Additional file 1
Statistical robustness and the companion study

This file accompanies the manuscript *Token allocation, not resampling, sets the square-canvas penalty in a small from-scratch audio transformer*. It contains the statistical checks the manuscript refers to and the companion study that separates the two corpus statistics the main text leaves confounded. Reference numbers [1] are those of the manuscript's reference list.

1 Statistical robustness

All quantities below are recomputed from the per-cell results reported in the manuscript; it reports the checks the manuscript cites it for.

Where the nuisance variation sits. Pairing is used throughout, so it matters how much of the dispersion of a paired difference is between seeds and how much is within a seed. Using the three per-seed means of each arm, the between-seed share of the variance of the difference is 0.43 for the 96.9% versus 88.7% contrast, 0.12 for the 66.0% versus 88.7% contrast, 0.61 for the time-only contrast, 0.23 for the frequency-only contrast and 0.28 for the square-canvas contrast. The fold-level term dominates for the contrasts that carry a claim, which is why the fold-paired units are used rather than seed means.

The budget axis, including the arm excluded by the inclusion rule. The sweep's inclusion rule fixes the budget near 500 tokens and therefore excludes one measured configuration, the native canvas with a 16×16 patch: an 8×31 grid of 248 tokens at a 79.5% time share, which reaches 72.30% over three seeds and five folds. The same arm sits about 4 pp below the linear interpolation of the sweep at its time share, so it is not part of the dose–response curve, and the exclusion rule (fixed before the results were produced) applies to it.

Resolution of the two cross-corpus measurements. With five paired folds and a per-fold standard deviation of 1.30 pp, the ESC-50 penalty of -10.20 pp is resolved to 2.16 pp at 80% power. With ten paired folds and a standard deviation of 3.47 pp, the UrbanSound8K penalty of -0.38 pp is resolved to 3.45 pp. The second measurement therefore excludes a penalty larger than about 3.5 pp on that corpus while remaining consistent with a smaller one, and the dispersion of that penalty is itself poorly determined at ten folds, which is why the second corpus is reported as a boundary condition and not as a second estimate of the same quantity.

Permutation tests. A sign-flip permutation test is assumption-light but resolution-limited: with five paired folds the smallest attainable p is 0.0625 and with three seeds it is 0.25, so for those contrasts the t interval is the reported one. The companion study below uses a design with more paired units and reports the permutation test alongside every contrast.

2 Companion study: separating the corpus statistics

The main text reports that the allocation penalty is present on ESC-50 and bounded on UrbanSound8K, and states plainly that two candidate explanations remain confounded: the corpus on which the penalty is measured has five times as many classes and a quarter of the training pool. Both quantities can be manipulated on a single corpus, holding the signal, the front end, the model and the evaluation rule fixed. This study does that, and then measures the mechanism that the main text would otherwise only assert.

2.1 Class information in each representation

Table 1: Class information carried by different representations of the same clips. A linear probe is fitted on four official folds of ESC-50 and evaluated on the fifth. Values are from the companion probe, whose classifier and mel pipeline differ from those of the probes quoted in the main text; the two agree on the ordering of the representations.

Representation	Dimensions	Probe accuracy (%)
Full spectrogram	64 128	25.95
Frequency-averaged (temporal content only)	501	10.45
Time-averaged (spectral content only)	128	33.25
Token means of the 4×63 lattice	252	21.50
+ dispersion along time	+252	25.65
+ dispersion along frequency	+252	27.45
Token means of the 16×32 lattice	512	28.70
+ dispersion along time	+512	35.40
+ dispersion along frequency	+512	39.45

2.2 Design

The companion study is a local reconstruction of the same design, and reaches a comparable parameter count with a deeper trunk than the manuscript platform: a hierarchical axial transformer of 10.80 M parameters (stage widths 128/256/384, depths 3/8/2, four, eight and twelve heads), the same front end, and the same recipe (300-epoch schedule shortened to 100 epochs per run). One model is trained per (arm, fold) and evaluated once on the held-out fold, which is the manuscript's protocol. Pool reduction is a nested class-stratified subsample, so the smaller pool is a subset of the larger one and the evaluation set never changes. Two requirements are imposed on every contrast reported here, and they are checked before the arms are run rather than after. The two arms of a contrast must hold token budgets within 5% of each other, and their patch shapes must be reciprocal, so that the manipulated quantity is the split of the budget between the axes rather than the size of the budget or the sampling density of one axis. A pair that satisfies neither is reported as a pilot and excluded from the results. The pairs that satisfy both requirements, all on ESC-50, are:

