CAN TEXT CONNECT IMAGES AND AUDIO WITHOUT PAIRED TRAINING DATA? A CONTROLLED STUDY

Anonymous authors
Paper under double-blind review

ABSTRACT

Adding a new kind of data to an AI system usually requires paired examples and joint retraining. We test a simpler route. One stage learns from images and captions. A separate stage learns from audio and captions. We then compare images with sounds, even though these small connectors never trained on matched image–sound pairs. On a noisy 4,411-row proxy built from AudioCaps video thumbnails, the best system reaches a 4.13% average retrieval score. That is about 34 times chance, but only 16% of the score from ImageBind, a large system trained directly on image–audio pairs. The proxy contains a signal, but it is too weak for practical search. Changing the text model does not fix the gap. The data source and audio encoder can change results sharply, and a more flexible lightweight method does not win consistently. The lesson is simple: this staged approach can produce a weak match without matched image–audio pairs in its connector training, but text does not guarantee a useful connection. Builders must test their data, encoder, and evaluation pairs.

1 INTRODUCTION

AI models that work with pictures, text, and sound are often called *multimodal* models. They are usually trained with examples that directly pair these kinds of data. Adding sound to an image system may therefore require matched image–sound examples and another round of joint training.

We ask whether there is a simpler route. Imagine that a photo of a dog and a recording of its bark are never shown together during training. The photo can still be paired with the caption "a dog barking," and the recording can be paired separately with a similar caption. Can text connect them later?

Our experiment follows that idea. First, we train small image and text connectors on image–caption pairs. We then keep them unchanged, or *frozen*. Next, we train an audio connector on audio–caption pairs. During this second stage, the audio side never sees an image. At test time, we ask whether a sound can find the thumbnail from the same video. This no-pair restriction applies to the small connectors we train; we did not audit all data used earlier to pretrain CLIP and CLAP.

This setting is stricter than systems such as ImageBind, which use direct paired data when adding a modality (Girdhar et al., 2023). It could also be useful: different teams could train their small connectors separately and combine them later. Our central question is:

> *Can separately trained image and audio systems connect through text, and is the resulting connection strong enough to use?*

We find an above-chance matching signal, but not a reliable search system. The best version reaches 4.13% on our average retrieval score: much better than the 0.121% chance score, but far below the 25.66% ImageBind reference. Changing the text model does not close the gap. Matching the training source to the test source helps, and changing the audio encoder can have a large effect. A CLAP-based caption ranking and a LoRA-style proxy do not behave as expected.

The main lesson is that using text does not guarantee a useful connection between separately trained systems. A builder must check the data source, encoder, and test pairs rather than assume that two good text links will automatically combine.

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.

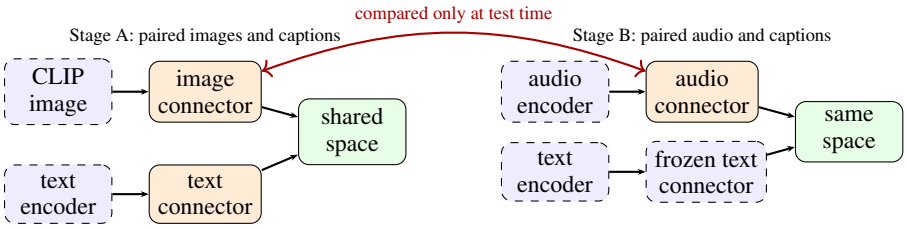

Figure 1: Stage A learns from images and captions. Stage B learns from audio and captions while Stage A stays unchanged. Only at test time do we compare images with audio. Dashed boxes are kept unchanged.

2 Related Work

Systems trained together. CLIP and ALIGN connect images with language (Radford et al., 2021; Jia et al., 2021). AudioCLIP trains with image, audio, and text together (Guzhov et al., 2022), while ImageBind connects six kinds of data through images (Girdhar et al., 2023). These systems aim for strong joint representations. We instead ask what happens when the image and audio parts are trained separately.

