LEARNING WITHOUT LISTENING: SILENT TWINS FOR WEAK-EVIDENCE AUDIO-LANGUAGE FINE-TUNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Post-training of audio language models usually uses supervision with definite answers, yet the available sound may not sufficiently constrain these targets. In this setting, models may still obtain substantial in-domain gains from non-acoustic task signals. Using AVQA fine-tuning as an example, we observe that a large in-domain improvement does not consistently translate into cross-task gains; moreover, replacing all training audio with equal-duration silence still yields substantial in-domain gains. The matched silent twin thus provides a measurable reference for these training-induced changes. In our experiments, we find that normal and silent fine-tuning induce highly overlapping changes across multiple audio language models and transfer benchmarks, and both damage questions that the base model originally answered reliably. We propose STS (Silent-Twin Subtraction): it uses the weight update produced by silent fine-tuning as a reference, subtracts a tunable proportion of it from the normal fine-tuning update, and uses a second coefficient to control the overall magnitude of the combined update. Across models and benchmarks, silent twins reveal what fine-tuning learns without listening, and STS consistently outperforms conventional fine-tuning on every transfer benchmark we evaluate.

1 INTRODUCTION

Post-training improves an audio-language model's ability to answer questions, while the supervision fixes the answer rather than the evidence the model must use (Tang et al., 2024; Chu et al., 2024; Xu et al., 2025; Li et al., 2025; Rouditchenko et al., 2025; Wu et al., 2026). A target answer can be supported by the recording, the wording of the question, the options, or regularities shared across examples. We use *weak evidence* to describe supervision whose target is not sufficiently constrained by the available sound. In this setting, understanding a post-training gain requires examining which changes can arise without the original training audio and how those changes affect performance on other tasks.

Answering this question calls for a training-level comparison. Test-time audio ablations measure a trained model's reliance on a clip (He et al., 2026; Foo et al., 2026), while question-only and hypothesis-only baselines expose predictable labels (Goyal et al., 2017; Gururangan et al., 2018). Training without audio directly tests a different training condition (Rouditchenko et al., 2025), although removing the audio tokens also changes the input sequence. A complementary issue concerns the readout: average accuracy can conceal gains on previously difficult questions that offset losses on reliable ones. We therefore pair a format-preserving training control with a question-level comparison of the resulting models.

We introduce the *silent twin*, a matched run in which every training waveform is replaced by equal-duration silence (Figure 2). Targets, training conditions and audio-token positions are preserved, and both models are evaluated on real audio. On AVQA fine-tuning (Yang et al., 2022), the twin obtains 79% of the normal run's in-task gain. The relationship extends to individual questions: on MMAR (Ma et al., 2025), normal fine-tuning produces substantial gains and losses beneath a nearly unchanged mean, and the twin reproduces much of this movement, including losses on Base-stable questions. The noise-corrected overlap is 0.81 on MMAR and ranges from 0.73 to 0.87 across three architectures (Table 3). Normal fine-tuning remains stronger than its twin, establishing an additional benefit from real-audio training alongside these shared changes.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

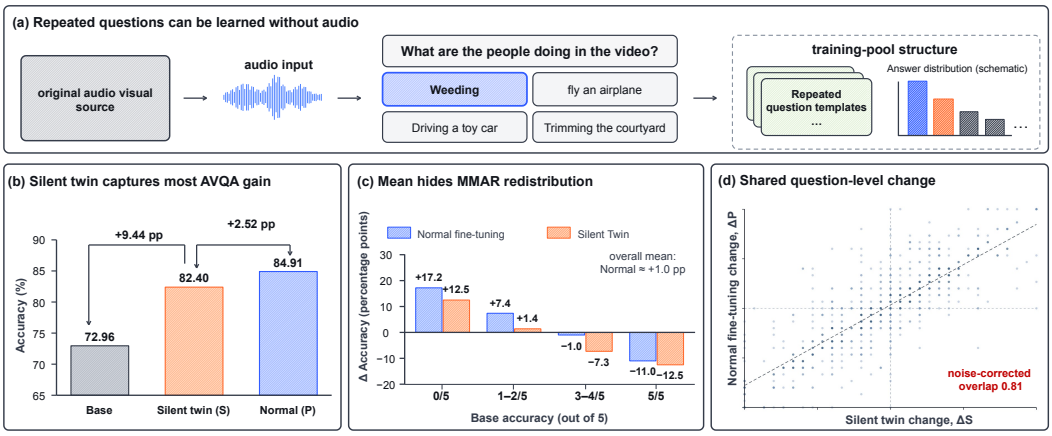


Figure 1: Silent twins reproduce much of what AVQA fine-tuning changes. (a) An AVQA training item and the pool structure. (b) AVQA test accuracy (greedy benchmark readout). (c) Change in per-question MMAR accuracy, grouped by how often Base answers correctly in five samples. (d) Per-question changes of normal and silent training on the 1,000 MMAR questions. Panels (c)–(d) use the sampled question-level readout; noise-corrected overlap = 0.81 (Section 3.2).

The matched silent update also provides a concrete reference for weight-space correction. We propose *silent-twin subtraction* (STS), which subtracts a weighted silent update from the normal update and independently controls the magnitude of the result: $\theta_0 + \beta(\Delta_P - \lambda\Delta_S)$. The coefficient λ sets the subtraction strength, and β sets the overall adaptation strength. This design controls the trained update directly while preserving the evaluation inputs and decoding procedure. An exact decomposition separates module-wise rescaling from an orthogonal residual, allowing us to examine the roles of scale and direction alongside transfer performance and retained AVQA gain.

The study combines cross-model measurement with a systematic evaluation of the correction. We reproduce the question-level phenomenon on three architectures and evaluate STS on Qwen2.5-Omni-7B (Xu et al., 2025). With coefficients selected per benchmark, STS improves over normal fine-tuning on all four transfer benchmarks and exceeds tuned scaling on three under within-benchmark cross-validation. The selected operating points quantify how transfer gains vary with in-task retention. Our contributions are as follows.

- We introduce a matched silent-twin control and a noise-corrected question-level readout for studying adaptation without the original training audio (Sections 3.1 and 3.2).
- We show that normal and silent fine-tuning induce highly overlapping question-level changes, including losses on Base-stable questions, across three architectures and three benchmarks (Sections 4.2 and 4.3).
- We develop STS and evaluate its transfer performance, coefficient selection and in-task retention against scaling and weight-space baselines (Section 4.4).

2 Related Work

Audio-language post-training spans instruction tuning (Tang et al., 2024; Chu et al., 2024; Xu et al., 2025) and reinforcement learning for question answering (Li et al., 2025; Rouditchenko et al., 2025; Wu et al., 2026), with AVQA used by several training recipes (Yu et al., 2026). Understanding the resulting changes connects three lines of work: input-reliance probes, training-time blind controls, and methods that edit fine-tuning updates.

2.1 Probing Input Reliance

Test-time probes ask whether a trained model relies on a given input, by withholding that input at evaluation and measuring what survives. Partial-input baselines are the classical instance of this idea.

Goyal et al. (2017) show that a question-only model already answers much of VQA, and Agrawal et al. (2018) shift the answer priors between splits to expose models that rely on them. Gururangan et al. (2018) and Poliak et al. (2018) find the same pattern for hypothesis-only inference, and Feng et al. (2019) show that a failed partial-input baseline does not certify a dataset free of artifacts. Audio-language studies carry this logic over to sound. They remove the audio (Foo et al., 2026), contradict it with the accompanying text (Wang et al., 2025a) or add irrelevant silence (Li et al., 2026), and AudioMCQ (He et al., 2026) splits training data by answers under silent audio. Related contrasts compare hidden states with and without the image (Long et al., 2026), or sharpen decoding against a noised image (Leng et al., 2024) or an all-zero waveform (Hsu et al., 2025). All of these tools probe a fixed model or item. They measure what a trained model relies on, not what a training update needed.

2.2 Training-time blind controls

Training-time blind controls ask what the supervision alone can teach, by withholding the input during training. Omni-R1 (Rouditchenko et al., 2025) drops the audio in reinforcement learning and finds that text-only fine-tuning also improves audio benchmarks. In vision, models are trained on black or blank images (Xiang et al., 2026; Zhang et al., 2026) or on randomly mismatched ones (Zhang et al., 2026), preference pairs are stripped of their images (Wang et al., 2024), and a text-only twin of a 3D language model filters benchmark questions and reweights the training loss (Ma et al., 2026). These controls do compare training updates, but the audio control changes the input format, and none of them compares the two runs' question-level changes or uses the control run as a weight-space reference.

2.3 Debiasing and weight-space editing

Debiasing and weight-space editing alter how adaptation affects model behavior. Training-time bias experts include RUBi (Cadene et al., 2019) and methods that fit the main model to examples poorly explained by a bias-only predictor (Clark et al., 2019; He et al., 2019). After training, Task Arithmetic (Ilharco et al., 2023) combines or negates task vectors, with related operations on parameter-efficient modules (Zhang et al., 2023). EWR (Daheim et al., 2024) and Bias Vector (Shirafuji et al., 2025) use negative experts, Ext-Sub (Hu et al., 2024) extracts a deficiency component before subtraction, and Fierro & Roger (2026) use the difference between fine-tuning runs. Other methods adjust the existing update through interpolation (Wortsman et al., 2022), layer-dependent scaling (Wang et al., 2025b), sparsification (Yu et al., 2024), or audio-model LoRA rescaling (Hu et al., 2026). These approaches provide the weight-space operations against which we evaluate a reference obtained from matched silent training.

Our contribution connects the training control, its behavioral readout and the resulting weight reference. The silent twin preserves the normal run's input format, the question-level analysis measures their shared changes, and STS tests how this matched reference can be used to correct the normal update.

3 Silent twins and silent-twin subtraction

3.1 Construction

For each normal fine-tuning run P, we train a silent twin S. The two runs share the examples, answers, example order, initial weights, optimizer, loss denominator, step budget and LoRA configuration; the only change is that each training waveform is replaced by zeros of the same duration (Figure 2). The audio encoder still processes an input, and the number and positions of audio tokens are unchanged, which differs from dropping the audio tokens (Rouditchenko et al., 2025), since that also changes sequence length and positions. Both models are evaluated on the real audio.

