm{pair}}$ from Eq. (13).

Paired bank ordering. Let $H_m = [a_{m,1}; \ldots; a_{m,r}]$ denote the original bank rows (parent measurements). We form the paired bank by interleaving each parent with its residual candidate:

$$H_m^{\mathrm{pair}} = \begin{bmatrix} p_{m,1} \\ d_{m,1} \\ \vdots \\ p_{m,r} \\ d_{m,r} \end{bmatrix} \in \mathbb{R}^{2r \times d}, \tag{26}$$

where $p_{m,j} = a_{m,j}/\|a_{m,j}\|_2$ and $d_{m,j}$ is the residual candidate for the j-th measurement pair.

Block-diagonal rotation $R_m(q)$. For each pair j, define the 2×2 rotation block

$$R_{m,j}(q) = \begin{bmatrix} \cos\alpha_{m,j}(q) & \sin\alpha_{m,j}(q) \\ 0 & 1 \end{bmatrix}. \tag{27}$$

Then $R_m(q) = \mathrm{diag}(R_{m,1}(q), \ldots, R_{m,r}(q)) \in \mathbb{R}^{2r\times 2r}$. Note that $R_m(q)$ is not an orthogonal rotation matrix: the second row $[0, 1]$ retains the residual row unchanged while the first row mixes the parent into the residual direction.

Repeated row-scale L_m. The matrix $L_m = \mathrm{diag}(\Lambda_{m,1}, \ldots, \Lambda_{m,r}, \Lambda_{m,1}, \ldots, \Lambda_{m,r}) \in \mathbb{R}^{2r\times 2r}$ applies the same scale $\Lambda_{m,j}$ to both the parent and residual rows of pair j.

Diagonal gate matrix $\Gamma_m^{\mathrm{pair}}(q)$. The gate matrix $\Gamma_m^{\mathrm{pair}}(q) = \mathrm{diag}(\gamma_{m,1}(q), \ldots, \gamma_{m,2r}(q)) \in \mathbb{R}^{2r\times 2r}$ places the gate value for each candidate on the diagonal. Parent and residual gates are independent.

Permutation J. The fixed permutation matrix $J \in \{0, 1\}^{2r\times 2r}$ restores parent-first ordering after the block operations. Specifically, J maps the interleaved output of $\Gamma_m^{\mathrm{pair}} L_m R_m H_m^{\mathrm{pair}}$ back to the canonical order where the first r rows correspond to parent measurements and the last r rows to residual measurements. Since J is a permutation, $J^\top J = I$.

A.4 Bank-bounded parameterization: construction details

This subsection gives the explicit construction of the matrices in the factorization $\widetilde{A}_m(q) = J\Gamma_m^{\mathrm{pair}}(q) L_m R_m(q) H_m^{\mathrm{pair}}$ from Eq. (13).

Paired bank ordering. Let $H_m = [a_{m,1}; \ldots; a_{m,r}]$ denote the original bank rows (parent measurements). We form the paired bank by interleaving each parent with its residual candidate:

$$H_m^{\mathrm{pair}} = \begin{bmatrix} p_{m,1} \\ d_{m,1} \\ \vdots \\ p_{m,r} \\ d_{m,r} \end{bmatrix} \in \mathbb{R}^{2r\times d}, \tag{28}$$

where $p_{m,j} = a_{m,j}$ and $d_{m,j}$ is the residual candidate for the j-th measurement pair.

Block-diagonal rotation $R_m(q)$. For each pair j, define the 2×2 rotation block

$$R_{m,j}(q) = \begin{bmatrix} \cos\alpha_{m,j}(q) & \sin\alpha_{m,j}(q) \\ 0 & 1 \end{bmatrix}. \tag{29}$$

Then $R_m(q) = \mathrm{diag}(R_{m,1}(q), \ldots, R_{m,r}(q)) \in \mathbb{R}^{2r\times 2r}$. Note that $R_m(q)$ is not an orthogonal rotation matrix: the second row $[0, 1]$ retains the residual row unchanged while the first row mixes the parent into the residual direction.

Repeated row-scale L_m. The matrix $L_m = \mathrm{diag}(\Lambda_{m,1}, \ldots, \Lambda_{m,r}, \Lambda_{m,1}, \ldots, \Lambda_{m,r}) \in \mathbb{R}^{2r\times 2r}$ applies the same scale $\Lambda_{m,j}$ to both the parent and residual rows of pair j.

Diagonal gate matrix $\Gamma_m^{\mathrm{pair}}(q)$. The gate matrix $\Gamma_m^{\mathrm{pair}}(q) = \mathrm{diag}(\gamma_{m,1}(q), \ldots, \gamma_{m,2r}(q)) \in \mathbb{R}^{2r\times 2r}$ places the gate value for each candidate on the diagonal. Parent and residual gates are independent.

Permutation J. The fixed permutation matrix $J \in \{0, 1\}^{2r\times 2r}$ restores parent-first ordering after the block operations. Specifically, J maps the interleaved output of $\Gamma_m^{\mathrm{pair}} L_m R_m H_m^{\mathrm{pair}}$ back to the canonical order where the first r rows correspond to parent measurements and the last r rows to residual measurements. Since J is a permutation, $J^\top J = I$.

A.5 Budget estimator and controls

For two modalities, write $x_m = \bar{w}_m - r$. The shared-total objective in Eq. (11) becomes

$$\mathcal{L}_b = \frac{\lambda_b}{2}(x_a + x_v)^2. \tag{30}$$

This objective constrains the sum of modality means while allowing their allocation to vary. It does not impose a per-example total-support cap. The random-bank and gradient-bank gate controls in

Table 10 instead retain their modality-wise target r, using the penalty $\lambda_b \sum_m (\bar{w}_m - r)^2$ ("per-mod." in the table). Comparisons with APERTURE therefore account for this difference in budget control.

Let $S(q) = \sum_{m,c} \pi_{m,c}(q)$ and $\mu = \mathbb{E}[S(q)]$. For an i.i.d. minibatch of size n and $b = b_{\text{tot}}$, the minibatch estimator $\overline{S}_{\mathcal{B}} = n^{-1} \sum_{q \in \mathcal{B}} S(q)$ satisfies

$$\mathbb{E}[(\overline{S}_{\mathcal{B}} - b)^2] = (\mu - b)^2 + \frac{\mathrm{Var}(S)}{n}. \tag{31}$$

Thus the training penalty controls the population mean together with a variance-dependent minibatch term. Expected support during training and deterministic support at evaluation are distinct quantities. The result tables report mean deterministic support, which is the quantity used for budget matching. We calibrate the gate logits with a scalar on the validation split and freeze that scalar before test evaluation, using the same rule for every matched arm.

A.6 EQUIVALENT MATRIX FACTORIZATION

For each pair, we define

$$R_j(\alpha_{m,j}) = \begin{bmatrix} \cos \alpha_{m,j} & \sin \alpha_{m,j} \\ 0 & 1 \end{bmatrix}. \tag{32}$$

We permute the bank into paired order and define

$$\begin{aligned} H_m^{\text{pair}} &= [p_{m,1}; d_{m,1}; \ldots; p_{m,r}; d_{m,r}], \\ R_m(q) &= \mathrm{blockdiag}_{j=1}^{r} R_j(\alpha_{m,j}(q)), \\ L_m &= \mathrm{diag}((\Lambda_{m,j}, \Lambda_{m,j})_{j=1}^{r}), \\ \Gamma_m^{\text{pair}}(q) &= \mathrm{diag}((\gamma_{m,j}(q), \gamma_{m,r+j}(q))_{j=1}^{r}). \end{aligned}$$

With the fixed permutation matrix J that restores parent-first ordering,

$$\widetilde{A}_m(q) = J \Gamma_m^{\text{pair}}(q) L_m R_m(q) H_m^{\text{pair}}. \tag{33}$$

This form separates gating, scale, rotation, and the bank. It also shows that a conditioner supplied with the complete $2r$ bank code can reproduce the same linear projection by mixing coordinates after the bank projection. In this paper, *before the bottleneck* therefore means before candidates outside the original r-row host slice become inaccessible, not necessarily before every linear operation.