Text as a connector. CLAP connects audio with language (Elizalde et al., 2023), and LanguageBind connects several data types through language (Zhu et al., 2024). These results show that text can organize different kinds of data. We study the narrower case in which one connection is frozen before the other is trained.

Small trainable connectors. Adapters can extend a frozen model without retraining every parameter (Sung et al., 2022). We compare a learned linear connector, a fixed random connector, and a low-rank change to projected text features that we call a LoRA-style proxy. This proxy is not LoRA applied to the CLIP or CLAP encoder. The appendix gives their exact definitions and cites the random-projection methods on which the second design is based.

3 How the Study Works

Two separate stages. Stage A uses COCO image–caption pairs (Lin et al., 2014). It learns small image and text connectors and then freezes them. Stage B learns an audio connector from audio–caption pairs while reusing the frozen text connector. The image–audio test score is never used to train the system or choose a checkpoint. We use CLIP for images and text and CLAP for environmental audio unless a comparison says otherwise.

Data and comparisons. The main audio data come from AudioCaps (Kim et al., 2019). We use WavCaps to check whether training and testing on the same source matters (Mei et al., 2024). AVCaps provides a small second test set (Sudarsanam et al., 2025). SPEECH-COCO lets us compare CLAP with the Whisper speech encoder while keeping everything else fixed (Havard et al., 2017; Radford et al., 2023). Exact sample counts, training settings, repeated runs, and model definitions are in Appendix A.

What the score means. For each query, we ask whether the correct match appears among the first 1, 5, or 10 results. We test both search directions and average the six values. We call this the *mean-recall score*. Higher is better.

Why this is only a proxy test. AudioCaps provides audio and captions, not a hand-checked image for each sound. We use one default YouTube thumbnail from the same video. Of 4,411 rows, 45 use a gray placeholder because the thumbnail could not be retrieved. A matching video ID does not guarantee that the thing making the sound is visible. The test can reveal an exact-row matching signal, but it cannot establish useful semantic search by itself.

4 RESULTS

4.1 THE PROXY TEST CONTAINS A SIGNAL, BUT THE SCORE IS LOW

Table 1: Image–audio mean recall on the 4,411-row AudioCaps proxy at the largest tested setting (512 dimensions). Modular results are mean ± standard deviation over five repeated runs. ImageBind is an older same-gallery scale reference, not a matched baseline.

Method	Image–audio pairs in connector training	Mean-recall score
Chance	no	0.121%
Linear connector	no	$\mathbf{4.13} \pm 0.26\%$
Shared random connector	no	$3.84 \pm 0.21\%$
LoRA-style proxy	no	$3.93 \pm 0.26\%$
ImageBind-Huge (Girdhar et al., 2023)	yes, at large scale	25.66%

The best modular score is about 34 times chance, showing that the complete staged system contains a measurable matching signal. It is also only about 16% of the ImageBind score. ImageBind is larger and uses direct image–audio training, so this is not a fair contest between methods. It is a scale check showing that our system is not ready for practical retrieval.

4.2 WHAT CHANGES THE RESULT—AND WHAT DOES NOT

Changing the text model does not fix the gap. One concern was that CLIP text might be the wrong connector for audio. We repeated both stages with CLAP text while keeping the examples, training budget, method, and random seeds fixed. Across the tested model sizes, CLAP text does not provide a consistent improvement. This does not show that CLIP text is always better; it shows that a simple switch to CLAP text does not solve this problem. Appendix B reports the complete comparison.

Matching the data source helps. A model trained on AudioCaps does best when tested on AudioCaps, while one trained on WavCaps does best when tested on WavCaps. In both directions, matching the source more than doubles the audio–text score. This shows that the source matters, but not whether the reason is caption style, recording conditions, topics, or something else. Exact values are in Appendix C.

The proposed caption score does not support the quality story. We ranked AudioCaps pairs by how well CLAP thought each sound matched its caption. If this score predicted retrieval performance in the proposed way, the top group should have helped most. Instead, the bottom group did better. We have no human quality labels, so this does not show which captions are better. It shows only that the CLAP score does not provide the proposed ordering of retrieval results. Appendix C gives the full results.