The twin records the adaptation that the matched recipe produces under silence, and evaluating both runs on real audio makes their difference a controlled comparison of training inputs. A change that the twin reproduces is attainable without the training audio; as with partial-input baselines, the inference is one-sided, and a change the twin misses is not thereby shown to need the sound (Feng et al., 2019). The twin is a reference for the update, not a decomposition of credit, since redundant

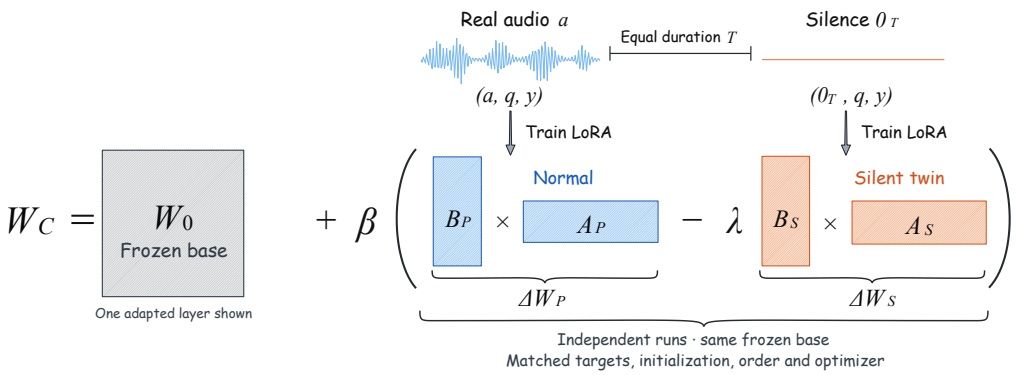


Figure 2: A silent twin repeats a fine-tuning run with each training waveform replaced by equal-duration silence; targets, initialization, example order and optimizer are shared. For the layer shown, W_0 is the frozen base weight, and $\Delta W_P = B_P A_P$, $\Delta W_S = B_S A_S$ are that module's blocks of Δ_P, Δ_S; the corrected weight W_C is one module of $\theta(\beta, \lambda)$ (Eq. 1).

information in the question and the audio has no unique split (Liang et al., 2023). Silence is also not a neutral input at inference (Li et al., 2026), so we use it only as a training condition and evaluate on the original recordings.

3.2 Question-level readout

We use complementary protocols for behavior and benchmark performance. The *sampled question-level readout* draws five samples per question at temperature 0.6 with a fixed option order, providing the repeated observations used to estimate question-level changes. The *greedy benchmark readout* uses greedy decoding and official scoring, averaging MMAR and MMAU-mini over five option orders. Both protocols evaluate the same adapters; each comparison uses a single readout throughout.

Three statistics characterize the changes from Base. The noise-corrected *RMS shift* measures their magnitude, and the *reproducible share* measures how much squared movement repeats across training seeds. The *overlap* is a normalized, noise-corrected inner product that captures alignment between the two arms' question-level changes. Products of distinct samples remove the corresponding sampling-noise terms. An independent Base panel defines the Base-stable questions, and separate Base samples measure ordinary accuracy differences within those strata (Appendix A).

3.3 The silent update as a weight reference

STS composes the normal and silent updates directly. Both runs start from the same Base weights θ_0 and adapt the same modules, giving their LoRA updates Δ_P and Δ_S shared coordinates. The corrected weights are

$$\theta(\beta, \lambda) = \theta_0 + \beta\,(\Delta_P - \lambda \Delta_S), \tag{1}$$

The two coefficients give separate controls over the composition. For a fixed λ, varying β changes the magnitude of the combined update while preserving its direction. Varying λ changes the contribution of the silent reference before this overall scaling. The family includes pure scaling at $\lambda = 0$ (Wortsman et al., 2022), the normal update at $(\beta, \lambda) = (1, 0)$, and Base at $\beta = 0$. These anchors support direct comparison under the same evaluation procedure.

Algorithm 1 implements the composition directly in the LoRA factors. The source adapters have ranks r_P, r_S and use $\alpha/r = 1$. Stacking their A factors and concatenating their weighted B factors gives $B_{C,m}A_{C,m} = \beta B_{P,m}A_{P,m} - \beta\lambda B_{S,m}A_{S,m}$. Setting $\alpha_C = r_C = r_P + r_S$ preserves this scaling. The same two trained adapters support every coefficient choice, and composition requires no additional optimization.

For each adapted module m, we split the silent update into its projection onto the normal update and a residual, $\Delta_{S,m} = c_m \Delta_{P,m} + R_m$ with $c_m = \langle \Delta_{S,m}, \Delta_{P,m} \rangle_F / \|\Delta_{P,m}\|_F^2$ and $\langle R_m, \Delta_{P,m} \rangle_F = 0$.

Algorithm 1 Matched training and exact LoRA composition for STS

Require: Base weights θ_0, data $\mathcal{D} = \{(a_i, q_i, y_i)\}$, matched seed ξ and training recipe $\mathcal{H}$, coefficients β, λ
Ensure: A single STS adapter for evaluation with real audio

$\mathcal{D}_S \leftarrow \{(0_{\mathrm{len}(a_i)}, q_i, y_i) : (a_i, q_i, y_i) \in \mathcal{D}\}$
$\{A_{P,m}, B_{P,m}\}_m \leftarrow \mathrm{TrainLoRA}(\theta_0, \mathcal{D}; \xi, \mathcal{H})$
$\{A_{S,m}, B_{S,m}\}_m \leftarrow \mathrm{TrainLoRA}(\theta_0, \mathcal{D}_S; \xi, \mathcal{H})$
for each adapted module m **do**
    $A_{C,m} \leftarrow \mathrm{vcat}(A_{P,m}, A_{S,m})$ ▷ Stack rows
    $B_{C,m} \leftarrow \mathrm{hcat}(\beta B_{P,m}, -\beta\lambda B_{S,m})$ ▷ Join columns
end for
$r_C \leftarrow r_P + r_S; \quad \alpha_C \leftarrow r_C$
return $\{A_{C,m}, B_{C,m}\}_m$ with rank r_C and scale $\alpha_C / r_C = 1$

Substituting into Eq. (1) gives the exact identity

$$\theta(\beta, \lambda) = \theta_0 + \sum_m \big[\beta(1 - \lambda c_m)\, \Delta_{P,m} - \beta\lambda R_m\big]. \tag{2}$$

Equation (2) expresses STS as module-wise rescaling plus an orthogonal residual. This motivates two comparisons: tuned uniform scaling tests the value of adding the silent reference, while matched module-wise rescaling isolates the residual at a given operating point. Section 4.4 reports the former; Appendix E gives the decomposition controls.

4 Experiments

We first measure the in-task gains attainable under silent training, then examine shared question-level changes across models, and finally evaluate STS through its transfer performance, coefficient selection and in-task retention.

4.1 Setup

Models and training. We fine-tune Qwen2.5-Omni-7B (Xu et al., 2025) with rank-32 LoRA (Hu et al., 2022) on the 112 attention projections of the Thinker's language model. AVQA (Yang et al., 2022) provides 40,174 training rows. Each run takes 1,255 steps at global batch 32, and each arm has three seeds. We repeat the full-pool recipe on Qwen2-Audio-7B-Instruct (Chu et al., 2024) and MiDashengLM-7B (Dinkel et al., 2025).

Evaluation. The in-task evaluation is the official AVQA test split of 16,790 items. Transfer is measured on MMAR (Ma et al., 2025), MMAU-mini (Sakshi et al., 2025), MMSU (Wang et al., 2026) and AIR-Bench (Yang et al., 2024). Accuracy tables use the greedy benchmark readout and question-level statistics the sampled question-level readout (Section 3.2). Two Base runs in the main batch differ by at most 0.08 points.

Coefficient selection. Each benchmark selects one configuration shared by the three training seeds. We report the best observed test-set configuration and five-fold cross-validation within the benchmark, selecting on four recording-grouped folds and evaluating on the fifth. STS and scaling use the same selection rule. The regular grid covers $\beta \in \{0.1, \ldots, 1\}$ and $\lambda \in \{0, 0.1, \ldots, 1.5\}$, with six additional scaling points: 166 configurations in total, including 16 pure scales. AIR-Bench uses a 56-cell grid on all its items. LiNeS (Wang et al., 2025b), Ext-Sub (Hu et al., 2024) and DARE on the LoRA factor A (Yu et al., 2024) each use six configurations. Appendices D and F specify the grids, fold construction and evaluation batches.

Corpus. Of the 16,790 test items, 98.6% repeat a training question string and 120 reuse a training clip; excluding the latter moves every arm by at most 0.2 points. The split thus tests new recordings under familiar wording (Appendix B).

Table 1: In-task gain and transfer after AVQA fine-tuning of Qwen2.5-Omni-7B (greedy benchmark readout, mean of three training seeds). **Bold** marks changes that stand clear of evaluation noise (Appendix A).

	In-task	Transfer			
	AVQA	MMAR	MMAU-mini	MMSU	AIR-Bench
Base accuracy (%)	72.96	61.58	74.96	63.66	69.98
Change from Base (points)					
Normal fine-tuning P	**+11.95**	−0.11	**−5.28**	+0.74	**+1.00**
Silent twin S	**+9.44**	**−3.52**	**−7.27**	−0.21	**−0.61**
$P - S$	**+2.52**	**+3.41**	**+1.99**	**+0.95**	**+1.62**

4.2 Silence recovers most of the in-task gain

AVQA fine-tuning produces a large in-task gain with a different pattern across transfer tasks. Accuracy rises from 72.96% to 84.91% on AVQA, while MMAU-mini falls by 5.28 points and AIR-Bench rises by 1.00 point; MMAR and MMSU change by -0.11 and $+0.74$ points (Table 1). The silent twin allows us to compare this adaptation with the changes produced by the same supervision under silence.