A.7 REFERENCE CROSS-MODAL INTERACTION

In the complete Reference host, the non-text code Z_m from the local branch in Eq. (1) queries only the prompt-text states Z_t^{ctx}:

$$\mathrm{Att}(Z_m, Z_t^{\text{ctx}}, Z_t^{\text{ctx}}) = \mathrm{softmax}\left(\frac{Z_m (Z_t^{\text{ctx}})^\top}{\sqrt{r}} \right) Z_t^{\text{ctx}}. \tag{34}$$

There are no additional query, key, or value projections in this low-rank interaction. Answer positions are excluded from keys and values under teacher forcing and generation, with the same prompt mask for all arms. APERTURE applies the same operation to Z_m^{rot} and retains the frozen host coefficient λ_m. The extension code enters through B_{ext} without feeding extension coordinates into the text branch. An experiment that changes Eq. (34) is an interaction ablation rather than a matched test of direction allocation.

A.8 EXACT RECOVERY OF THE REFERENCE ADAPTER AT INITIALIZATION

Proposition 1 (Deterministic host recovery). *Suppose $\alpha_{m,j} = 0$, deterministic parent gates equal one, $B_{\text{par}} = B_0$, $B_{\text{ext}} = 0$, $\widetilde{A}_t = A_t$, and λ_m equals its reference-host value. Then the deterministic* APERTURE *output equals the reference-host output for every input.*

Proof. At zero angle, Eq. (8) gives $u_{m,j} = p_{m,j}$. Since $p_{m,j} = a_{m,j} / \|a_{m,j}\|_2$ and $\Lambda_{m,j} = \|a_{m,j}\|_2$, multiplying $p_{m,j}$ by $\Lambda_{m,j}$ recovers $a_{m,j}$. Thus the first r rows in Eq. (9) equal A_m. The second r rows may produce a code, but their contribution is annihilated by $B_{\text{ext}} = 0$. The rotated code, text code, attention function, coefficient λ_m, and B_{par} are therefore exactly those of the Reference host. Adding the unchanged frozen-backbone output preserves equality. □

This statement concerns deterministic inference gates at initialization. Stochastic training samples need not reproduce the host exactly. After initialization, $\widetilde{A}_t$ and B_{par} continue training from the reference-host checkpoint, while B_{ext} is trained from zero; the entire $B_R = [B_{\mathrm{par}} \mid B_{\mathrm{ext}}]$ is trainable.

A.9 Parameter and compute costs

Let there be L adapted sites and $M = |\mathcal{M}|$ non-text modalities. Ignoring biases, the shared question projection contributes $d_c d_t$ trainable parameters, and the per-site angle and gate heads contribute $M(rd_c + 2rd_c)$, or $3LMrd_c$ in total. The directly trained text projections contribute Lrk, and the complete widened output matrices contribute $2Ldr$. Thus, the APERTURE allocation phase optimizes

$$P_{\mathrm{train}} = d_c d_t + 3LMrd_c + Lrk + 2Ldr, \tag{35}$$

excluding biases. The question projection and allocator heads are new. Within B_R, only the extension half adds parameters relative to the inherited output matrix, contributing Ldr; $\widetilde{A}_t$ and B_{par} continue from the host but remain trainable. Storing the $2r$ bank rows and the r paired host-row norms requires $LMr(2k+1)$ fixed scalars.

Equation (35) is a bias-free analytical count of the modules specified above. Comparing this count with parameter percentages in Table 12 requires a common accounting scope: the same site multiplicities, tensor-sharing rules, bias convention, stage-specific trainability, and percentage denominator. Static-Alloc$^+$ and APERTURE use the same tensor shapes and trainable modules by design, so their parameter counts match under the same accounting convention.

For N_m non-text tokens at one site, the dense candidate projection costs $O(N_m k\, 2r)$ and the output multiplication costs $O(N_m d\, 2r)$, in addition to the host attention and the conditioning heads. The dominant projection width therefore matches a static full-bank control, but the allocator has nonzero overhead. Compute and latency comparisons therefore depend on precision, batch size, and sequence lengths as well as parameter count; gate sparsity alone does not reduce these dense projection costs.

A.10 Theoretical characterization

This appendix collects the theoretical foundations underlying the question-conditioned measurement allocation framework introduced in Section 3. We proceed in three parts: (i) the measure-theoretic setup that ensures the optimization problem is well-posed, (ii) the host projection bottleneck that justifies the fixed-bank design, and (iii) the policy-class inclusion result that establishes the expressive advantage of question-conditioned allocation. None of these results guarantees that training will realize an accuracy gain; they characterize the structure of the feasible policy space.

A.11 Measure-theoretic setup

To ensure that the optimization problem in Section 3 is well-posed, we state the underlying measurability and integrability assumptions explicitly. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be the probability space supporting the joint data distribution over examples $Z = (x, q, y)$, where x is the multimodal input, q is the question, and y is the target. Denote by Q the marginal distribution of q.

We assume that the task utility function $\mathcal{U}(Z, a)$ is jointly measurable in (Z, a) and integrable with respect to the joint distribution for every feasible action a. This ensures that the expectation $\mathbb{E}[\mathcal{U}(Z, \pi(Q))]$ in the formulation of Section 3 is well-defined for any measurable policy π. We further restrict the policy class to measurable maps $\pi : q \mapsto a(q)$, which is standard in statistical decision theory and excludes pathological non-measurable selectors.

Regarding attainment: defining $V_b^\star$ as a supremum does not require that an optimal policy $\pi_b^\star$ exists. When the supremum is attained, we denote any attaining policy by $\pi_b^\star$; however, no attainment claim is needed for the policy-inclusion result in Proposition 2, which holds regardless of whether the optimum is achieved. This distinction is important because the feasible policy class $\mathcal{P}_b$ is defined by an inequality constraint on expected cost, and the utility functional $\mathcal{U}$ need not be continuous or concave in the allocation action a.

A.12 HOST PROJECTION BOTTLENECK

We now connect measurement access to the adapter's fixed input projection. The key observation is that the question cannot recover an arbitrary measurement already discarded by the adapter's input projection.

Lemma 1 (Host projection bottleneck)**.** *For fixed side information q, any conditioner that observes only $(A_m x, q)$ satisfies*

$$f(A_m(x+v), q) = f(A_m x, q), \qquad v \in \ker(A_m). \tag{36}$$

Moreover, a linear code $\widetilde{A}(q)x$ is recoverable from $(A_m x, q)$ for every $x \in \mathbb{R}^k$ if and only if

$$\mathrm{rowspan}(\widetilde{A}(q)) \subseteq \mathrm{rowspan}(A_m). \tag{37}$$

Lemma 1 explains why the candidate measurements are exposed before the original host bottleneck: reweighting $A_m x$ cannot recover a linear measurement that the host projection has already discarded. The statement is local to this adapter branch. It does not bound information in the frozen backbone, on a restricted data manifold, or in architectures that provide the conditioner with additional projections. The proof is in Appendix A.14. If questions are statistically correlated with the modality input, they may still provide a predictor for a discarded measurement; the lemma only rules out exact recovery for every x with q held fixed.

A.13 CONDITIONAL POLICY INCLUSION

Allowing the question to choose measurements expands the policy class under the same average budget. This is formalized as follows.

Proposition 2 (Conditional policy inclusion)**.** *Let $\mathcal{P}_b$ and $\mathcal{S}_b$ denote the question-conditioned and static policy classes defined in Section 3 under the same average-support budget. Then $\mathcal{S}_b \subseteq \mathcal{P}_b$, and therefore*

$$\sup_{\pi \in \mathcal{P}_b} \mathbb{E}[\mathcal{U}(Z, \pi(Q))] \geq \sup_{\pi \in \mathcal{S}_b} \mathbb{E}[\mathcal{U}(Z, \pi(Q))]. \tag{38}$$

The inclusion alone does not establish a strict inequality. A strict gain requires that no feasible static action attain the conditional optimum. Under an average-cost constraint, a conditional policy can also distribute costly actions across only a subset of questions, even when the expected utility of each action is independent of the question. Thus, question-dependent action utility is not a necessary condition for a strict gain. The result is a policy-class statement rather than a performance guarantee: it does not imply that the learned allocator reaches the conditional optimum, that the question reveals instance-level modality validity, or that an average-support constraint is equivalent to a per-example cap. The budget-feasible policies represented by the finite neural allocator used by APERTURE form a subset of $\mathcal{P}_b$; the proposition is therefore an expressivity comparison and does not assert that the parameterized training procedure realizes every measurable policy in the abstract class.

Finally, the fixed-bank restriction should not be interpreted as claiming a strictly larger linear function class than all post-projection alternatives. Equation (13) shows that a conditioner given the complete bank code can reproduce the same projection by question-dependent coordinate mixing. The contribution is therefore a bank-bounded, host-compatible, and explicitly auditable parameterization of conditional measurement allocation, rather than a uniqueness claim about dynamic adapters.