The LoRA-style proxy is not a consistent winner. The linear connector has the highest mean score on AudioCaps, while the LoRA-style proxy has the highest mean on the smaller AVCaps test. The differences are small, and the ordering changes with the test set. A separate test also shows that its starting values do not explain the pattern. We therefore claim neither a LoRA advantage nor a known cause. See Appendix D.

The audio encoder can change the result sharply. In a test using spoken captions, replacing the general-audio CLAP encoder with the speech-focused Whisper encoder raises the image–audio score by more than fourteen times at the largest setting. Everything else is held fixed. This shows that the encoder can matter greatly in one controlled speech test, not that Whisper is always better. Exact values are in Appendix E.

Across our existing experiments, image–audio scores are associated with both the image–text and audio–text links more strongly than with the audio–text link alone. Because this pattern is fit and

measured on the same experiments, we treat it only as a description; Appendix F contains the equation, fit values, and exclusions.

5 What We Learned

- **What worked:** the full staged system produces an above-chance exact-row matching signal without image–audio pairs in its connector training.
- **What did not fix it:** changing the text model and using a more flexible LoRA-style proxy do not produce a reliable gain. A proposed model-based caption score does not predict the retrieval ordering.
- **What remains uncertain:** the source and encoder clearly change results, but our tests do not isolate why. The main gallery uses noisy thumbnails and repeated video IDs, and we have not completed a human check of whether each sound source is visible. New datasets, models, and natural image–audio pairs require new tests.

These limits matter to anyone building a modular system. Two components may line up without new image–audio pairs in their connector training, but the connection is too fragile to trust without testing the particular data source, encoder, and evaluation pairs.

6 Conclusion

Our staged text-based system leaves an above-chance signal on the proxy test, but it does not guarantee a useful connection. Because we have no randomized text-anchor control, we do not claim that text alone caused the signal. The score remains low, the benchmark is noisy, and several intuitive fixes fail. The practical result is therefore a warning as much as a possibility: modular connection without new image–audio pairs in the connector stages may work in principle, but it is not reliable in practice without careful testing.

AI use statement

Generative AI tools were used in three ways. First, they aided and polished the writing, including drafting, simplifying, and editing the manuscript. Second, they supported retrieval and discovery, including finding candidate related work; the authors checked the cited sources and did not treat AI output as a citation source. Third, they supported research ideation and execution, including inspecting existing project artifacts, reconstructing reported aggregates, organizing analyses, editing tables, and creating the protocol diagram. The tools did not generate new experimental data. The authors remain responsible for every claim, citation, figure, and disclosure.

Ethics statement

This work uses previously released datasets and introduces no new human-subject study. AudioCaps comes from online videos, so missing thumbnails, removed content, and unequal visibility can bias the evaluation. We do not infer personal attributes or release new video content. We report the unfinished human thumbnail audit rather than replacing it with automated labels.

Reproducibility statement

Appendix A reports data splits, training settings, seeds, and checkpoint selection. The supplementary manifest identifies all 169 run artifacts. The supplied script rebuilds the reported summaries from 538 archived seed records, but it does not rerun training or recreate frozen features. Full replay requires the archived anonymous inputs.

References

Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. CLAP: Learning audio concepts from natural language supervision. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 1–5, 2023. doi: 10.1109/ICASSP49357.2023.10095889.

Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. ImageBind: One embedding space to bind them all. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15180–15190, 2023.

Andrey Guzhov, Federico Raue, Jörn Hees, and Andreas Dengel. AudioCLIP: Extending CLIP to image, text and audio. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 976–980, 2022. doi: 10.1109/ICASSP43922.2022.9747631.

William Havard, Laurent Besacier, and Olivier Rosec. SPEECH-COCO: 600k visually grounded spoken captions aligned to MSCOCO data set. In *GLU 2017 International Workshop on Grounding Language Understanding*, pp. 42–46, 2017. doi: 10.21437/GLU.2017-9.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 4904–4916. PMLR, 2021.

William B. Johnson and Joram Lindenstrauss. Extensions of Lipschitz mappings into a Hilbert space. *Contemporary Mathematics*, 26:189–206, 1984. doi: 10.1090/conm/026/737400.