The silent twin gains 9.44 points on AVQA, recovering 79% of the normal run's gain. A substantial improvement is therefore attainable without the original training waveforms. Real-audio training provides an additional gain: normal fine-tuning exceeds the twin by 2.52 points on AVQA and by 0.95 to 3.41 points on the four transfer benchmarks. The paired contrasts in Table 1 separate these two observations: a large silent-training gain and a consistent additional benefit from real-audio training.

The training pool provides predictable structure alongside its recordings. Its 774 question strings repeat 52 times on average, and a question-only majority lookup reaches 52.3% leave-one-out accuracy against 25% chance, with balanced gold positions. Fine-tuning also improves answer formatting: responses that match no option fall from 45 of $16{,}790$ for Base to 7 for the twin and 8 for normal training. AVQA was annotated for audio-visual question answering; its metadata marks 99.3% of our rows as jointly audio-visual, and a vision-language model answers 76% of evaluation questions from a single muted frame (Kim & Seo, 2026). These corpus-level observations motivate the matched training comparison (Appendix B).

Table 2: Training controls under the greedy benchmark readout (one batch). Base accuracy in %; other rows give the change from Base in points, three-seed means. **Bold** marks changes that stand clear of evaluation noise. The second Base run differs from the first by at most 0.16 points.

	AVQA	MMAR	MMAU-mini	MMSU
Base (%)	73.03	61.56	74.98	63.68
Normal fine-tuning P	**+11.90**	−0.31	**−5.28**	+0.79
Silent twin S	**+9.40**	**−3.69**	**−7.31**	−0.24
Text-only training	**+9.26**	**−3.18**	**−6.60**	−0.63
Audio-contribution filter, 1,255 steps	**+10.24**	−0.23	**−7.78**	+0.44
Audio-contribution filter, 803 steps	**+10.57**	+0.64	**−6.84**	**+1.37**

The additional training controls reinforce this interpretation (Table 2). Text-only training gains 9.26 points on AVQA and stays within 0.8 points of the twin on the three transfer benchmarks evaluated in this batch. Similar gains with and without silent input support using the format-preserving twin as the matched reference. A second control applies the audio-contribution split of AudioMCQ (He et al., 2026) with one model, retaining rows that Base answers incorrectly under silence. Both the full-budget and one-pass versions improve AVQA while leaving a substantial MMAU-mini loss. Appendix G details these supervised controls and their relation to the cited recipes.

4.3 SILENT TWINS REPRODUCE THE QUESTION-LEVEL REARRANGEMENT

Question-level analysis exposes changes that the mean score obscures. Under the sampled question-level readout (Section 3.2), normal fine-tuning changes mean MMAR accuracy by $+1.0$ point, while the noise-corrected RMS shift is 0.332 and 88% of its squared magnitude reproduces across training seeds. Figure 1c shows the redistribution: questions that Base answers incorrectly in all five panel samples gain 17.2 points, whereas Base-stable questions lose 11.0 points.

The silent twin follows much of the same question-level pattern. Its RMS shift is 0.306, 93% of the squared movement reproduces across seeds, and accuracy on the same Base-stable questions falls by 12.5 points. The two arms' noise-corrected overlap is 0.81 (Figure 1d). Thus, both the redistribution and the loss on previously reliable questions are present in a run trained without the original audio.

Table 3: Silent twins across three architectures on the full pool (three seeds per arm, sampled question-level readout). Overlap is the noise-corrected overlap of the two arms' question-level changes (Section 3.2); the other columns are changes in points, the last two on Base-stable questions.

				Base-stable	
Model	Benchmark	Overlap	$P - S$	P	S
Qwen2.5-Omni-7B	MMAR	0.81	+3.9	−11.0	−12.5
	MMAU-mini	0.81	+3.2	−10.9	−13.6
	MMSU	0.73	+1.5	−3.1	−2.9
Qwen2-Audio-7B-Instruct	MMAR	0.85	+6.7	−21.1	−31.9
	MMAU-mini	0.77	+9.0	−12.3	−28.6
	MMSU	0.79	+4.1	−8.0	−16.7
MiDashengLM-7B	MMAR	0.87	+4.5	−13.5	−22.1
	MMAU-mini	0.86	+6.6	−11.9	−21.1
	MMSU	0.81	+2.1	−5.1	−7.4

The finding holds across three architectures (Table 3; full readout in Appendix C). Qwen2-Audio differs from Qwen2.5-Omni in both the audio encoder and the language model, and MiDashengLM in the audio encoder. Their overlaps lie between 0.77 and 0.87, on par with Qwen2.5-Omni-7B (0.73 to 0.81). Both arms lose accuracy on Base-stable questions in every row, the twin loses more in eight of the nine, and normal fine-tuning exceeds its twin in every row, by 1.5 to 9.0 points.

Whether normal fine-tuning inherits this loss depends on the training pool (Table 4). On a second pool selected for strong audio contribution (Appendix G), the silent twin still loses 9.8 points on Base-stable questions, but the normal run loses only 1.4, and the overlap of the two runs falls from 0.81 to 0.65. Where the labels demand the sound, the audio-bearing run no longer follows its twin.

Table 4: Normal fine-tuning and its silent twin on MMAR for two training pools (sampled question-level readout, three seeds per arm; changes from Base in points; Base-stable columns as in Table 3). The AVQA row repeats Table 3.

	Normal P		Silent twin S			
Training pool	all	Base-stable	all	Base-stable	Overlap	$P - S$
AVQA	+1.0	−11.0	−2.9	−12.5	0.81	+3.9
AudioMCQ strong-audio	+3.2	−1.4	−0.6	−9.8	0.65	+3.8

4.4 SILENT-TWIN SUBTRACTION

With benchmark-specific coefficients, STS improves transfer over normal fine-tuning on all four evaluated benchmarks (Table 5, top): by 2.33 points on MMAR, 5.77 on MMAU-mini, 0.94 on MMSU and 0.92 on AIR-Bench. The largest gain recovers the MMAU-mini loss introduced by fine-tuning and brings accuracy above Base. Among the evaluated configurations, STS also reaches the highest accuracy on MMAR, MMAU-mini and MMSU compared with LiNeS, Ext-Sub and DARE on A.

Table 5: Transfer accuracy of corrections to the AVQA-fine-tuned Qwen2.5-Omni-7B (%, greedy benchmark readout, three-seed means). Upper block: best configuration of each method on each benchmark's test set. Lower block: selected out of fold by grouped five-fold cross-validation; parentheses give the retained AVQA gain. **Bold**: best entry per block. Protocols: Section 4.1 and Appendix F.

	Transfer accuracy (%)			
Method	MMAR	MMAU-mini	MMSU	AIR-Bench
Best configuration on each test set				
Base	61.58	74.96	63.66	69.98
Normal fine-tuning P	61.47	69.68	64.40	70.98
LiNeS (Wang et al., 2025b)	63.19	74.77	64.82	–
Ext-Sub (Hu et al., 2024)	62.70	73.00	65.14	–
DARE on A (Yu et al., 2024)	62.80	73.31	64.83	–
Scaling	63.24	75.04	64.88	71.87
STS (ours)	**63.80**	**75.45**	**65.34**	**71.90**
Grouped 5-fold cross-validation				
Scaling	62.84 (74%)	74.59 (28%)	64.50 (84%)	**71.87** (84%)
STS (ours)	**63.80** (72%)	**75.12** (0%)	**64.83** (61%)	71.83 (82%)

STS outperforms scaling in each of the three training seeds on MMAR, MMAU-mini and MMSU (Table 6), with mean out-of-fold gains of 0.96, 0.53 and 0.33 points. Both families select coefficients on four recording-grouped folds and are evaluated on the fifth. MMAR selects $(0.4, 0.3)$ in all five folds, with seed-level gains of $+0.68$, $+1.26$ and $+0.94$ points. On MMAU-mini and MMSU, the gains remain positive across seeds despite variation in the selected coefficients. On AIR-Bench, STS and scaling score 71.83% and 71.87%, respectively. On the strong-audio pool of Appendix G, where the normal run carries none of its twin's loss, subtracting the twin lowers MMAR accuracy by 3.5 points in every seed: STS helps where normal fine-tuning has inherited the silent-attainable loss.

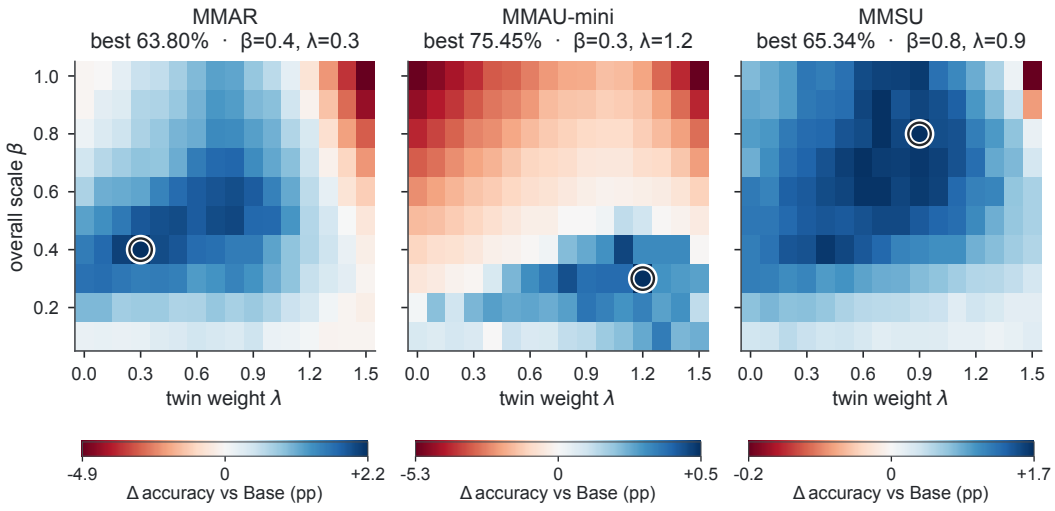


Figure 3: Response surfaces of STS on three transfer benchmarks: change in accuracy from Base (greedy benchmark readout) over the 160 regular cells of the (β, λ) grid, panel-specific color scale ($\lambda = 0$: pure scaling). The circled cell is the best test-set configuration with $\lambda > 0$; the same three cells score 81.58, 72.37 and 78.00% on AVQA.