A.14 PROOF OF THE ACCESS LEMMA

Proof of Lemma 1. If $v \in \ker(A_m)$, then $A_m(x+v) = A_m x$, so applying the same conditioner with all side inputs and any stochastic seed fixed gives Eq. (36).

If $\widetilde{A}(q) = M(q)A_m$, then $\widetilde{A}(q)x = M(q)(A_m x)$, so the requested code is recoverable from $A_m x$. Conversely, suppose a linear code $\widetilde{A}(q)x$ is recoverable from $A_m x$ for every x. Whenever $A_m x = A_m x'$, recoverability requires $\widetilde{A}(q)x = \widetilde{A}(q)x'$. Hence $\ker(A_m) \subseteq \ker(\widetilde{A}(q))$. By the fundamental theorem of linear algebra, this kernel inclusion is equivalent to $\mathrm{rowspan}(\widetilde{A}(q)) \subseteq \mathrm{rowspan}(A_m)$, which in turn is equivalent to the existence of $M(q)$ with $\widetilde{A}(q) = M(q)A_m$. □

A.15 AVERAGE-BUDGET CONDITIONAL POLICY RESULT

Proof of Proposition 2. For any static policy $\pi_s(q) = a_0$ with $\mathbb{E}_Q[C(a_0)] \le b$, define the conditional policy $\pi_c(q) = \pi_s(q) = a_0$. It is measurable, has the same expected cost, and therefore belongs to $\mathcal{P}_b$. Hence $\mathcal{S}_b \subseteq \mathcal{P}_b$. Taking the supremum of the same utility functional over a superset gives Eq. (38); no assumption about training or attainment of either supremum is required. □

The policy class in Eq. (38) contains every static policy satisfying the same average-support constraint. Therefore its supremum cannot be smaller than the supremum over static policies. Equality can occur when a static policy is conditionally optimal almost surely, or when the restricted bank and budget leave no useful question-dependent choice. The result is a policy-class inclusion statement; it does not certify optimization, semantic demand, or the quality of the local-gradient proxy.

For a learned policy $\widehat{\pi} \in \mathcal{P}_b$, the relevant theoretical quantity is its regret relative to the best policy in the evaluated bank-constrained class, not relative to an oracle that observes the complete example:

$$R_{\mathrm{cond}} = \sup_{\pi \in \mathcal{P}_b} \mathbb{E}[\mathcal{U}(Z, \pi(Q))] - \mathbb{E}[\mathcal{U}(Z, \widehat{\pi}(Q))]. \tag{39}$$

A learned conditional policy can outperform a learned static policy either because the conditional class contains a better policy or because its training procedure finds a higher-utility policy within its class. The matched controls are needed to distinguish these possibilities.

A.16 CONSTRUCTION OF THE MATCHED CONTROLS

This section specifies four controls that assess alternative explanations for the primary comparison and explains how to interpret their reported results.

Static-Alloc$^+$ (parameter-matched, question-blind). Let $\bar{e} = \mathbb{E}_{q \sim \mathrm{train}}[\mathrm{MaskedMean}(E_0(q))] \in \mathbb{R}^{d_t}$, estimated once on the training split and then frozen. Static-Alloc$^+$ replaces Eq. (6) by $c_q^+ = W_\downarrow \bar{e}$ and leaves every other tensor untouched, so $W_\downarrow$, $W_{\alpha,m}$, $b_{\alpha,m}$, $W_{\eta,m}$, $b_{\eta,m}$, $\widetilde{A}_t$, B_{par}, and B_{ext} all retain their shapes and trainability. The trainable parameter count is therefore identical to that of APERTURE, while $I(c_q^+; Q) = 0$ for a fixed trained model. This is a parameter-count-matched, question-blind control, not an exact optimization match: throughout training $\nabla_{W_\downarrow} \mathcal{L}$ is an outer product with the single fixed vector $\bar{e}$, so this gradient has rank at most one. The alternative of holding the allocator *output* at a learned constant removes this gradient path and therefore changes the optimization structure. Static-Alloc (constant c_q, no $W_\downarrow$) is retained as the more standard capacity-controlled variant; the two controls bracket capacity and conditioning effects without claiming identical optimization geometry.

Full-Bank Mixer at matched downstream width. By Eq. (33), a conditioner given the complete $2r$ bank code for each modality can reproduce APERTURE's linear projection. This arm tests how conditional allocation is parameterized within the same bank. Its input support is $2r$ per modality by construction and cannot be matched without changing the control. We instead match the width of the code passed downstream, which bounds its rank: for each modality m, the question-dependent mixing matrix $\Omega_m(q) \in \mathbb{R}^{b_m \times 2r}$ has b_m output rows. We set b_m to the nearest integer in $\{0, \ldots, 2r\}$ to APERTURE's calibrated deterministic mean support in that modality, with ties rounded downward, and report the achieved support difference against the same matching tolerance. The bank, optimizer, schedule, and update budget are held fixed. We report the arm at both $2r$ rows per modality (full-width reference) and b_m rows per modality (the control labeled "rank-matched"). Its total input support is $2Mr$, and its total downstream width is $\sum_m b_m$. The full-bank factorization establishes representability; it does not by itself establish equality of the rank-constrained classes. The accuracy comparison in Table 1 evaluates the full-width and rank-matched controls. Matching downstream width does not reproduce APERTURE's per-example support distribution or imply equal parameter counts or execution costs.

LoRA-gradient bank. Equation (20) uses $G_{m,i} = X_{m,i}^\top \Delta_{m,i}$, whereas the transposed gradient contribution to the reference input factor from the local LoRA term is $X_{m,i}^\top \Delta_{m,i} B_0$. The ablation row builds $\overline{\Sigma}_m^{\mathrm{LoRA}}$ from the latter and solves the same residual eigenproblem (4). Since $\mathrm{rank}(B_0) \le r$,

right-multiplication by B_0 restricts the statistic to an output subspace of dimension at most r. The B_0-free statistic removes this restriction and may retain directions that become useful after B_{ext} is trained. This is a hypothesis about why the two constructions differ, not a proof that the default bank is better. In Table 10, the LoRA-gradient bank scores below the default B_0-free bank on both MUSIC-AVQA (76.20 versus 77.55, a 1.35-point difference) and AVE (74.45 versus 75.58, a 1.13-point difference). This single ablation does not establish that the B_0-free statistic is generally superior.

Frozen output path. The control holds $B_{\text{par}} = B_0$ at its Stage 1 value and trains the allocator and B_{ext}. It measures the effect of freezing the inherited output map while retaining the extension branch. The primary matched comparison keeps the same trainable output path in the question-blind and question-conditioned arms.

A.17 Diagnostic projectors and numerical references

For a row matrix U, we use the Gram-aware projector

$$\Pi_U = U^\top (UU^\top)^\dagger U, \qquad E_{m,i}(U) = \operatorname{tr}(\Pi_U \Sigma_{m,i}), \qquad e_{m,i}(U) = \frac{E_{m,i}(U)}{\operatorname{tr}(\Sigma_{m,i})}. \tag{40}$$

Ratios are defined only for examples with $\operatorname{tr}(\Sigma_{m,i}) > 0$. The bank gap is $g_{\text{bank},m,i} = 1 - e_{m,i}(H_m)$; it is the fraction of total local gradient energy outside the bank, not an accuracy guarantee. Let $F_{m,i}(b)$ be the supremum of the unnormalized energy over at most b rows of the bank, with allowed paired rotations. Let $E^{\text{orient}}_{m,i}$ be the supremum of the energy over row matrices containing one allowed rotated parent measurement per pair. Then

$$E_{m,i}(P_m) \le E^{\text{orient}}_{m,i} \le F_{m,i}(r) \le E_{m,i}(H_m). \tag{41}$$

The orientation and allocation headrooms are the first two differences in (41), divided by $\operatorname{tr}(\Sigma_{m,i})$. They measure accessible local gradient energy, not model accuracy.

For $\epsilon = 0.1$, we define the candidate-constrained width

$$k_{\epsilon,m,i} = \min\{b \in \{0, \dots, 2r\} : F_{m,i}(b) \ge (1-\epsilon) E_{m,i}(H_m)\}. \tag{42}$$

This is a bank-energy covering width, not spectral effective rank, semantic demand, or the rank of the complete weight update. For S3, we enumerate all $k_a, k_v \in \{0, \dots, 2r\}$ with $k_a + k_v \le b_{\text{tot}}$, evaluate $F_{a,i}(k_a) + F_{v,i}(k_v)$, and identify the near-optimal splits. We normalize displayed energies by $\sum_m \operatorname{tr}(\Sigma_{m,i})$ so audio and visual shares have a common denominator; this does not make them a scale-invariant semantic ground truth. We retain tied splits and exclude zero-signal or missing-modality examples.