Daniel M. Kane and Jelani Nelson. Sparser Johnson–Lindenstrauss transforms. *Journal of the ACM*, 61(1):1–23, 2014. doi: 10.1145/2559902.

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*, pp. 119–132. Association for Computational Linguistics, 2019. doi: 10.18653/v1/N19-1011.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In *European Conference on Computer Vision*, volume 8693 of *Lecture Notes in Computer Science*, pp. 740–755. Springer, 2014. doi: 10.1007/978-3-319-10602-1_48.

Xinhao Mei, Chutong Meng, Haohe Liu, Qiuqiang Kong, Tom Ko, Chengqi Zhao, Mark D. Plumbley, Yuexian Zou, and Wenwu Wang. WavCaps: A ChatGPT-assisted weakly-labelled audio captioning dataset for audio-language multimodal research. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:3339–3354, 2024. doi: 10.1109/TASLP.2024.3419446.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, 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 *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 28492–28518. PMLR, 2023.

Parthasaarathy Sudarsanam, Irene Martín-Morató, Aapo Hakala, and Tuomas Virtanen. AVCaps: An audio-visual dataset with modality-specific captions. *IEEE Open Journal of Signal Processing*, 6: 691–704, 2025. doi: 10.1109/OJSP.2025.3578296.

Yi-Lin Sung, Jaemin Cho, and Mohit Bansal. VL-Adapter: Parameter-efficient transfer learning for vision-and-language tasks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5227–5237, 2022.

Bin Zhu, Bin Lin, Munan Ning, Yang Yan, Jiaxi Cui, HongFa Wang, Yatian Pang, Wenhao Jiang, Junwu Zhang, Zongwei Li, Wancai Zhang, Zhifeng Li, Wei Liu, and Li Yuan. LanguageBind: Extending video-language pretraining to N-modality by language-based semantic alignment. In *International Conference on Learning Representations*, 2024. URL `https://openreview.net/forum?id=QmZKc7UZCy`.

A Reproducibility and Evaluation Details

Training. The refreshed R1–R8 suite uses frozen feature tensors and projection-head training only. CLIP is `openai/clip-vit-base-patch32` at revision `3d74acf9`; CLAP is `laion/clap-htsat-unfused` at revision `8fa0f1c6`; the speech intervention uses `openai/whisper-small` and a hashed 125,000-row feature artifact. Default CLIP image/text features are the 768/512-dimensional vision/text pooler outputs; CLAP audio uses `get_audio_features`. R1 instead uses official `get_text_features` outputs for both anchors. Inputs are projected and then ℓ_2-normalized by every head.

Both phases minimize symmetric InfoNCE with a learned logit scale initialized to $1/0.07$. AdamW uses learning rate 10^{-3}, weight decay 10^{-4}, cosine decay, unit gradient clipping, and early-stopping patience 8. Phase-A/Phase-B maxima are 60/40 epochs; train/evaluation batches are 2,048/4,096. Phase A selects image-to-text validation R@10; Phase B selects audio-to-text validation R@10; image–audio test performance is never used for selection. The linear model uses bias-free modality matrices. Shared JL follows sparse Johnson–Lindenstrauss random projections (Johnson & Lindenstrauss, 1984; Kane & Nelson, 2014): it pads every input to 768, then applies one fixed matrix with seed 42, $\epsilon = 0.1$, and 10 independently signed entries of magnitude $1/\sqrt{10}$ per input column before learned bias-free $m \times m$ heads. The LoRA proxy uses rank 16, scale $1/16$, Gaussian down-projection initialization, and a zero-initialized up projection; it is learned with the image head in Phase A and frozen in Phase B while the audio matrix is learned.

Budgets and seeds. R1, R4, and R7 use 45,178 AudioCaps Phase-B rows and seeds 0–4. R2 quality arms each contain 10,681 AudioCaps rows; the source crossing uses 39,940 rows per source after overlap filtering. R5 uses 100,000 SPEECH-COCO train pairs, a 20,000-pair Phase-B validation split, 5,000 held-out evaluation pairs, and seeds 0–2. Embedding dimensions are 64, 128, 256, and 512, except the R5 matched encoder control, which uses 128, 256, and 512.