The response surfaces identify distinct operating regions (Figure 3). MMAR favors a modest subtraction near $\beta = 0.4$. MMAU-mini favors smaller overall updates and stronger subtraction, peaking at $(0.3, 1.2)$, while MMSU reaches its maximum at $(0.8, 0.9)$. The two coefficients therefore con-

Table 6: STS gains across training seeds (greedy benchmark readout). Left: most frequently selected coefficient pair and its fold count. Right: out-of-fold STS-minus-scaling differences; positive values favor STS. Each seed's score averages all held-out questions.

	Coefficient selection		STS gain over scaling (points)			
Benchmark	Most selected (β, λ)	Folds selecting this pair	Average	Seed 1	Seed 2	Seed 3
MMAR	$(0.4, 0.3)$ $(0.3, 1.2)$ or	5/5	$+0.96$	$+0.68$	$+1.26$	$+0.94$
MMAU-mini	$(0.4, 1.1)$	2/5 each	$+0.53$	$+0.14$	$+1.04$	$+0.42$
MMSU	$(0.8, 0.9)$	2/5	$+0.33$	$+0.06$	$+0.34$	$+0.60$
AIR-Bench	$(0.5, 0.3)$	3/5	-0.04	$+0.05$	-0.08	-0.10

trol a task-dependent balance rather than a single fixed correction. Appendix E also evaluates the common operating point $(0.5, 1)$ across benchmarks.

In-task retention provides a second coordinate for evaluating these operating points (Table 5, bottom). On MMAR, STS retains 72% of the AVQA gain against 74% for scaling, so its largest cross-validated advantage comes at nearly equal retention. On MMSU the corresponding shares are 61% and 84%. On MMAU-mini, scaling retains 28%, while the selected STS cells return AVQA performance to approximately Base. Reporting both transfer accuracy and retention makes the consequence of each coefficient choice explicit.

The decomposition helps interpret how the silent reference changes the update. Module m receives the scale $\beta(1-\lambda c_m)$, with greater shrinkage where the silent and normal updates align more strongly. At the common operating point $(0.5, 1)$, matched module-wise rescaling retains 81% of the AVQA gain; adding the residual changes MMAR and MMSU by $+0.07$ and $+0.09$ points. The residual adds 1.14 points on MMAU-mini, where this operating point was selected, and changes AIR-Bench by -0.39 points. These controls distinguish the contributions of scale and direction at the evaluated operating point (Appendix E).

5 Conclusion

Matched silent twins make a substantial part of audio-language post-training directly observable. On AVQA, silent training recovers most of the in-task gain, and question-level measurements reveal shared changes and losses on Base-stable questions across three architectures. These findings motivate evaluating post-training through both aggregate gains and the individual decisions that change. The performance difference between the matched runs quantifies the additional gain obtained with real-audio training under the same recipe.

STS turns this training reference into a concrete weight-space intervention. Its benchmark-specific operating points improve on normal fine-tuning across the evaluated transfer tasks and exceed tuned scaling on three under grouped cross-validation, with distinct in-task retention profiles. Together, the control and the intervention connect measurement of adaptation with adjustment of the resulting update. Reporting silent-twin performance alongside normal fine-tuning, and transfer performance alongside in-task retention, provides a more informative account of what audio post-training achieves. A silent twin costs one additional fine-tuning run, and we recommend training one beside every audio post-training run.

Limitations. Our experiments cover multiple-choice audio understanding under supervised LoRA fine-tuning on one pool, with the phenomenon measured on three architectures and the correction on one; the pool's test split shares most of its question strings with training. Coefficients are selected on each target benchmark by cross-validation whose folds were fixed after the grids had been inspected, with three training seeds. The gain over scaling is benchmark-specific and paid for in in-task accuracy, and restricting the training audio the model can observe does not enlarge it (Appendix H); the strong-audio comparison rests on one pool and one benchmark. Open-ended responses and reinforcement-learning recipes are left to future work.

AI USE STATEMENT

In this work, we used generative AI tools for retrieval and discovery, to search for and summarize related work; for research execution, to implement and refine analysis and evaluation code, which we reviewed before each run; to aid and polish the writing; and to draft sections of the paper, which we revised. We have not used generative AI tools to generate data or experimental results. We have reviewed all AI-assisted work and take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

The study uses only publicly released datasets, benchmarks and model checkpoints. It examines audio-only training on AVQA, a dataset designed for audio and video; the audit does not assess the dataset's value in its intended setting. No new participant data were collected for this study.

REPRODUCIBILITY STATEMENT

Sections 3 and 4.1 describe the control, the readouts and the coefficient selection; Appendices A, D, E and F give the protocol, the estimator, the grids, every intervention arm and the baselines. The supplementary material contains the fine-tuning code; the evaluation code, benchmark manifests with content hashes, per-question outputs and coefficient grids will be released with the paper.

REFERENCES

Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. Don't just assume; look and answer: Overcoming priors for visual question answering. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4971–4980, 2018. doi: 10.1109/CVPR.2018.00522.

Remi Cadene, Corentin Dancette, Hedi Ben-younes, Matthieu Cord, and Devi Parikh. RUBi: Reducing unimodal biases for visual question answering. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, 2019.

Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, et al. Qwen2-Audio technical report. *arXiv preprint arXiv:2407.10759*, 2024.

Christopher Clark, Mark Yatskar, and Luke Zettlemoyer. Don't take the easy way out: Ensemble based methods for avoiding known dataset biases. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2019. arXiv:1909.03683.

Nico Daheim, Nouha Dziri, Mrinmaya Sachan, Iryna Gurevych, and Edoardo M. Ponti. Elastic weight removal for faithful and abstractive dialogue generation. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pp. 7096–7112, 2024. doi: 10.18653/v1/2024.naacl-long.393.

Heinrich Dinkel, Gang Li, Jizhong Liu, Jian Luan, Yadong Niu, Xingwei Sun, Tianzi Wang, Qiyang Xiao, Junbo Zhang, and Jiahao Zhou. MiDashengLM: Efficient audio understanding with general audio captions. *arXiv preprint arXiv:2508.03983*, 2025.

Shi Feng, Eric Wallace, and Jordan Boyd-Graber. Misleading failures of partial-input baselines. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2019. arXiv:1905.05778.

Constanza Fierro and Fabien Roger. Steering language models with weight arithmetic. In *International Conference on Learning Representations (ICLR)*, 2026. URL `https://arxiv.org/abs/2511.05408`. arXiv:2511.05408.

Leonardo Haw-Yang Foo, Chih-Kai Yang, Chen-An Li, Ke-Han Lu, and Hung-yi Lee. All that glitters is not audio: Rethinking text priors and audio reliance in audio-language evaluation. In *Interspeech*, pp. 3825–3830, 2026. doi: 10.21437/Interspeech.2026-913. arXiv:2604.24401.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6325–6334, 2017. doi: 10.1109/CVPR.2017.670.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel R. Bowman, and Noah A. Smith. Annotation artifacts in natural language inference data. In *Proceedings of NAACL-HLT*, 2018. arXiv:1803.02324.

Haolin He, Xingjian Du, Renhe Sun, Zheqi Dai, Yujia Xiao, Mingru Yang, Jiayi Zhou, Xiquan Li, et al. Measuring audio's impact on correctness: Audio-contribution-aware post-training of large audio language models. In *International Conference on Learning Representations (ICLR)*, 2026. URL https://openreview.net/forum?id=sJ0jUO9Mxr. Numbers cited from arXiv:2509.21060v4.

He He, Sheng Zha, and Haohan Wang. Unlearn dataset bias in natural language inference by fitting the residual. In *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo)*, 2019. arXiv:1908.10763.

Tzu-wen Hsu, Ke-Han Lu, Cheng-Han Chiang, and Hung-yi Lee. Reducing object hallucination in large audio-language models via audio-aware decoding. In *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 1–7, 2025. doi: 10.1109/ASRU65441.2025.11434595. arXiv:2506.07233.

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 (ICLR)*, 2022. arXiv:2106.09685.

Weiteng Hu, Yin Cao, and Jun Yang. Reasoning-oriented post-training and inference-time LoRA rescaling for audio-dependent question answering. *arXiv preprint arXiv:2608.23092*, 2026.

Xinshuo Hu, Dongfang Li, Baotian Hu, Zihao Zheng, Zhenyu Liu, and Min Zhang. Separate the wheat from the chaff: Model deficiency unlearning via parameter-efficient module operation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 18252–18260, 2024. doi: 10.1609/aaai.v38i16.29784.

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Suchin Gururangan, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. Editing models with task arithmetic. In *International Conference on Learning Representations (ICLR)*, 2023. arXiv:2212.04089.

Geewook Kim and Minjoon Seo. Do modern video-LLMs need to listen? a benchmark audit and scalable remedy. In *Interspeech*, pp. 1126–1130, 2026. doi: 10.21437/Interspeech.2026-2532. arXiv:2509.17901.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with PagedAttention. In *Proceedings of the 29th Symposium on Operating Systems Principles (SOSP)*, pp. 611–626, 2023. doi: 10.1145/3600006.3613165.

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13872–13882, 2024. doi: 10.1109/CVPR52733.2024.01316.

Chen-An Li, Tzu-Han Lin, and Hung-yi Lee. When silence matters: The impact of irrelevant audio on text reasoning in large audio-language models. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 17757–17761, 2026. doi: 10.1109/ICASSP55912.2026.11463785. arXiv:2510.00626.

Gang Li, Jizhong Liu, Heinrich Dinkel, Yadong Niu, Junbo Zhang, and Jian Luan. Reinforcement learning outperforms supervised fine-tuning: A case study on audio question answering. *arXiv preprint arXiv:2503.11197*, 2025.

Paul Pu Liang, Yun Cheng, Xiang Fan, Chun Kai Ling, Suzanne Nie, Richard Chen, Zihao Deng, Nicholas Allen, Randy Auerbach, Faisal Mahmood, Ruslan Salakhutdinov, and Louis-Philippe Morency. Quantifying & modeling multimodal interactions: An information decomposition framework. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pp. 27351–27393, 2023.