At $r = 4$, we enumerate the four pair states (neither, parent, residual, both) subject to the budget. We optimize bounded angles from multiple deterministic starts and include the zero-angle host state. The resulting $\widehat{F}$ curves are numerical feasible references, not certified global optima. We define near-optimality as a relative energy difference of 10^{-3} from the best feasible value. Let $\mathcal{S}_{\text{near}}$ denote the support masks within this tolerance for a fixed budget and example; this tolerance is distinct from the covering-width tolerance $\epsilon = 0.1$. The reported bank coverage, headrooms, covering widths, demand shares, and near-optimal-support counts are in Table 15. Because support counts measurements rather than independent directions, k_ϵ must not be interpreted as spectral rank or as evidence of question-dependent mask diversity. We compute these diagnostics on the training-derived $\mathcal{C}_2$ after freezing policies and checkpoints. They neither use test labels nor train the allocator.

B Experimental details

Experimental results and sources. The tables below report point estimates under the stated evaluation protocol. Values marked † are quoted from the cited publications and retain their source scope; parameter shares marked ‡ remain unaudited draft values. Other entries report this study's results. Point estimates support descriptive comparisons but do not establish statistical significance. Appendix B.10 defines the evaluation scope, aggregation rules, and interpretation.

Table 3: Evaluation suite and inclusion rationale. ✓ marks benchmarks outside the published comparison suite (Wei et al., 2025). Each addition is included because it discriminates between allocation policies in a way the existing suite cannot; benchmark count alone does not strengthen an allocation claim.

Scenario	Datasets (✓ = outside the published suite)	What the additions contribute
audio–vis.–text	MUSIC-AVQA, AVE	MUSIC-AVQA provides interpretable question groups and carries the primary S3 analysis; AVE evaluates event localization. The local-gradient proxy is defined in Section 5.4
visual–text	$\text{MME}_{\text{percep}}$, MMBench, POPE, SEED-Bench, GQA ✓, MMStar ✓	GQA supplies question-type strata for **S1**; MMStar is image-dependent by construction and tests performance when visual evidence is required
speech–text	$\text{MMAU}_{\text{mini-speech}}$, $\text{AIR-Bench}_{\text{speech}}$	With one non-text stream, **S3** is vacuous; these benchmarks assess generality only and are never averaged into audio–visual results

Sources and harness. We obtain datasets from ModelScope unless noted otherwise and evaluate through `lmms-eval` where supported. Quoted source values retain their original evaluation protocols. The dataset source identifiers are `lmms-lab/MME`, `lmms-lab/MMBench`, `lmms-lab/POPE`, `lmms-lab/SEED-Bench`, `lmms-lab/GQA`, `evalscope/MMStar`, `AI-ModelScope/LLaVA-Instruct-150K`, `gamma-lab-umd/MMAU-test-mini`, and `evalscope/AIR-Bench`. MUSIC-AVQA and AVE are obtained from their authors' original releases; third-party ModelScope mirrors are used to cross-check file counts. Within each audio–visual evaluation split, all methods use the same examples. The dataset sizes below describe the released benchmarks. Quoted results retain their source evaluation scope, as discussed in Appendix B.10.

B.1 Datasets and split audits

Scenarios and datasets. We follow the three-scenario organization of the published comparison (Wei et al., 2025), retain quoted results as external references, and add benchmarks that distinguish allocation policies.

Audio–visual–text is the primary setting: MUSIC-AVQA (Li et al., 2022), whose 33 templates provide interpretable question groups and therefore support the central claim and the **S3** analysis, with AVE (Tian et al., 2018) for event localization. *Visual–text* fine-tunes on LLaVA-Instruct-150K (Liu et al., 2023) and evaluates on the four benchmarks of the published suite plus GQA (Hudson & Manning, 2019) and MMStar (Chen et al., 2024). *Speech–text* uses LibriSpeech (Panayotov et al., 2015) with $\text{MMAU}_{\text{mini-speech}}$ and $\text{AIR-Bench}_{\text{speech}}$; with one non-text stream **S3** is vacuous, so it probes generality only; speech results are not averaged with audio–visual results. Reproductions use `lmms-eval` where a task exists; quoted results retain their source evaluation protocols.

Stage-1 pretraining. Stage 1 trains the modality projectors and the reference multimodal adapter before any task-specific allocation. The visual projector uses the Video-LLaVA mixture (Lin et al., 2023) (665K image–text and 100K video–text captioning instructions), the audio projector uses AudioCaps (Kim et al., 2019), and the speech projector uses GigaSpeech-M (Chen et al., 2021). Stage 1 also trains the modality-specific down-projections, the shared output projection, and the reference cross-modal interaction. The resulting reference checkpoint is held fixed during bank construction and supplies the common initialization for every arm; this matters for interpreting Lemma 1, which concerns the adapter branch only. Table 4 summarizes the configuration used in the comparison.

Stage-2 fine-tuning. Stage 2 builds the frozen candidate bank from training-only calibration examples and then trains the question projection $W_{\downarrow}$, rotation and gate heads, text input projection, and widened output maps. The question embedding encoder E_0 remains frozen. The reference adapter remains available as the parent branch for identity initialization and for the Reference-cont control, which receives the same Stage-2 update budget without the allocation mechanism. Modality

encoders, the backbone, and the alignment modules remain frozen during Stage 2, as specified in Section 4.4.

MUSIC-AVQA. 45,867 QA pairs over 9,288 videos and 33 templates. Templates are mapped to interpretable audio, visual, or audio–visual groups using the official annotation. These groups are used for stratification; the local-gradient demand share used for allocation analysis is computed separately on $\mathcal{C}_2$. The released question split places questions from the same video in both train and test; we keep this *official* split for comparability with published numbers, and additionally construct a *video-disjoint* split by moving all questions of a randomly chosen subset of test videos, together with all their train questions, into a held-out set. The video-disjoint split seed is 42; the held-out-video count is 1,857, with train/validation/test video counts of 6,000 / 1,431 / 1,857; the resulting QA counts are 29,620 / 7,052 / 9,195.

AVE. 4,143 ten-second videos and 28 event classes. We use the official split and report segment-level accuracy as the primary metric, together with event-level mAP in Table 13.

Visual–text. LLaVA-Instruct-150K for instruction tuning; evaluation on $\text{MME}_{\text{percep}}$ (Fu et al., 2023), MMBench (dev) (Liu et al., 2024b), POPE (random, popular, adversarial; 3,000 questions each) (Li et al., 2023b), SEED-Bench (image subset) (Li et al., 2023a), and GQA testdev-balanced (12,578 questions). GQA structural types (verify, query, choose, logical, compare) define the stratification in Table 14.

Backbone configurations. The host study separates primary mechanism tests from transfer evaluations. *Tier 1* is LLaMA-2-7B-Chat (Touvron et al., 2023) with CLIP ViT-L/14 (Radford et al., 2021), BEATs (Chen et al., 2022), and Whisper (Radford et al., 2023), each behind a 32-query Q-Former projector we pre-train in Stage 1; the control, ablation, and diagnostic results use this host. *Tier 2* is Qwen3-8B (Yang et al., 2025), the primary transfer host, which receives the full APERTURE versus Static-Alloc$^+$ pair so the transfer claim rests on a matched comparison. *Tier 3* comprises LLaMA-3-8B-Instruct (Grattafiori et al., 2024) and Qwen2-7B-Instruct (Yang et al., 2024a), which provide supplementary transfer comparisons. Encoders stay fixed throughout, since changing them would change what the bank spans and confound the matched allocation comparison.

Split-audit protocol. The audit uses example, image/video, question-stem, and split identifiers with SHA-256 hashes to check option-permutation duplicates and overlap between test examples and training, $\mathcal{C}_1$, $\mathcal{C}_2$, validation, or the controlled projector pretraining mixture. The strict video-disjoint evaluation additionally requires held-out video IDs to be absent from these sources. The official MUSIC-AVQA split retains its documented video overlap. These checks address the identified datasets and splits; they do not establish the absence of all overlap with backbone pretraining data. The interpretation of each evaluation split, including the use of $\mathcal{C}_2$, is discussed in Appendix B.10.

B.2 HYPERPARAMETERS

Training and calibration. All arms share the reference-host alignment checkpoint and rank $r = 4$, are warm-started from that checkpoint, and receive the same Stage 2 update budget as *Reference-cont*; optimizer, schedule, and epochs are fixed before Stage 2 and shared. The bank uses $|\mathcal{C}_1| = 1024$ training-only examples; validation selects the checkpoint and calibrates the scalar gate offset. Test evaluation uses the checkpoint and calibration choices fixed on validation.