Splits and leakage audit. AudioCaps contributes 45,178/2,223/4,411 train/validation/test rows and 45,178/445/883 unique YouTube identifiers, respectively. Pairwise identifier overlap among all three splits is exactly zero; ordered and sorted-unique identifier hashes are in the manifest. WavCaps is a deterministic 46,000-row, source-clean subset restricted to provenance label `WavCaps/WavCaps` (two decode failures during construction), split 40,500/500/5,000 using seeds 2026 and 4242. Before the R2 crossing, sources appearing in either evaluation gallery are removed, leaving 44,725 AudioCaps training rows, 439 AudioCaps validation sources, 40,440 WavCaps training rows, and 500 WavCaps validation rows; the two test sets share 10 normalized sources. Seeded equal matching then selects 39,940 training rows per source. Hashes and the frozen overlap counts are supplied.

Paired uncertainty. Seeds are first averaged within a condition cell. Cells are keyed by evaluation suite, training source, text anchor, intervention, method, dimension, and budget. Complete pairs are formed by dropping only the controlled arm field, yielding 8 anchor, 8 quality, 4 initialization, and 3 encoder blocks. Ten thousand percentile resamples use NumPy seed 20260728. Because these are a few fixed, correlated design cells rather than draws from a target population, every reported paired interval is labeled descriptive. The accompanying trace records every individual paired effect. The script recomputes means and standard deviations from the seed records, refits the displayed models and intervals from hashed preconstructed blocks and cells, and checks the paired intervals against the archived R8 summary. It does not derive those blocks or cells from training outputs.

Image construction. AudioCaps test rows use the dataset audio, caption, and YouTube ID plus the deterministic default thumbnail `0.jpg` for the same ID. Retrieval failures keep their row with a gray placeholder, so methods share an identical 4,411-row ordering. No multiple-frame selection, temporal localization, or visibility filtering is performed. The unique-video diagnostic retains the first test occurrence for each of 883 IDs. These rules prevent silent pool changes but do not validate semantic image–audio correspondence.

B CLIP-TEXT VERSUS CLAP-TEXT ANCHOR

For completeness, the descriptive paired CLAP-minus-CLIP percentile interval across the eight fixed method-by-dimension blocks is $[-0.006971, 0.000055]$. These blocks are fixed design choices, not a population sample; the interval is a sensitivity summary rather than a population confidence interval.

Table 2: R1 image–audio mean recall (mean±SD, five seeds). Rows are matched on data, method, dimension, and seed.

Method	Anchor	64	128	256	512
Linear probe	CLIP	$.03677 \pm .00337$	$.03583 \pm .00370$	$.03910 \pm .00377$	$.03872 \pm .00256$
	CLAP	$.02934 \pm .00246$	$.02936 \pm .00170$	$.02994 \pm .00150$	$.03283 \pm .00094$
Shared JL	CLIP	$.00473 \pm .00044$	$.01689 \pm .00047$	$.02925 \pm .00123$	$.03677 \pm .00144$
	CLAP	$.01056 \pm .00094$	$.01985 \pm .00103$	$.02720 \pm .00092$	$.02944 \pm .00077$

Shared JL favors CLAP at 64 and 128 dimensions but not at 256 or 512; the linear probe favors CLIP throughout. This interaction motivates the pooled paired summary and precludes a universal anchor ranking. A discovered evaluation-path limitation excludes R1 text-leg recalls: training uses the declared anchor-specific projected tensors, but the evaluation helper substitutes the cache's default raw CLIP text tensor for both arms. This does not enter image–audio retrieval, so the anchor comparison in this table is retained; $\overline{R}_{\mathrm{IT}}$ and $\overline{R}_{\mathrm{AT}}$ from R1 are not interpreted. The standard R2/R4/R5/R7 paths evaluate their declared cached text tensors and are unaffected.