Lin Long, Changdae Oh, Seongheon Park, and Sharon Li. Understanding language prior of LVLMs by contrasting chain-of-embedding. In *International Conference on Learning Representations (ICLR)*, 2026. URL https://openreview.net/forum?id=wnTnwWKu72. arXiv:2509.23050.

Xianzheng Ma, Tao Sun, Shuai Chen, Yash Bhalgat, Jindong Gu, Angel X. Chang, Iro Armeni, Iro Laina, Songyou Peng, and Victor Adrian Prisacariu. Do 3D large language models really understand 3D spatial relationships? In *International Conference on Learning Representations (ICLR)*, 2026. URL https://openreview.net/forum?id=3vlMiJwo8b. arXiv:2603.23523.

Ziyang Ma, Yinghao Ma, Yanqiao Zhu, Chen Yang, Yi-Wen Chao, Ruiyang Xu, Wenxi Chen, Yuanzhe Chen, et al. MMAR: A challenging benchmark for deep reasoning in speech, audio, music, and their mix. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, volume 38, pp. 74837–74868, 2025. doi: 10.52202/085713-2252.

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. Hypothesis only baselines in natural language inference. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics (*SEM)*, 2018. arXiv:1805.01042.

Andrew Rouditchenko, Saurabhchand Bhati, Edson Araujo, Samuel Thomas, Hilde Kuehne, Rogerio Feris, and James Glass. Omni-R1: Do you really need audio to fine-tune your audio LLM? In *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 1–7, 2025. doi: 10.1109/ASRU65441.2025.11434780. arXiv:2505.09439.

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. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://openreview.net/forum?id=TeVAZXr3yv.

Daiki Shirafuji, Makoto Takenaka, and Shinya Taguchi. Bias vector: Mitigating biases in language models with task arithmetic approach. In *Proceedings of the International Conference on Computational Linguistics (COLING)*, 2025. arXiv:2412.11679.

Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, and Chao Zhang. SALMONN: Towards generic hearing abilities for large language models. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://openreview.net/forum?id=14rn7HpKVk.

Cheng Wang, Gelei Deng, Xianglin Yang, Han Qiu, and Tianwei Zhang. When audio and text disagree: Revealing text bias in large audio-language models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 4878–4888, 2025a. doi: 10.18653/v1/2025.emnlp-main.246.

Dingdong Wang, Junan Li, Jincenzi Wu, Dongchao Yang, Xueyuan Chen, Tianhua Zhang, and Helen Meng. MMSU: A massive multi-task spoken language understanding and reasoning benchmark. In *International Conference on Learning Representations (ICLR)*, 2026. URL https://openreview.net/forum?id=yHzCDP1tXw.

Fei Wang, Wenxuan Zhou, James Y. Huang, Nan Xu, Sheng Zhang, Hoifung Poon, and Muhao Chen. mDPO: Conditional preference optimization for multimodal large language models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 8078–8088, 2024. doi: 10.18653/v1/2024.emnlp-main.460.

Ke Wang, Nikolaos Dimitriadis, Alessandro Favero, Guillermo Ortiz-Jimenez, Francois Fleuret, and Pascal Frossard. LiNeS: Post-training layer scaling prevents forgetting and enhances model merging. In *International Conference on Learning Representations (ICLR)*, 2025b. arXiv:2410.17146.

Mitchell Wortsman, Gabriel Ilharco, Jong Wook Kim, Mike Li, Simon Kornblith, Rebecca Roelofs, Raphael Gontijo-Lopes, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, and Ludwig Schmidt. Robust fine-tuning of zero-shot models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7959–7971, 2022.

Shu Wu, Chenxing Li, Wenfu Wang, Hao Zhang, Hualei Wang, Meng Yu, and Dong Yu. Audio-Thinker: Guiding large audio language model when and how to think via reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 33962–33970, 2026. doi: 10.1609/aaai.v40i40.40689. Extended version: arXiv:2508.08039v3.

Kun Xiang, Terry Jingchen Zhang, Zirong Liu, Bokai Zhou, Yueling Tang, Junjie Yu, Jiacong Lu, Shangrui Huang, Heng Li, Likui Zhang, Kunkun Liu, Changzheng Zhang, Yangle Fang, Boqiang Guo, Hui-Ling Zhen, Dandan Tu, Yinya Huang, and Xiaodan Liang. SeePhys Pro: Diagnosing modality transfer and blind-training effects in multimodal RLVR for physics reasoning. *arXiv preprint arXiv:2605.09266*, 2026.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, et al. Qwen2.5-Omni technical report. *arXiv preprint arXiv:2503.20215*, 2025.

Pinci Yang, Xin Wang, Xuguang Duan, Hong Chen, Runze Hou, Cong Jin, and Wenwu Zhu. AVQA: A dataset for audio-visual question answering on videos. In *Proceedings of the 30th ACM International Conference on Multimedia (ACM MM)*, pp. 3480–3491, 2022. doi: 10.1145/3503161.3548291.

Qian Yang, Jin Xu, Wenrui Liu, Yunfei Chu, Ziyue Jiang, Xiaohuan Zhou, Yichong Leng, Yuanjun Lv, Zhou Zhao, Chang Zhou, and Jingren Zhou. AIR-Bench: Benchmarking large audio-language models via generative comprehension. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 1979–1998, 2024. doi: 10.18653/v1/2024.acl-long.109.

Fangxu Yu, Tao Feng, Dehai Min, Zinan Lin, Weijia Xu, Michael Xu, Philip S. Yu, Ge Liu, et al. Reinforcement learning with evolving rubrics as rewards for audio reasoning. *arXiv preprint arXiv:2608.02831*, 2026.

Le Yu, Bowen Yu, Haiyang Yu, Fei Huang, and Yongbin Li. Language models are super mario: Absorbing abilities from homologous models as a free lunch. In *International Conference on Machine Learning (ICML)*, pp. 57755–57775, 2024.

Gengwei Zhang, Jie Peng, Zhen Tan, Mufan Qiu, Hossein Nourkhiz Mahjoub, Vaishnav Tadiparthi, Kwonjoon Lee, Yanyong Zhang, and Tianlong Chen. Understanding the role of hallucination in reinforcement post-training of multimodal reasoning models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 25326–25335, 2026. arXiv:2604.03179.

Jinghan Zhang, Shiqi Chen, Junteng Liu, and Junxian He. Composing parameter-efficient modules with arithmetic operations. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. arXiv:2306.14870.

A Protocol and question-level estimator

A.1 Two readouts

The sampled question-level readout uses Transformers with five samples per question at temperature 0.6 and a fixed option order, supplying RMS shift, reproducible share, overlap and Base-stable comparisons. The greedy benchmark readout uses vLLM (Kwon et al., 2023), greedy decoding and official scoring, averaging MMAR and MMAU-mini over five option orders. Both evaluate the same adapters, with contrasts formed within each protocol. For example, normal fine-tuning changes MMAR by $+1.0$ point under the sampled readout and -0.11 under the greedy readout.

A.2 Generation and engine

Greedy generation uses at most 512 new tokens and a context of $16{,}384$ tokens. Audio longer than 300 s is cut to 300 s, as the Qwen2.5-Omni feature extractor does. One process occupies one GPU. The engine loads adapters at run time, and no merged checkpoint is written. We check every adapter's digest before use, because the engine silently ignores an adapter whose module names do not match instead of rejecting it. Raw generations are stored once and scored when tables are built.

A.3 Paired comparisons

A second Base run measures process-to-process variation. In tables of paired changes, bold marks a contrast whose sign is stable under resampling of the evaluation questions: the three training seeds are first averaged within each question, and questions are then resampled with the arms kept paired. Seed-to-seed variation is reported separately. In the method-comparison table, bold instead marks the largest accuracy within each block.

A.4 Question-level estimator

Let $X_{a,s,i,t} \in \{0,1\}$ mark a correct answer for arm a, training seed s, item i and decoding seed t, and let B^{pan} be an independent Base panel. Define

$$C_i(U,V) = |D|^{-1} \sum_{(t,u)\in D} U_{it} V_{iu},$$

where D contains only pairs of distinct decoding-seed indices, and $Q_i(U,V) = C_i(U,V) - C_i(U,B^{\mathrm{pan}}) - C_i(V,B^{\mathrm{pan}}) + C_i(B^{\mathrm{pan}},B^{\mathrm{pan}})$. For each arm let $T_a = \langle Q_i(X_{a,s},X_{a,s})\rangle_{i,s}$ and $G_a = \langle Q_i(X_{a,s},X_{a,s'})\rangle_{i,s<s'}$, with $\langle\cdot\rangle$ the mean over the indicated indices. We report $\mathrm{RMS}_a = \sqrt{\max(T_a,0)}$ and the reproducible share G_a/T_a. The overlap is

$$O_{P,S} = \frac{\langle Q_i(X_{P,s},X_{S,s'})\rangle_{i,s,s'}}{\sqrt{G_P G_S}},$$

whose numerator averages all nine training-seed pairs. These are uncentered second moments across items, not Pearson covariances. Independent Base samples, not B^{pan}, supply ordinary accuracy differences. B^{pan} also defines the Base-stable strata. Restricting the numerator to distinct training-seed pairs changes the main MMAR estimate from 0.806 to 0.803.

A.5 Training

LoRA with rank 32, $\alpha = 32$ and dropout 0.05 adapts the query, key, value and output projections of every self-attention layer of the Thinker's language model (112 modules in Qwen2.5-Omni-7B). Each run takes $1{,}255$ optimizer steps at global batch 32 (8 GPUs, gradient accumulation 4) with a constant learning rate of 10^{-4}. The loss is cross-entropy on the answer tokens and the end-of-sequence token under a global token denominator. Both arms share the schedule, and the last 14 rows of the ordered pool are not exposed. The replications use the same design.