Uncertainty reporting template. The primary MUSIC-AVQA study declares three independent seeds (42, 43, and 44). The per-seed scores, arithmetic mean, standard deviation, video-clustered 95% confidence interval, and pre-specified paired test are required outputs, but the underlying predictions and run records are not present in the current workspace. Other benchmarks use their official aggregation; the displayed table entries are point estimates until the seed-level records are archived.

Reporting scope. We interpret the point-estimate differences as descriptive comparisons. Comparisons with quoted values do not establish evaluation on identical available examples; the matched internal comparisons use the same subsets.

Table 4: Hyperparameters. Stage-1 pretrained values are fixed before Stage 2; APERTURE-specific values are selected on validation once and frozen. AV and VT denote audio–visual–text and visual–text, respectively.

Group	Name	Value	Note
Host	rank r	4	Reference host; 2 and 8 in the sensitivity run
Host	adapted sites	W_q, W_v of every layer	fixed in Reference host
Host	query tokens per modality	32	Q-Former
Optim.	optimizer / lr / warm-up	AdamW / 10^{-4} / 3%	cosine decay
Optim.	epochs	3 (AV), 1 (VT), 3 (speech)	Stage-2 schedule
Optim.	batch size	8	Stage-2 training
Optim.	precision	bf16	training arithmetic
APERTURE	$\lvert\mathcal{C}_1\rvert$, $\lvert\mathcal{C}_2\rvert$	1024, 1024	training-only, disjoint
APERTURE	d_c (question code)	64	$W_\downarrow$ Xavier, no bias
APERTURE	angle bound	$\pi/4$	(8)
APERTURE	(γ_0, ζ), β	$(-0.10, 1.10)$, 2/3	hard-concrete
APERTURE	parent gate bias	3.00	$\eta_{\text{ext}} <$ $\log(-\gamma_0/\zeta)$ for identity
APERTURE	λ_b, warm-up	0.02, linear warm-up	budget penalty
APERTURE	support tolerance	± 0.10 / modality, ± 0.20 joint	matching rule
Reporting	seeds	42, 43, 44; otherwise 42	primary MUSIC-AVQA; other evaluations

B.3 VISUAL–TEXT RESULTS

With one non-text stream, S3 is vacuous and the comparison isolates orientation (S1) and within-modality width (S2). Table 5 therefore carries the same matched controls as the primary table but no cross-modal split arm, and it is read as supporting evidence for bank-bounded conditional projection rather than as evidence about cross-modal allocation.

Relative to Static-Alloc$^+$, APERTURE improves $\text{MME}_{\text{percep}}$ by 5.04 score units; MMBench changes by -0.37 points, and POPE, SEED-Bench, GQA, and MMStar accuracy change by 2.63, 3.39, 1.54, and 0.70 percentage points, respectively. The POPE difference is computed from the displayed subset accuracies before rounding their means. The reported test supports are 4.75 for Static-Alloc$^+$ and 4.62 for APERTURE; equality of achieved support is not established by these entries. The parameter-matching design is described in Appendix A.16.

B.4 EXTERNAL BASELINE REFERENCES

Table 6 positions **APERTURE** relative to the adaptation families referenced in this work. All rows except the last two correspond to prior methods whose implementations and protocols are taken as given; among them, Conditional MixLoRA and Multimodal PEFT (MokA) are the sources of the published comparison values quoted below.

Published comparison values used for context are quoted from Wei et al. (2025) (arXiv:2506.05191v3) under the protocol described in Section 5.1. The quoted audio–visual LoRA results come from its Table 1, and the visual–text LoRA and MokA results come from its Table 2. They are not used to initialize Stage 1 or Stage 2 and are not treated as measurements of the present model. The source citation is retained wherever a quoted value appears.

Table 8 gives the Stage-1/Stage-2 transfer blocks condensed in Table 1; its POPE column records separate visual–text evaluations. Each host reports results under the same two-stage protocol, and each transfer comparison uses the same Static-Alloc$^+$ control as the primary study.

Relative to each host's Static-Alloc$^+$, the measured MUSIC-AVQA/AVE accuracy differences are $+2.37/+1.19$ percentage points for Qwen3-8B, $+0.78/+1.21$ for LLaMA-3-8B, and $+2.99/+1.87$

Table 5: Visual–text results (accuracy, %; $\text{MME}_{\text{percep}}$ is a 0–2000 score). † quoted from Table 2 of the published comparison (Wei et al., 2025). MokA† is the source model, distinct from our internally evaluated Reference host. Unquoted rows are experimental point estimates. Static-Alloc, Static-Alloc$^+$, and APERTURE share the candidate bank and use the same validation support target; achieved test supports are reported separately and require a matching audit. Static-Alloc$^+$ uses the parameter-matching design (Appendix A.16). With one non-text stream, S3 is vacuous and the budget controls that stream's support. Panel (a) compares benchmarks in the published suite; panel (b) gives our GQA and MMStar evaluations. Bold marks the best score in each column.

(a) Published-suite benchmarks and matched controls

Method	MME_{p}	MMBench	POPE	SEED	$\bar{r}_{\text{test}}$
LoRA†	908.52	50.64	70.28	39.71	4
MokA†	1025.86	52.74	74.23	40.45	4
Reference-cont	1027.22	51.56	77.99	42.32	4
Uniform-wide	1020.47	50.79	74.28	39.58	8
Static-Alloc	1024.82	52.72	74.25	40.42	4.81
Static-Alloc$^+$	1025.18	**54.96**	75.41	41.74	4.75
APERTURE	**1030.22**	54.59	**78.04**	**45.13**	4.62

(b) Additional visual–text benchmarks

Method	GQA	MMStar
Reference-cont	59.32	44.27
Uniform-wide	60.58	45.48
Static-Alloc	61.27	46.46
Static-Alloc$^+$	62.89	46.77
APERTURE	**64.43**	**47.47**

Table 6: Locus of conditionality in representative PEFT methods.

Method family	Conditioning signal	What is conditioned / selected
LoRA / AdaLoRA / DoRA / RandLoRA	task or training-time schedule	low-rank parameterization of weight updates (rank, magnitude, or basis)
Conditional MixLoRA	multimodal input	LoRA parameters or updates conditioned on the input modality
MoSLoRA	task or training	mixture of low-rank subspaces
LD-MoLE / LoRA-Mixer	token or layer	routing or mixing over LoRA experts within projection layers
LoRA-MoE / hypernetwork-based	input or task	selection, mixing, or generation of adapter parameters
Multimodal PEFT (MokA, Phi-4 Multimodal)	modality identity	modality-specific vs. shared adapter modules and pathways
Full-Bank Mixer (**control**)	question	conditional mixing over the *full* candidate bank; no measurement selection
APERTURE	**question**	**selection of measurements exposed to a fixed adapter under a shared average budget**

for Qwen2-7B in this table. Qwen3-8B, Qwen2-7B, and LLaMA-3-8B improve on both metrics. Benefits therefore depend on the host and task; point estimates do not establish statistical significance.

In the separate visual–text evaluations, APERTURE improves POPE accuracy over each host's Static-Alloc$^+$ by 1.12, 1.16, and 1.09 percentage points for Qwen3-8B, LLaMA-3-8B, and Qwen2-7B, respectively.

Table 7: Complete LLaMA-2-7B audio–visual–text baseline block (accuracy, %). The five external rows omitted from the compact main-text Table 1 are retained here with the same values and protocol. $\bar{r}_{\text{test}}$ is the mean deterministic active support per non-text modality at $r = 4$; † marks a value quoted from the cited source.

		Matched			Accuracy (%)		
LLM	**Method**	**B**	**S**	**P**	**MUSIC-AVQA**	**AVE**	$\bar{r}_{\text{test}}$
LLaMA-2-7B	LoRA†				73.41	69.84	4
	MoELoRA				72.41	70.84	4
	MoSLoRA				71.78	69.72	4
	Conditional MixLoRA				73.54	70.43	4
	LD-MoLE				73.73	71.10	4
	LoRA-Mixer				73.80	71.36	4
	Reference host				72.71	71.68	4
	Reference-cont				73.54	71.02	4
	Random bank		✓	✓	72.95	67.45	4
	Uniform-wide	✓			74.35	71.71	8
	Full-Bank Mixer ($2r$)	✓			72.43	70.26	8
	Full-Bank Mixer (rank-matched)	✓	✓		74.82	73.88	8
	Shuffle-train	✓	✓	✓	72.16	70.08	4
	Static-Alloc	✓	✓		73.20	71.55	4
	Static-Alloc$^+$	✓	✓	✓	73.37	73.35	4
	Static-Joint	✓	✓	✓	73.58	72.50	4
	APERTURE	✓	✓	✓	**77.55**	**75.58**	4

Table 8: Transfer-host results (accuracy, %). Each host reports Reference host, Reference-cont, Static-Alloc$^+$, and APERTURE under the same Stage-1/Stage-2 protocol. Static-Alloc$^+$ uses the same support-calibration tolerance and the parameter-matching design of Appendix A.16. Bold marks the best unrounded score within each host and metric; comparisons across host blocks are not used to select bold entries.