C QUALITY AND SOURCE CONTROLS

Table 3: R2 AudioCaps quality subsets at $m = 512$ (mean±SD, five seeds; 10,681 training pairs per arm). "Top" and "bottom" refer only to matched-pair CLAP cosine similarity.

Method	Arm	$\overline{R}_{\mathrm{AT}}$	$\overline{R}_{\mathrm{IA}}$
Linear probe	top	$.17014 \pm .00279$	$.03154 \pm .00314$
	random	$.20217 \pm .00202$	$.03700 \pm .00320$
	bottom	$.17586 \pm .00371$	$.03658 \pm .00186$
Shared JL	top	$.15958 \pm .00302$	$.02314 \pm .00151$
	random	$.19355 \pm .00109$	$.03058 \pm .00145$
	bottom	$.17413 \pm .00196$	$.03294 \pm .00141$

Table 4: R2 crossed source results at $m = 512$ (mean±SD, five seeds). A dash means no image gallery exists for that evaluation suite.

Method	Train	Evaluate	$\overline{R}_{\mathrm{AT}}$	$\overline{R}_{\mathrm{IA}}$
Linear probe	AudioCaps	AudioCaps	$.22012 \pm .00306$	$.04031 \pm .00167$
	WavCaps	AudioCaps	$.10542 \pm .00313$	$.02822 \pm .00513$
	AudioCaps	WavCaps	$.03909 \pm .00147$	—
	WavCaps	WavCaps	$.09543 \pm .00496$	—
Shared JL	AudioCaps	AudioCaps	$.21506 \pm .00472$	$.03886 \pm .00198$
	WavCaps	AudioCaps	$.09447 \pm .00277$	$.02668 \pm .00336$
	AudioCaps	WavCaps	$.04047 \pm .00091$	—
	WavCaps	WavCaps	$.09729 \pm .00225$	—

The random subset is not an intermediate "quality" level: it is a size control. Its behavior further illustrates that a model-derived cosine score does not order the subsets as a human caption-quality scale would require. The score is the cosine of each matched, normalized CLAP audio/text pair.

Quartiles use the frozen implementation's `torch.argsort` ordering; no additional tie-breaking rule is imposed.

Across eight fixed method-by-dimension blocks, the bottom-minus-top descriptive paired intervals are $[0.00270, 0.00642]$ for image–audio mean recall and $[0.00827, 0.01489]$ for audio–text mean recall. These intervals summarize the fixed blocks and do not identify a causal caption-quality effect.

D LoRA-Proxy Controls

Table 5: R4 inherited versus independent (control) audio initialization for the LoRA proxy (image–audio mean recall, five seeds).

Initialization	64	128	256	512
Independent	$.02957 \pm .00207$	$.03322 \pm .00123$	$.03682 \pm .00266$	$.03949 \pm .00291$
Inherited	$.02856 \pm .00281$	$.03351 \pm .00193$	$.03741 \pm .00378$	$.03803 \pm .00292$

Only the audio projection differs immediately before Phase B; all non-audio tensors are byte-identical at the intervention point. Because downstream gradients can diverge, final tensor equality is neither expected nor required. The descriptive interval crossing zero should be read as failure to detect an initialization effect under this design, not proof of equivalence.

E Matched Speech Encoder Control

Table 6: R5 SPEECH-COCO Shared-JL results (mean±SD, three seeds). Image/text Phase A, rows, gallery, projection, dimension, and optimization are held fixed.

Encoder	Metric	128	256	512
CLAP	$\overline{R}_{\mathrm{AT}}$	$.02539 \pm .00028$	$.05643 \pm .00092$	$.09289 \pm .00257$
	$\overline{R}_{\mathrm{IA}}$	$.00211 \pm .00012$	$.00384 \pm .00015$	$.00696 \pm .00085$
Whisper-small	$\overline{R}_{\mathrm{AT}}$	$.13819 \pm .00042$	$.40458 \pm .00168$	$.68818 \pm .00982$
	$\overline{R}_{\mathrm{IA}}$	$.00886 \pm .00054$	$.04460 \pm .00215$	$.10020 \pm .00633$