A.6 Benchmark versions

Our scores use the prompts, decoding and scoring rules specified above. MMAU-mini uses the v05.15.25 annotation file, and MMSU uses the question file preceding the 2026 parquet release,

which revised some answers and reordered the options. Published numbers obtained with other prompt, decoding or annotation versions provide protocol-specific reference points.

B CORPUS AND TEST-SPLIT AUDIT

The training pool has 40,174 eligible rows and 774 distinct question strings, each repeated 52 times on average. The question-only majority-answer lookup reaches 52.3% leave-one-out accuracy against a 25.0% chance floor, and gold positions are balanced, so the lookup does not exploit option order. The metadata field `question_relation` marks 99.3% of rows as jointly audio-visual. This is a field value of the original annotation, not evidence about individual clips. The most frequent question string, "What are the people doing in the video?", occurs 4,294 times in the 40,174 rows. The item in Figure 1(a) is one of them.

On the official test split (16,790 items), 98.6% of items repeat a training question string and 120 reuse a training clip. Excluding those 120 items, or the 343 items whose option list repeats a string, leaves every arm within 0.2 points. Fine-tuning also teaches the answer format: Base gives 45 responses that match no option, P gives 8, S 7 and the fixed subtraction F of Appendix E 57.

The 1,000 AVQA-derived items of AIR-Bench (Yang et al., 2024) repeat a training question string in 98.7% of cases, and the lookup is correct on 547 of them. Table 7 splits the arms by whether the lookup is correct. On these 1,000 items the overall difference between normal and silent training (+0.50) is not resolved. The 16,790-item test split resolves it (+2.52).

Table 7: AVQA-derived AIR-Bench items split by whether the question-only lookup is correct. Change from Base in points.

	Lookup correct ($n = 547$, Base 82.3)	Lookup wrong ($n = 453$, Base 68.7)
Normal P	+4.63	+6.03
Silent twin S	+5.67	+3.68
Scale $0.5P$	+5.06	+6.11
STS F	−0.49	+2.13
$P - S$	−1.04	+2.35

C REPLICATION SETTINGS

C.1 FULL READOUT

Table 8 expands the Qwen2.5-Omni-7B rows of Table 3 under the sampled question-level readout. RMS, reproducible share and overlap are estimated over each complete benchmark. The independent Base panel defines the strata used for the ordinary accuracy comparisons. MMAR and MMSU use five Base-panel samples. MMAU-mini uses ten independent Base samples, so its four stratum columns correspond to $0/10$, 1–$4/10$, 5–$9/10$ and $10/10$. MMSU uses the official letter-only prompt and scorer on all 5,000 questions. The $P - S$ entries are differences between the two arms' seed-mean accuracies.

D COEFFICIENT SELECTION

D.1 GRIDS

The regular search uses $\beta \in \{0.1, \ldots, 1\}$ and $\lambda \in \{0, 0.1, \ldots, 1.5\}$ plus six extra scaling points, giving 166 configurations with three seeds each. AIR-Bench uses the same β values and $\lambda \in \{0, 0.3, 0.6, 0.9, 1.2\}$ plus the six extra scales, giving 56 configurations. All 24,683 AIR-Bench items are scored with the official rule, and its folds group 20,914 recordings. The AIR-Bench comparison and predictions were specified before evaluation of this grid. The observed selections favor small subtraction weights and scales in 0.25–0.75, with out-of-fold accuracies of 71.83% for STS and 71.87% for scaling.

Table 8: Full sampled question-level readout of the Qwen2.5-Omni-7B rows of Table 3 (three training seeds; changes are seed means in points). Strata group questions by how many samples of an independent Base panel are correct (n per stratum in each block header). RMS, Share and Overlap are as defined in Section 3.2.

Arm	Acc. (%)	Δ all	Δ by Base-panel stratum				RMS	Share	Overlap
			0/5	1–2/5	3–4/5	5/5			
Qwen2.5-Omni-7B, 40,174 rows, 3 seeds; MMAR; Base 57.6%; strata n = 240 / 194 / 168 / 398									
Normal P	58.6	+1.0	+17.2	+7.4	−1.0	−11.0	0.332	0.88	
Silent twin S	54.7	−2.9	+12.5	+1.4	−7.3	−12.5	0.306	0.93	
$P - S$		+3.9							0.81
Qwen2.5-Omni-7B, 40,174 rows, 3 seeds; MMAU-mini; Base 73.9%; strata n = 114 / 141 / 158 / 587									
Normal P	67.9	−6.0	+9.7	+12.3	−15.5	−10.9	0.315	0.89	
Silent twin S	64.8	−9.2	+7.7	+9.1	−21.3	−13.6	0.319	0.92	
$P - S$		+3.2							0.81
Qwen2.5-Omni-7B, 40,174 rows, 3 seeds; MMSU, official protocol; Base 62.73%; strata n = 1,068 / 872 / 614 / 2,446									
Normal P	63.85	+1.1	+9.6	+5.9	−3.5	−3.1	0.221	0.90	
Silent twin S	62.35	−0.4	+5.4	+2.2	−4.3	−2.9	0.169	0.91	
$P - S$		+1.5							0.73

D.2 Test-set optima

The full search family contains the scaling subset. Table 9 distinguishes the best scaling cell from the best strictly subtractive cell ($\lambda > 0$), including the AVQA row. The selected configurations describe different operating points: applying the MMAU-mini choice $(0.3, 1.2)$ gives 62.09% on MMAR, 64.15% on MMSU and 72.37% on AVQA, compared with 72.96% for AVQA Base. On MMAR the best scaling cells, β=0.3 and β=0.39, tie at 63.24%, and the paired seed differences between $(0.4, 0.3)$ and $(0.3, 0)$ are $+0.42$, $+1.06$ and $+0.20$ points (Figure 4).

Table 9: Test-set optima of the coefficient grid (greedy benchmark readout, accuracy in %, three-seed means). "Best STS" is the best cell with $\lambda > 0$; the difference is the paired three-seed mean.

	Best scaling		Best STS		
Benchmark	accuracy	β	accuracy	(β, λ)	Difference
MMAR	63.24	0.3 or 0.39	63.80	$(0.4, 0.3)$	$+0.56$
MMAU-mini	75.04	0.1	75.45	$(0.3, 1.2)$	$+0.41$
MMSU	64.88	0.5	65.34	$(0.8, 0.9)$	$+0.46$
AIR-Bench	71.87	0.4	71.90	$(0.5, 0.3)$	$+0.03$
AVQA (training task)	84.91	0.9	84.73	$(0.9, 0.1)$	-0.18

D.3 Cross-validated selection

We evaluate within-benchmark selection using one fixed five-fold partition. The partition was specified after inspection of the full grids. A benchmark-specific hash assigns recording groups to folds: MMAR has 911 recordings, while MMAU-mini and MMSU have one recording per question. Each family selects the highest mean accuracy over the four selection folds and the three seeds, breaking ties toward smaller λ and then smaller β. Base is included as a candidate alongside the grid cells. Each question contributes once to the out-of-fold score; the selected cells can differ across folds. Table 6 gives selection frequencies and paired seed contrasts.

D.4 Selection across folds

STS exceeds scaling by 0.96, 0.53 and 0.33 points on MMAR, MMAU-mini and MMSU, with positive differences across the three training seeds. MMAR selects $(0.4, 0.3)$ in every fold; MMSU and MMAU-mini select four and three distinct cells. On AIR-Bench, three folds select $(0.5, 0.3)$ and two select scaling, giving a mean contrast of -0.04 points. The resulting in-task retention shares are reported with the aggregate scores in Table 5. This evaluation uses the existing trained adapters and one fixed partition of each benchmark.

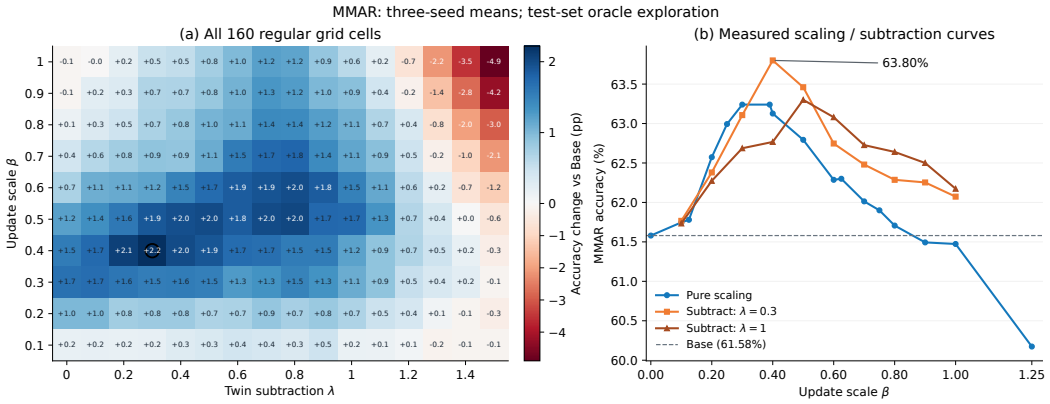


Figure 4: MMAR coefficient grid (test-set values; greedy benchmark readout). (a) Change from Base over the 160 regular (β, λ) cells; the circle marks the test-set maximum, $(0.4, 0.3)$ at 63.80%. (b) All measured pure-scaling points, including the six extra scales and Base, with the λ=0.3 and λ=1 slices; connecting segments are visual guides.

E FIXED COEFFICIENT AND DECOMPOSITION CONTROLS

This appendix reports the fixed coefficient $F = (0.5, 1)$, the controls that separate the two terms of the decomposition in Section 3.3, and the full results of every intervention arm.

Table 10: Fixed coefficients (same readout and batch as Table 1). Change from Base in points. "Retained" is the share of P's mean AVQA gain. $F = (0.5, 1)$ was selected on MMAU-mini and then frozen. Bold as in Table 1.