Tier	**LLM**	**Method**	**MUSIC-AVQA**	**AVE**	**POPE**	$\bar{r}_{\text{test}}$
3	Qwen2-7B	Reference host	72.89	70.26	70.52	4.00
		Reference-cont	73.14	70.94	70.81	4.00
		Static-Alloc$^+$	73.65	71.75	72.36	4.00
		APERTURE	**76.64**	**73.62**	**73.45**	4.00
3	LLaMA-3-8B	Reference host	78.99	75.49	76.15	4.00
		Reference-cont	79.17	75.92	77.32	4.00
		Static-Alloc$^+$	78.85	76.45	79.56	4.00
		APERTURE	**79.63**	**77.66**	**80.72**	4.00
2	Qwen3-8B	Reference host	80.89	78.44	81.52	4.00
		Reference-cont	82.26	77.61	83.67	4.00
		Static-Alloc$^+$	81.45	78.19	84.22	4.00
		APERTURE	**83.82**	**79.38**	**85.34**	4.00

B.5 SPEECH–TEXT TRANSFER

The speech scenario has one non-text stream, so only direction and width are conditioned. We report it as a transfer check under the frozen default configuration, evaluating on $\text{MMAU}_{\text{mini-speech}}$ (Sakshi et al., 2024) and $\text{AIR-Bench}_{\text{speech}}$ (Yang et al., 2024b). Relative to Static-Alloc$^+$, APERTURE increases the reported MMAU and AIR-Bench scores by 1.85 and 2.55 percentage points, respectively.

Table 9: Speech–text transfer results (LLaMA-2-7B) under the two-stage protocol. Both benchmark scores are expressed on a 0–100 percentage scale. Bold marks the best score in each column.

Method	$\text{MMAU}_{\text{mini-speech}}$ (%)	$\text{AIR-Bench}_{\text{speech}}$ (%)
Reference host	60.24	58.48
Reference-cont	61.58	58.66
Static-Alloc^{+}	61.84	58.96
APERTURE	**63.69**	**61.51**

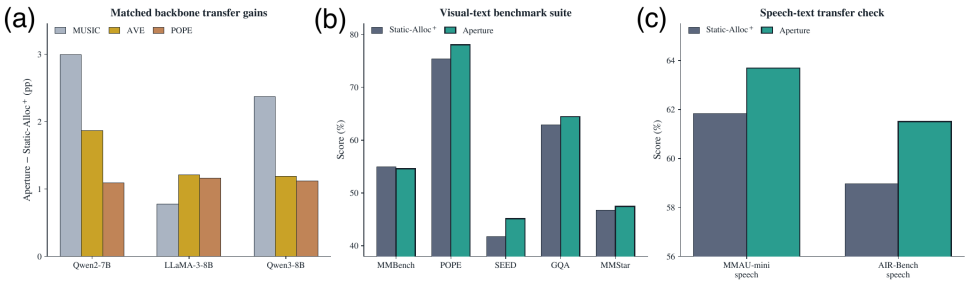

Figure 3: Transfer checks across backbone and scenario hosts. (a) Matched-backbone improvements of APERTURE over Static-Alloc^{+} on MUSIC-AVQA, AVE, and POPE, in percentage points. (b) Visual–text benchmark scores for the same matched comparison. (c) Speech–text transfer scores on MMAU-mini-speech and AIR-Bench-speech. All bars are point estimates under the two-stage protocol.

B.6 MECHANISM ABLATION

The following ablations separate the bank statistic, question conditioner, rotation, gating, and output-path choices. Because these components train under one loss, the rows are configuration-level comparisons rather than additive component attributions.

Table 10: Mechanism ablation on LLaMA-2-7B. Δ_{host} is the percentage-point change in MUSIC-AVQA official accuracy against the Reference host. $\bar{r}_{\text{ach}}$ is achieved audio+visual support.

	Configuration							MUSIC-AVQA		AVE
#	Variant	Eq.	Bank	Rot.	Gate	Bud.	Out	$\bar{r}_{\text{ach}}$	Acc.	Δ_{host}	Acc.
1	host, parent only	(2)	parent	–	–	fixed	–	4+4	72.71	–	71.68
2	random bank + gates	(24)	random	–	yes	per-mod.	full	4+4	73.05	+0.34	70.96
3	gradient bank + gates	(4)	gradient	–	yes	per-mod.	full	4+4	73.80	+1.09	71.20
4	gradient bank + rotation	(8)	gradient	yes	–	fixed	full	4+4	75.10	+2.39	74.60
5	**APERTURE**	(11)	gradient	yes	yes	joint	full	4.10+3.91	**77.55**	**+4.84**	**75.58**
6	pooled-state question code	(6)	gradient	yes	yes	joint	full	4+4	74.15	+1.44	74.40
7	activation-PCA bank	(20)	activation	yes	yes	joint	full	4+4	74.90	+2.19	74.25
8	LoRA-gradient bank	(20)	LoRA-grad	yes	yes	joint	full	4+4	76.20	+3.49	74.45
9	frozen output path	(15)	gradient	yes	yes	joint	frozen	4+4	74.90	+2.19	74.30

The random-to-gradient comparison keeps candidate width and per-modality targets fixed. The pooled-state, activation-PCA, and LoRA-gradient variants underperform APERTURE; row 9 also shows that freezing B_{par} while retaining B_{ext} is weaker. These results support the selected configuration without establishing a unique causal decomposition.

B.7 SENSITIVITY, ROBUSTNESS, AND COST

Table 11 compares rank, support budget, and modality masking under the same evaluation protocol. Table 12 reports parameter, latency, and memory costs; these are distinct from the active support budget.

Table 11: Sensitivity on MUSIC-AVQA (LLaMA-2-7B; accuracy, %). One factor varies at a time from the validation-selected default ($r = 4$, $b_{\text{tot}} = 8$, $|\mathcal{C}_1| = 1024$). Masking removes one modality at test time. The calibration size is fixed. The budget sweep specifies the stated target budgets.

Factor	Setting	Static-Alloc$^+$	APERTURE
rank r	2	71.44	**74.53**
	4 (default)	73.37	**77.55**
	8	74.39	**78.19**
budget b_{tot}	$\lvert\mathcal{M}\rvert r/2 = 4$	72.63	**74.25**
	$\lvert\mathcal{M}\rvert r = 8$ (default)	73.37	**77.55**
	$3\lvert\mathcal{M}\rvert r/2 = 12$	75.17	**77.74**
robustness	mask audio	54.29	**55.58**
	mask visual	55.11	**56.31**

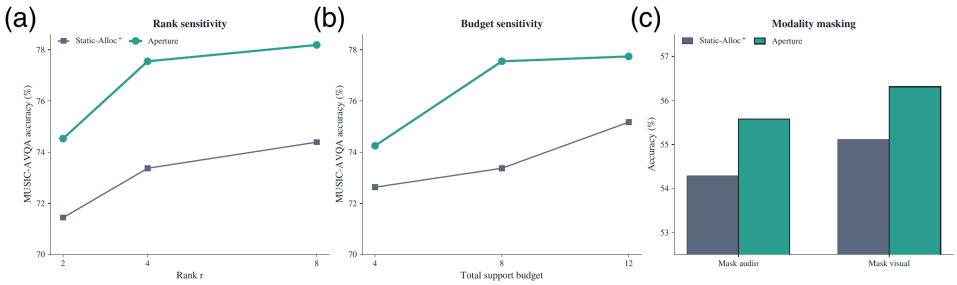

Figure 4: Sensitivity and modality-masking checks on MUSIC-AVQA (LLaMA-2-7B). (a) Rank sensitivity for $r \in \{2, 4, 8\}$. (b) Sensitivity to the total support budget $b_{\text{tot}} \in \{4, 8, 12\}$. (c) Accuracy after masking audio or visual input. Each panel varies one factor at a time from the validation-selected default; results are point estimates.