F Expanded Condition-Cell Analysis

The constrained summary used in the paper is

$$\widehat{\overline{R}_{\mathrm{IA}}} = \alpha \sqrt{\overline{R}_{\mathrm{IT}} \overline{R}_{\mathrm{AT}}}, \tag{1}$$

where $\overline{R}_{\mathrm{IT}}$ is image–text mean recall, $\overline{R}_{\mathrm{AT}}$ is audio–text mean recall, and α is fit through the origin. The reduced comparison uses $\beta \sqrt{\overline{R}_{\mathrm{AT}}}$. The coefficient α is a residual description of a specified set of conditions; it is not a causal or universal measure.

Table 7: R6/R8 in-sample condition-cell models ($n = 78$ after excluding 16 R1 cells with invalid text legs). Log fits use $\log(\max(R, 10^{-6}))$ during fitting and are evaluated in recall space.

Model	Pearson r	r^2	MAE
Constrained $\alpha \sqrt{\overline{R}_{\mathrm{IT}} \overline{R}_{\mathrm{AT}}}$	.730	.510	.01441
Constrained $\beta \sqrt{\overline{R}_{\mathrm{AT}}}$	.373	.131	.02361
Free $c + a \log \overline{R}_{\mathrm{IT}} + b \log \overline{R}_{\mathrm{AT}}$	.529	.088	.01654
Reduced $c + b \log \overline{R}_{\mathrm{AT}}$	.317	$-.021$	.02008

The free full model estimates $a = 0.320$ and $b = 1.002$, so it does not recover equal exponents. The constrained geometric form is retained as a low-variance summary, not because the exponent symmetry is empirically identified. The 78 retained cells include quality, source, initialization, encoder, method, dimension, and suite changes. This breadth improves stress testing but introduces protocol heterogeneity; all global fits should be read accordingly.

G Evidence and Excluded Analyses

The submitted supplementary bundle contains a machine-readable evidence manifest and compact repair ledger. To avoid repository linkage during double-blind review, the public manifest withholds the version-control commit and uses a neutral source-snapshot label. It retains non-VCS source-snapshot and resolved-configuration hashes for every run unit, plus SHA-256 for all 169 run-unit results, analyses, final machine validation, and the scale reference; seeds; suites; and gallery identities. The repair ledger hash-binds five original repair receipts and the final validation to the preregistered batch manifest. It verifies that all 169 planned unit identifiers occur exactly once in the final inventory and that the expected R1/R2/R4/R5/R7 unit, record, aggregate, and seed grids are complete. Thus recovery inclusion is exhaustive over the planned grid rather than selected by outcome.

The original governed-batch receipt marks 13 units administratively promotable. The remaining 156 are explicitly retained as *audited recovery runs*: they followed documented harness, split-binding, gallery-identity, or CPU-analysis repairs and therefore cannot inherit the original operations receipt. We do not relabel that flag. Instead, compact evidence eligibility requires a non-smoke succeeded result whose canonical result hash recomputes from its content, recorded 64-character source-snapshot and governed or recovery resolved-configuration identifiers, and inclusion in the passed final 169-unit validation. All 169 meet that compact rule. The bundle does not resolve the snapshot/configuration identifiers against the underlying bytes. The ledger supports protocol consistency only through the scientific invariants declared in the bound receipts; it does not re-execute training. The unit trace links all 538 seed records and the preconstructed paired blocks and condition cells to paths and hashes.

One input family and two analyses are deliberately excluded. The condition-cell models reject all 16 R1 anchor-control cells because the known evaluation-path substitution invalidates their text legs; the evidence trace records every included and excluded cell. Separately, the 50-row thumbnail visibility sheet has no authorized human labels and remains `incomplete_pending_human`, while the calibration-budget analysis contains no data at budgets 10–100 and lacks complete equal-seed replicates at 250–2,000. Neither incomplete analysis can support a paper claim. The refreshed absolute-scale collector also attached zero compatible reference rows, so the ImageBind row is explicitly retained as a separately hashed older same-gallery scale reference rather than presented as part of the refreshed matched suite.