	In-task		Transfer			
	AVQA	retained	MMAR	MMAU-mini	MMSU	AIR-Bench
Normal P	**+11.95**	100%	−0.11	**−5.28**	+0.74	**+1.00**
Scale $0.5P$	**+10.82**	91%	+1.21	**−1.68**	**+1.22**	**+1.85**
Scale $0.25P$	**+7.55**	63%	+1.41	−0.29	**+0.97**	**+1.47**
STS F	**+2.87**	24%	**+1.72**	+0.02	**+1.33**	**+1.45**

E.1 TRANSFER AT A FIXED COEFFICIENT

The common operating point $F = (0.5, 1)$ was selected on MMAU-mini. Relative to $0.5\Delta_P$, it improves that benchmark by 1.70 points and changes MMAR, MMSU and AIR-Bench by $+0.51$, $+0.11$ and -0.40 points (Table 10). Its AVQA gain retention is 24%, compared with 91% for $0.5\Delta_P$. These paired results locate the fixed correction in both transfer accuracy and in-task retention.

E.2 CONTRIBUTION OF THE ORTHOGONAL RESIDUAL

Holding the per-module rescaling of F fixed and removing only the residual isolates the residual's contribution. The residual adds $+1.14$ points on the development benchmark. On transfer it adds $+0.07$ on MMAR and $+0.09$ on MMSU, and it changes AIR-Bench by -0.39. On the development benchmark the residual's two leading singular directions carry its whole effect. On AIR-Bench they do worse than norm-matched random directions. On AVQA, the matched rescaling, a norm-matched random residual and a spectrum-preserving shuffle of the residual all keep about 81% of the in-task gain. Their mean transfer gains are 1.47 to 1.51 points. The two leading residual directions reach 1.57 but keep only 40% (Table 13).

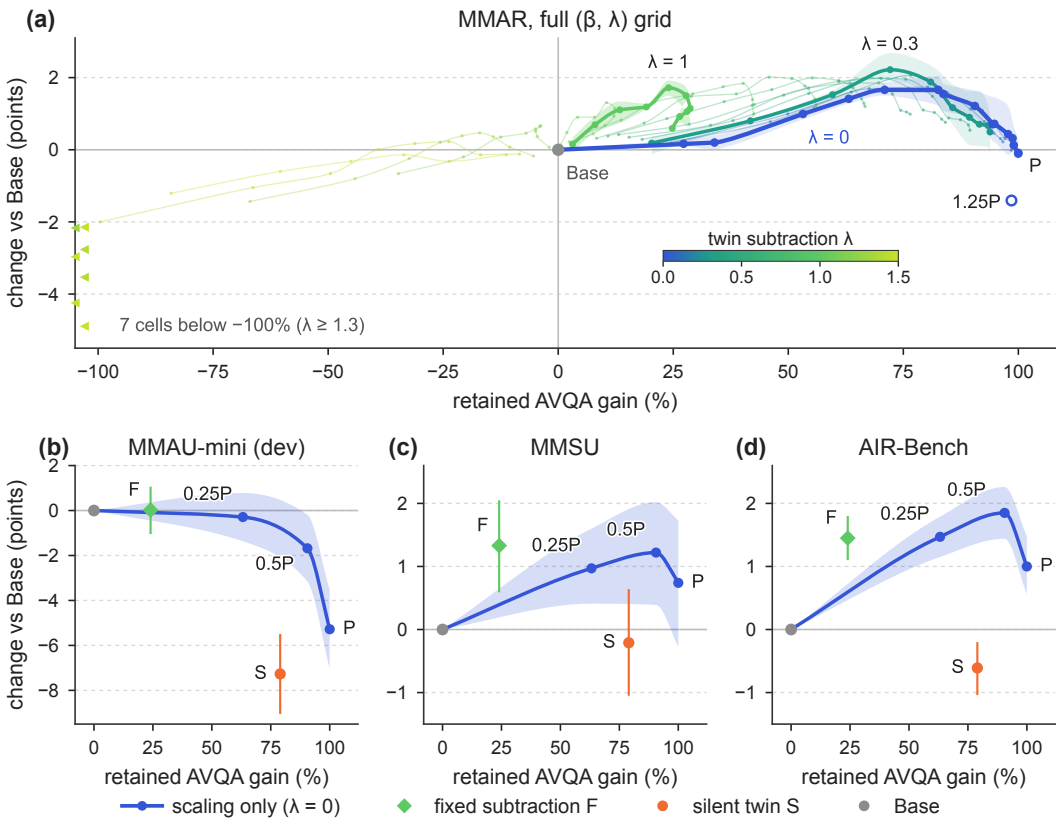


Figure 5: STS against scaling by retained AVQA gain. (a) All 166 cells of the MMAR coefficient grid (test-set values); x-axis: share of the AVQA gain retained; y-axis: change from Base. Each line joins one λ's cells in order of β (color); the blue line is pure scaling. Bold lines mark $\lambda = 0.3$ (the cross-validation-selected slice) and $\lambda = 1$ (the frozen F). Shading spans the three training seeds; seven cells below -100% are drawn as triangles at the left edge. (b–d) The fixed arms of Table 10 on the other three benchmarks, with $F = (0.5, 1)$; bars and shading show the spread over evaluation questions.

E.3 Full results

Table 11 gives the absolute accuracies, per-seed values and changes of every arm in Tables 1 and 10. The three AVQA accuracies of F span 3.4 points, compared with at most 0.8 for the other arms, so its retained gain is also reported for each seed. Table 12 gives the direction controls, evaluated in a separate batch.

E.4 Direction-control construction

In Table 12, matched rescaling retains the along-P component of F, while the direction controls vary its orthogonal residual. Random residuals are norm-matched. The scaling controls are recomposed from the same source adapters, and F reuses the adapter of Table 11. The rank sweep is evaluated on MMAU-mini, where $k = 2$ is fixed before transfer evaluation. Other top-, bottom- and random-rank controls are evaluated only on MMAU-mini. Dashes in the transfer-contrast rows mark comparisons outside the reported set; the additional MMAU-mini comparisons are given below.

E.5 Residual and directions

On the development benchmark of Table 12 (MMAU-mini, 1,000 questions $\times$ five option orders), Base scores 74.94, the matched rescaling 73.87 and the full subtraction 75.01. The top-1, top-2,

Table 11: Full version of Tables 1 and 10 (main batch, greedy benchmark readout; MMAR and MMAU-mini averaged over five option orders). Accuracy (%), per-seed accuracies and changes from Base (points). **Bold** as in Table 1. "Second run − first" is the run-to-run noise floor; "Retained" is the share of P's AVQA gain kept, per seed against the same seed's P gain.

	In-task	Transfer			
Arm	AVQA test 16,790 items	MMAR 1,000 × 5 orders	MMAU-mini 1,000 × 5 orders	MMSU 5,000	AIR-Bench 24,683
Base (%)	72.96	61.58	74.96	63.66	69.98
Base, second run (%)	72.97	61.66	74.96	63.62	70.01
second run − first	+0.01	+0.08	+0.00	−0.04	+0.03
Normal P (%)	84.91	61.47	69.68	64.40	70.98
per seed	85.11/84.92/84.71	61.22/62.18/61.02	69.76/71.08/68.20	64.06/64.74/64.40	71.00/71.31/70.63
vs Base	**+11.95**	−0.11	**−5.28**	+0.74	**+1.00**
Silent twin S (%)	82.40	58.06	67.69	63.45	69.36
per seed	82.70/82.57/81.92	57.42/57.94/58.82	66.64/68.02/68.42	63.62/63.46/63.28	69.14/69.63/69.31
vs Base	**+9.44**	**−3.52**	**−7.27**	−0.21	**−0.61**
retained % (per seed)	79.0 (80.2/80.4/76.3)				
Scale $0.5P$ (%)	83.78	62.79	73.28	64.88	71.83
per seed	83.73/83.72/83.90	62.70/63.22/62.46	73.24/73.64/72.96	64.90/65.02/64.72	71.79/71.98/71.71
vs Base	**+10.82**	+1.21	**−1.68**	**+1.22**	**+1.85**
retained % (per seed)	90.5 (88.6/90.0/93.1)				
Scale $0.25P$ (%)	80.51	62.99	74.67	64.63	71.44
per seed	80.20/80.47/80.85	63.00/63.12/62.86	74.54/75.06/74.42	64.76/64.52/64.62	71.29/71.54/71.50
vs Base	**+7.55**	+1.41	−0.29	**+0.97**	**+1.47**
retained % (per seed)	63.2 (59.6/62.8/67.1)				
STS $F = (0.5, 1)$ (%)	75.83	63.30	74.98	64.99	71.42
per seed	74.40/75.31/77.78	63.50/63.32/63.08	74.76/75.48/74.70	64.64/65.20/65.14	71.25/71.51/71.51
vs Base	**+2.87**	**+1.72**	+0.02	**+1.33**	**+1.45**
retained % (per seed)	24.0 (11.9/19.6/41.0)				
$P - S$	**+2.52**	**+3.41**	**+1.99**	**+0.95**	**+1.62**
$F - 0.5P$ (same step)	**−7.96**	+0.51	**+1.70**	+0.11	**−0.40**
$F - 0.25P$	**−4.68**	+0.31	+0.31	+0.36	−0.02
$F - P$	**−9.09**	+1.83	**+5.30**	+0.59	**+0.44**

Table 12: Direction controls for the fixed STS point $F = (0.5, 1)$ (greedy benchmark readout, three-seed means). Base accuracy is in %; other entries are changes from Base or the stated paired contrasts, in points. All comparisons use a separate evaluation batch. Bold as in Table 1.