Table 12: Cost on LLaMA-2-7B. Internally evaluated methods use identical hardware, batch, precision, and sequence length. Latency is forward time per sample relative to LoRA. † values retain the profiling protocol and LoRA normalization of the published comparison (Wei et al., 2025). The dense implementation evaluates all $2r$ candidates, so support sparsity does not reduce FLOPs. Static-Alloc$^+$ and APERTURE have identical trainable tensors by design. ‡ marks a retained draft parameter share that conflicts with this design and requires verification against the trainable-module inventory; both marked internal shares are pending a common counting-scope audit. The cost measures describe the dense implementation.

Method	Train. (%)	Latency (×)	Memory (×)
LoRA†	1.27	1.00	1.00
Reference host	1.05	1.02	1.01
Static-Alloc	1.06	1.03	1.02
Static-Alloc$^+$	1.08‡	1.04	1.03
Uniform-wide	1.15	1.08	1.06
Multiple LoRA	1.40	1.12	1.25
APERTURE	1.12‡	1.07	1.05

Cost accounting and profiling. The parameter percentages in Table 12 are reported trainable shares; marked draft values await a counting-scope audit. Equation (35) describes the module inventory under the stated sharing assumptions. The quoted LoRA value retains its source accounting scope, while comparisons among our evaluated methods use the same profiling settings. Active support counts and dense execution costs measure different properties of the implementation. Hardware profiling uses NVIDIA A100-80GB GPUs.

B.8 ADDITIONAL RESULT TABLES

Table 13 separates evaluation splits and metrics, Table 14 reports visual question strata, and Table 15 summarizes the bank and allocation diagnostics.

Table 13: Official accuracy and additional metrics (LLaMA-2-7B). MUSIC-AVQA official accuracy and AVE segment accuracy repeat Table 1. Video-disjoint accuracy uses a separate MUSIC-AVQA split; event-level mAP is a separate AVE metric. Scores are percentages; Δ is APERTURE minus Static-Alloc$^+$ in percentage points within each column. Bold marks the best score in each column and the explicitly labeled matched difference row. The video-disjoint and event-mAP columns are appendix-only.

	MUSIC-AVQA		AVE (official)	
Method	**Official Acc.**	**Video-disjoint Acc.**	**Segment Acc.**	**Event mAP**
Reference host	72.71	73.90	71.68	70.10
Static-Alloc$^+$	73.37	74.88	73.35	70.47
APERTURE	**77.55**	**76.97**	**75.58**	**71.15**
Δ (APERTURE $-$ Static-Alloc$^+$)	**4.18**	**2.09**	**2.23**	**0.68**

The accuracy difference between APERTURE and Static-Alloc$^+$ is $+4.18$ points on the official MUSIC-AVQA split and $+2.09$ on the video-disjoint split. On AVE, it is $+2.23$ points in segment accuracy and $+0.68$ points in event-level mAP.

Table 14: POPE by subset and GQA by question type (LLaMA-2-7B; scores in %). POPE random, popular, and adversarial are released subsets of 3,000 questions each. Overall POPE accuracy is the arithmetic mean of the accuracies on these three equally sized subsets. GQA scores are conditional on structural type; their unweighted mean is not the overall GQA accuracy. The Reference host row is an internal entry, distinct from the quoted MokA row in Table 5.

(a) POPE subset accuracy and F1

	Random		Popular		Adversarial	
Method	**Acc.**	**F1**	**Acc.**	**F1**	**Acc.**	**F1**
Reference host	75.42	74.82	74.12	72.81	74.62	70.82
Static-Alloc$^+$	76.26	74.74	74.81	73.61	75.16	71.88
APERTURE	**78.78**	**75.43**	**77.22**	**74.32**	**78.11**	**73.98**

(b) GQA accuracy within each structural type

Method	**Verify**	**Query**	**Choose**	**Logical**	**Compare**
Reference host	60.27	58.88	62.76	56.97	61.53
Static-Alloc$^+$	61.57	59.57	63.87	57.57	62.29
APERTURE	**63.11**	**62.69**	**65.66**	**60.77**	**65.64**

Table 15: Bank and allocation diagnostics on $\mathcal{C}_2$ (MUSIC-AVQA, LLaMA-2-7B, $r = 4$). Bank coverage and headrooms are normalized by total local gradient energy; k_ϵ covers a fraction of bank-captured energy. The numerical reference is the bank-constrained enumeration of Appendix A.17 (four support states per plane, hence $4^r = 256$ support masks per modality at $r = 4$, with angle refinement; feasible audio–visual combinations are enumerated subject to the joint budget) and is a diagnostic, not an accuracy bound. The quantities are defined in Appendix A.17.

Quantity	**Audio**	**Visual**
bank-captured energy $1 - g_{\mathrm{bank}}$	0.83	0.80
orientation headroom h^{orient} (mean)	0.17	0.20
allocation headroom h^{alloc} (mean)	0.13	0.16
bank-constrained covering width k_ϵ, $\epsilon = 0.10$ (median)	3	4
demand share s_m (mean over audio / visual / AV questions)	0.48	0.52
$\lvert\mathcal{S}_{\mathrm{near}}\rvert$ of near-optimal supports (median)	6	7

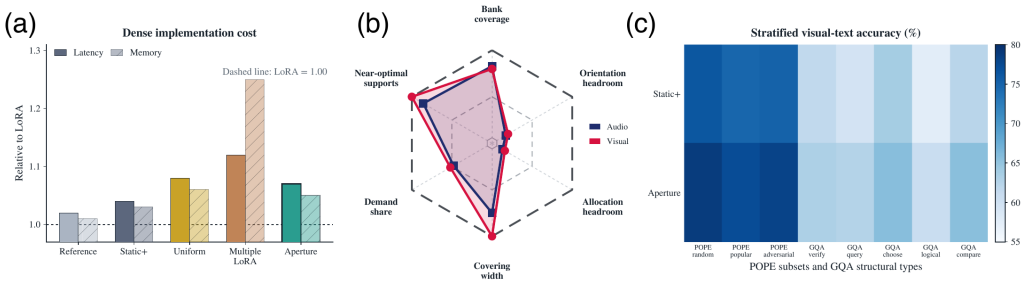

Figure 5: Supplementary cost and audit summaries. (a) Dense implementation latency and memory relative to LoRA. (b) Audio and visual bank coverage, headroom, demand share, covering width, and near-optimal support counts. (c) Stratified visual–text accuracy across POPE subsets and GQA structural types for Static-Alloc$^+$ and APERTURE. These panels summarize diagnostics and do not replace held-out test estimates.

Interpretation of diagnostics. The reported diagnostics do not provide a unique causal explanation for the accuracy differences. Low bank coverage limits what this bank can expose; a weak question-only probe limits the information available to the allocator; and a weak end-to-end result can additionally reflect optimization, calibration, or output-path effects. These observations are hypotheses for follow-up ablations, not a decision rule that identifies one cause from one scalar.

B.9 ALLOCATION AND SUPPORT ANALYSIS

The budget sweep in Table 11 reports total support targets $b_{\text{tot}} \in \{4, 8, 12\}$. Relative to Static-Alloc$^+$ at each target, APERTURE improves MUSIC-AVQA accuracy by 1.62, 4.18, and 2.57 percentage points, respectively. These operating points support comparisons at the reported budgets; they do not determine a continuous support frontier.

Table 2 reports question-group accuracy and aggregate allocation statistics, while Table 15 characterizes the candidate bank on $\mathcal{C}_2$. The audio and visual banks capture 83.00% and 80.00% of local gradient energy, respectively; their median covering widths are three and four at $\epsilon = 0.10$. These diagnostics quantify accessible local gradient energy; they do not by themselves identify the cause of an accuracy gain or establish a per-example semantic explanation.

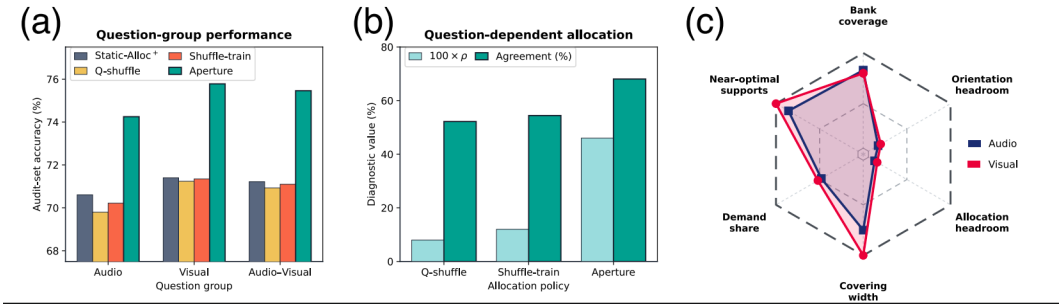

Figure 6: Allocation diagnostics on the training-derived audit set $\mathcal{C}_2$ (MUSIC-AVQA, LLaMA-2-7B). (a) Question-group accuracy for Static-Alloc$^+$, Q-shuffle, Shuffle-train, and APERTURE. (b) Learned allocation versus the independent local-demand proxy, showing $100 \times \rho$ and argmax agreement. (c) Audio and visual bank diagnostics shown as a radar summary. These audit results are diagnostic and are not additional held-out test estimates.

The cost measurements in Table 12 show that APERTURE has reported latency of $1.07\times$ and memory of $1.05\times$, compared with $1.04\times$ and $1.03\times$ for Static-Alloc$^+$. The retained draft trainable shares are 1.12% for APERTURE and 1.08% for Static-Alloc$^+$. These values conflict with the identical trainable-tensor design and require a common counting-scope audit; they are not verified measurements. Different latency and memory do not require different trainable parameter counts. The active support is an allocation statistic; the dense evaluation of all candidates does not provide a corresponding FLOP reduction.

B.10 Evaluation and interpretation

Evidence status. The tables report point estimates under the protocols specified in this paper. This checklist separates values reported in the manuscript from the supporting artifacts that must be archived for reproducibility. Values marked † elsewhere in the paper are quotations from the cited source and retain that source's protocol; they are not measurements of the present implementation.

Current numerical snapshot. The current manuscript records the following point estimates for the primary LLaMA-2-7B comparison: MUSIC-AVQA official accuracy is 73.37% for Static-Alloc$^+$ and 77.55% for Aperture ($+4.18$ points); the video-disjoint accuracy is 74.88% and 76.97% ($+2.09$ points); AVE segment accuracy is 73.35% and 75.58% ($+2.23$ points); and AVE event mAP is 70.47% and 71.15% ($+0.68$ points). For the visual–text suite, the corresponding Static-Alloc$^+$ / Aperture entries are MME $1025.18/1030.22$, MMBench $54.96/54.59$, POPE $75.41/78.04\%$, SEED-Bench $41.74/45.13\%$, GQA $62.89/64.43\%$, and MMStar $46.77/47.47\%$; achieved test supports are $4.75/4.62$. These are reported point estimates, not seed-level summaries. The diagnostic snapshot in Table 15 is $0.83/0.80$ bank-captured energy, $0.17/0.20$ orientation headroom, $0.13/0.16$ allocation headroom, $3/4$ covering width, $0.48/0.52$ demand share, and $6/7$ near-optimal support counts (audio/visual).

Scores and comparisons. Tables report experimental point estimates. Methods within an internal comparison must use the same examples, preprocessing, checkpoint selection rule, and evaluation code. The official MUSIC-AVQA evaluation, video-disjoint evaluation, and training-derived audit-set macro evaluation are distinct quantities: the first follows the released question split, the second holds out complete videos, and the third averages audio, visual, and audio–visual question groups with equal weight. Group-wise Macro entries are the arithmetic means of the three displayed group accuracies, rounded to two decimal places only after aggregation.

Aggregation and precision. POPE aggregate accuracy is defined as the arithmetic mean of the random, popular, and adversarial subset accuracies, each evaluated over 3,000 questions. For Static-Alloc$^+$ this gives $(76.26 + 74.81 + 75.16)/3 = 75.41\%$; for Aperture it gives $(78.78 + 77.22 + 78.11)/3 \approx 78.04\%$. The difference computed from the displayed subset accuracies before rounding the means is 2.63 percentage points. GQA structural-type scores describe their respective question groups, while the aggregate is question-weighted. The Reference host subgroup row is an internal measurement, distinct from the quoted MokA row. All decimal results use two decimal places for display; differences and relative changes are computed before rounding. Bold values identify the best unrounded scores within each stated comparison, rather than statistical significance.

Uncertainty and repeated runs. The primary MUSIC-AVQA comparison declares seeds 42, 43, and 44. The bootstrap must resample complete videos rather than individual questions; the split seed, video counts, question counts, and resampling implementation must be archived with the predictions. A single displayed value is a point estimate; it should not be interpreted as a mean or uncertainty estimate until these records are present.

Split and audit scope. The official MUSIC-AVQA split deliberately preserves the released video overlap for comparability. The additional video-disjoint split uses seed 42, 1,857 held-out videos, train/validation/test video counts of 6,000 / 1,431 / 1,857, and corresponding question counts of 29,620 / 7,052 / 9,195. The official split, video-disjoint split, and training-derived $\mathcal{C}_2$ demand analysis must never be pooled into one accuracy number.

Budgets and execution cost. Support counts the active candidate measurements; nominal budget targets in a sensitivity sweep and achieved mean support in a support-calibrated comparison serve different purposes. Calibration matches its deterministic mean on validation, and the joint objective controls the sum across modalities. The shared tensor parameterization defines the parameter-matching control. The bias-free module count in Eq. (35) is an analytical quantity. The current cost table reports trainable shares of 1.27% (LoRA†), 1.05% (Reference host), 1.06% (Static-Alloc), 1.08% (Static-Alloc$^+$), 1.15% (Uniform-wide), 1.40% (Multiple LoRA), and 1.12% (Aperture). The 1.08% and 1.12% internal shares conflict with the identical trainable-tensor design and remain pending a common counting-scope audit; they are not verified measurements. The quoted LoRA

percentage includes the source model's projectors, whereas Stage 2 here freezes alignment modules, so the external row is source context rather than a matched Stage-2 efficiency result. Dense evaluation of all $2r$ candidates does not imply FLOP savings from a sparse selected support.

Profiling requirements. Latency and memory are not inferable from the support budget. For each method, the device and model identifier, CUDA/PyTorch/compiler versions, batch size, sequence length, precision, warm-up steps, timed repetitions, synchronization rules, aggregation statistic, and raw profiler output or hash are required; the current protocol records an NVIDIA A100-80GB GPU, batch size 8, and bf16 arithmetic; the table reports relative latency/memory of $1.07/1.05$ for APERTURE and $1.04/1.03$ for Static-Alloc^{+}. The same protocol must be used for all internally compared methods, while † rows retain the profiling scope reported by their source.

Interpretation. The measured score differences support descriptive comparisons across the stated settings. Bank diagnostics, demand shares, orientation headroom, allocation headroom, and near-optimal-support counts quantify local reference-host behavior; they do not establish a causal explanation or a global optimum. The video-disjoint comparison and transfer checks should be interpreted with their uncertainty intervals and split audits.

Table 16: Evidence checklist for the experimental record. "Reported" means that the point estimate appears in the manuscript; "archive pending" marks the supporting artifact or audit record that still needs to be retained.

Evidence item	Required artifact	Required report	Status
Primary MUSIC-AVQA	Predictions for seeds 42, 43, 44 with checkpoint and code hashes	Per-seed accuracy, mean/std or CI, paired difference, video-clustered 95% interval	Reported; archive pending
AVE	Predictions and official evaluation output for every declared seed	Segment accuracy and event-level mAP with aggregation rule and uncertainty	Reported; archive pending
Video-disjoint split	Split seed, complete video manifest, SHA-256, and train/validation/test counts	Video and question counts plus overlap audit against training and calibration data	Reported; audit pending
$\mathcal{C}_1, \mathcal{C}_2$	Example IDs, hashes, construction code, and frozen calibration checkpoints	Group counts, no-test-overlap check, and demand/diagnostic computation	Reported; archive pending
Baselines and controls	Commit, configuration, checkpoint, parameter inventory, and support-calibration log	Matched support and parameter audit for every internal baseline	Reported; audit pending
Cost measurements	Hardware/software metadata and raw profiler output or hash	Warm-up, repetitions, synchronization, aggregation, latency, memory, and parameter share	Reported; profiler pending
Reproducibility package	Environment lockfile, commands, random seeds, and dataset/checkpoint hashes	End-to-end rerun record and generated metric files	Config reported; rerun pending

AI USE STATEMENT

Language-model tools, including RelatedWord, were used to assist with literature retrieval and background research informing the Related Work section, as well as with copy editing and LaTeX consistency checks. All sources, citations, claims, and AI-assisted outputs were independently reviewed and verified by the authors, who remain fully responsible for the content, experimental results, and interpretations presented in this paper.