Arm	MMAR	MMAU-mini	MMSU	AIR-Bench
Base (%)	61.72	74.94	63.68	70.03
second run − first	+0.00	−0.02	+0.06	−0.04
Matched rescaling	+1.42	−1.07	**+1.24**	**+1.82**
STS F (full subtraction)	**+1.49**	+0.07	**+1.33**	**+1.42**
Top-2 residual directions	**+1.70**	+0.04	**+1.43**	**+1.57**
Random rank-2	+1.45	−1.10	+1.13	+1.82
Spectrum shuffle	+1.46	−1.21	+1.21	+1.85
Scale $0.25P$	+1.24	−0.25	+1.01	+1.44
Scale $0.5P$	+1.07	−1.64	+1.05	+1.83
Residual: F − matched	+0.07	**+1.14**	+0.09	**−0.39**
Top-2 − random rank-2	+0.25	–	+0.30	**−0.25**
Top-2 − F	+0.21	–	+0.10	**+0.14**
Matched − $0.25P$	+0.18	–	+0.23	**+0.38**
Matched − $0.5P$	+0.35	–	+0.19	−0.01

top-4 and top-16 residual directions score 74.64, 74.98, 74.93 and 75.01. Relative to Base, the top-1, top-4, top-16 and bottom-2 changes are −0.30, −0.01, +0.07 and −1.09 points. A spectrum-preserving shuffle scores 73.73 and a norm-matched random rank-2 residual 73.84. On MMAR, MMSU and AIR-Bench, the top-2 directions minus the norm-matched random residual give +0.25, +0.30 and −0.25.

E.6 RANK COMPARISONS ON MMAU-MINI

Full subtraction exceeds spectrum shuffle and random rank-32 by 1.27 and 1.09 points, respectively. The top-versus-random contrasts at ranks 1, 4 and 16 are $+0.78$, $+1.22$ and $+1.25$ points. The top-versus-bottom contrasts are $+1.12$ at rank 4 and $+1.25$ at rank 16; these last two comparisons are not norm-matched.

E.7 AVQA DIRECTION-CONTROL RESULTS

In this batch, Base scores 72.98% on the 16,790 AVQA test items. The recomposed normal update ($\beta = 1$), matched rescaling, top-2 residual and full subtraction improve on Base by 11.91, 9.72, 4.77 and 2.82 points, respectively. Adding the residual to matched rescaling changes accuracy by -6.90 points. A second Base evaluation differs by $+0.01$ point. Absolute accuracies for the random, shuffled and scaling arms are reported in Table 13.

E.8 RETENTION AT EQUAL IN-TASK COST

Table 13 evaluates the same arms on the AVQA test split in the direction-control batch. The transfer column averages the changes on MMAR, MMSU and AIR-Bench.

Table 13: In-task retention and transfer for scaling and residual controls on the AVQA test split (direction-control batch, Table 12). "Retained" is the share of P's mean AVQA gain. "Transfer mean" is the mean change on MMAR, MMSU and AIR-Bench. –: not evaluated.

Arm	AVQA (%)	Retained (%)	Transfer mean
Scale 0.75	84.76	98.7	–
Scale 0.625	84.29	94.8	–
Scale 0.5	83.78	90.5	+1.32
Scale 0.39	82.83	82.6	–
Spectrum-preserving shuffle	82.72	81.6	+1.51
Matched rescaling	82.70	81.4	+1.49
Norm-matched random rank-2	82.70	81.4	+1.47
Scale 0.25	80.51	63.1	+1.23
Top-2 residual directions	77.75	40.0	+1.57
Scale 0.125	77.02	33.9	–
STS F (full subtraction)	75.80	23.6	+1.41

E.9 DAMAGE TO ALREADY-CORRECT ANSWERS

On the 11,547 AVQA test questions that both Base and P answer correctly, full subtraction loses 4.12 points and the matched rescaling 1.20. Unparseable answers rise from 7 under the matched rescaling to 56 under full subtraction.

F WEIGHT-SPACE BASELINES

We apply three weight-space methods to the same normal and silent LoRA updates. LiNeS (Wang et al., 2025b) rescales the normal update by a depth-dependent factor $\alpha + (\tau - \alpha)\, l/(L-1)$ for layer l of L. We use $\alpha \in \{0, 0.2, 0.4\}$ and $\tau \in \{0.6, 1\}$. Ext-Sub (Hu et al., 2024) projects each row of the silent update onto the bisector of the two rows' unit vectors and subtracts η times the remainder from the normal update. In closed form its result is $\beta\,[(1+\eta\kappa_i/2)\,\Delta_P - (\eta/2)\,\Delta_S]$, with $\kappa_i = \|\Delta_{S,i}\|/\|\Delta_{P,i}\|$ the norm ratio of the i-th rows. Ext-Sub is therefore the family of Section 3.3 with a row-dependent scale on Δ_P. We use $\beta \in \{0.5, 0.75, 1\}$ and $\eta \in \{0.5, 1\}$. DARE (Yu et al., 2024) randomly drops delta parameters and rescales the rest. We apply it to the LoRA factor A. A share p of its entries is dropped at random, and the rest are rescaled by $1/(1-p)$ before scaling by β. We use $p \in \{0.5, 0.9\}$. Every configuration is composed for the three training seeds.

Table 5 reports each method at its best evaluated test-set configuration. For LiNeS, the selected (α, τ) are $(0.2, 0.6)$, $(0, 0.6)$ and $(0.2, 1)$ on MMAR, MMAU-mini and MMSU. Ext-Sub selects

$\beta = 0.5$ with $\eta = 0.5$, 0.5 and 1; DARE on A selects $\beta = 0.5$ with $p = 0.9$, 0.9 and 0.5. The search uses six configurations per baseline, 16 for scaling and 166 for STS. STS has the highest observed accuracy on these three benchmarks. The baseline evaluation batch has Base accuracies 61.64, 74.98 and 63.88, while the grid batch has 61.58, 74.96 and 63.66; all use the greedy benchmark readout.

G TRAINING CONTROLS

Two further training arms follow the recipe of Appendix A on the same pool and schedule. They are evaluated in one batch with the normal run, the silent twin and a second Base run (Table 2). *Text-only* training removes the audio input altogether. It is the no-audio condition of Omni-R1 (Rouditchenko et al., 2025) under our supervised recipe. The user turn holds only the question and the options, and rows, order, answer tokens, LoRA configuration, steps and seeds match the normal run. *Strong-audio-contribution* training keeps only the rows that Base answers incorrectly under equal-duration silence. This applies the weak/strong split of AudioMCQ (He et al., 2026), a three-model majority under 30-second silence, with a single model. Of the $40{,}174$ rows, $25{,}720$ qualify. Base answers $14{,}365$ rows correctly under silence and gives no parseable silent answer on 89. The kept rows are cycled to fill the same $1{,}255$ steps, and we also evaluate the checkpoint after one pass (803 steps). Table 2 in the main text reports both controls.

Text-only and silent training yield similar benchmark gains, with differences below 0.8 points in Table 2. The silent twin additionally preserves the normal run's audio pathway and token positions. Selecting rows by audio contribution produces a smaller AVQA gain and a larger MMAU-mini loss than normal fine-tuning, while MMAR and MMSU remain close or improve. These controls instantiate the cited input interventions within our matched supervised recipe.

G.1 A STRONG-AUDIO TRAINING POOL

To test whether the loss on Base-stable questions is tied to the AVQA pool, we repeated the twin construction on the strong-audio-contribution subset of AudioMCQ (He et al., 2026), whose items were selected because models fail them under silent audio, with three seeds per arm and the readout, estimator and Base panel of Table 3. Table 4 in the main text places the MMAR results next to the AVQA row. On the strong-audio pool the normal run improves on Base by 3.2 points and loses 1.4 points on Base-stable questions, whereas its silent twin loses 9.8 points there and 0.6 overall; the normal run exceeds the twin by 3.8 points overall and by 8.5 on Base-stable questions, and the overlap of the two runs is 0.65. Subtracting this twin from the normal update lowers MMAR accuracy by 3.5 points in each seed.

H RESTRICTED OBSERVATION

H.1 OPERATOR

To vary the share of audio that training observes, we keep a share ρ of each recording's audio feature frames and replace the rest by silence. The audio front end of Qwen2.5-Omni represents a typical AVQA recording as about $1{,}000$ valid feature frames. For a dose ρ we keep $\lfloor \rho\, n \rfloor$ of the n valid frames, chosen in a fixed hashed block order with at most one block kept partially. Each remaining frame is replaced in place by the corresponding feature frame of equal-duration silence. Frame count, positions, attention mask, text tokens and supervision tokens are unchanged, so $\rho = 1$ is normal training and $\rho = 0$ is the silent twin. No label or model score enters the choice of frames. The share is a training-time observation budget, not a measure of acoustic evidence.

Table 14: Restricted observation ($1{,}256$-step protocol, greedy benchmark readout; Base: AVQA 72.98%, MMAR 61.74%). Change from Base in points, three-seed means.

Share of audio frames observable in training	AVQA test	MMAR
$\rho = 1$ (all frames)	$+11.90$	-0.45
$\rho = 0.5$ (half of the frames)	$+12.03$	-0.30
$\rho = 0$ (none; the silent twin)	$+9.35$	-3.63

H.2 PROTOCOL

Each dose is trained with three seeds on the full $40{,}174$-row pool for $1{,}256$ steps, the final partial batch included. The recipe of Appendix A is otherwise unchanged. This protocol has its own full-audio reference and is never mixed with the $1{,}255$-step runs of the main text. Both evaluations (Table 14) use the greedy benchmark readout on MMAR ($1{,}000$ questions, five option orders) and on the AVQA test split ($16{,}790$ items). Base scores 61.74% and 72.98%, and a second Base run differs by 0.10 and 0.08 points.

H.3 STS AT EACH DOSE

For each dose, we evaluate $\theta_0 + \gamma\Delta_{P_\rho}$ at nineteen scales from 0 to 1 and $\theta_0 + 0.4(\Delta_{P_\rho} - \lambda\Delta_S)$ at eleven λ values from 0 to 1, where Δ_S is the $\rho = 0$ update. The fixed $\beta = 0.4$ comes from the MMAR grid. Relative to this protocol's full-audio run, AVQA loss budgets of 3 and 5 points give STS-minus-scaling contrasts of $+0.47$ and $+0.37$ points at $\rho = 1$, and $+0.01$ and $+0.25$ at $\rho = 0.5$. These contrasts are smaller than the seed-to-seed variation. At $\rho = 0$, the update is $0.4(1-\lambda)\Delta_S$, so STS follows a scaling path; the best feasible cells coincide under the 5-point budget. The evaluated $\rho = 0$ family lies outside the 3-point budget. At $\lambda = 0.3$, the contrast changes by -0.23 points from $\rho = 1$ to $\rho = 0.5$.