- **P1, at the manuscript's budget.** Time-heavy $16 \times 8 \to 8 \times 63 = 504$ tokens ($\rho = 0.887$) against frequency-heavy $8 \times 16 \to 16 \times 32 = 512$ tokens ($\rho = 0.667$); budgets differ by 1.6%.
- **P3, at half that budget.** Time-heavy $32 \times 8 \to 4 \times 63 = 252$ tokens ($\rho = 0.940$) against frequency-heavy $16 \times 16 \to 8 \times 32 = 256$ tokens ($\rho = 0.800$); budgets differ by 1.6%.

P1 and P3 are also run with the class count reduced from fifty to ten, holding the pool, the folds and everything else fixed, to vary the granularity of the task rather than of the budget. The pilot that preceded these pairs compared 32×8 (252 tokens) with 8×16 (512 tokens). Its arms differ in budget

by a factor of two and its patch shapes are not reciprocal, so its difference measures the budget rather than the allocation and it is not used as evidence anywhere in this file.

2.3 Results

Table 2: The allocation contrast with the token budget held fixed inside each pair. The difference is the frequency-heavy arm minus the time-heavy arm, so a negative value reproduces the direction of the main sweep.

Pair	Tokens	ρ	Folds	Time (%)	Freq. (%)	Δ (pp)
P1 504 tokens	504 / 512	0.887 / 0.667	3	33.08	30.33	$-2.75\,\mathrm{pp}\ [-6.04, +0.54]$
P3 252 tokens	252 / 256	0.940 / 0.800	3	26.17	24.33	$-1.83\,\mathrm{pp}\ [-10.56, +6.89]$

For pair P1, whose arms hold 504 and 512 tokens, the frequency-heavy arm differs from the time-heavy arm by $-2.75\,\mathrm{pp}\ [-6.04, +0.54]$ over 3 paired folds (the interval includes zero; exact sign-flip $p = 0.250$ against a minimum attainable p of 0.250 at 3 folds, so the interval is the reported statistic; MDE 4.10 pp at 80% power). For pair P3, whose arms hold 252 and 256 tokens, the frequency-heavy arm differs from the time-heavy arm by $-1.83\,\mathrm{pp}\ [-10.56, +6.89]$ over 3 paired folds (the interval includes zero; exact sign-flip $p = 0.750$ against a minimum attainable p of 0.250 at 3 folds, so the interval is the reported statistic; MDE 10.88 pp at 80% power). The two pairs hold the budget fixed within each pair and differ from each other by a factor of two, so the comparison across them separates the effect of the allocation from the effect of the budget. The pilot reported below satisfies neither requirement.

The pilot, recorded but not used. Before the pairs were run, the same contrast was measured with 32×8 (252 tokens, $\rho = 0.940$) against 8×16 (512 tokens, $\rho = 0.667$) over three folds: $+4.08\,\mathrm{pp}$ $[-0.14, +8.31]$, with per-fold differences +3.50, +6.00, +2.75. The two arms differ in budget by a factor of two, so that number measures the budget and not the allocation; it is recorded here because it is the reason the pairs above hold the budget fixed, and it is not used as evidence anywhere.

2.4 Deviations from the main study

The companion study differs from the main study in four stated ways: the local rebuild reaches 10.80 M parameters with stage depths 3/8/2 rather than the manuscript's reported depths; both allocation arms run at a smaller matched budget (252 and 512 tokens) because a single 4 GB GPU has to complete the whole grid; each run uses 100 epochs rather than 300; and three of the five official ESC-50 folds are used, with UrbanSound8K folds reconstructed per source recording because the archive available here does not ship the official metadata file. Effect sizes and their signs, not absolute accuracies, are what the companion study carries across.

References

[1] Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, Uszkoreit J, Houlsby N, An image is worth 16x16 words: Transformers for image recognition at scale, in Proc. Int. Conf. Learn. Represent. (ICLR 2021). openreview.net/forum?id=YicbFdNTTy. arXiv:2010.11929.

[2] Gong Y, Chung Y-A, Glass J, AST: Audio spectrogram transformer, in Proc. Interspeech 2021 (ISCA, 2021), pp. 571–575.

[3] Gong Y, Lai C-I, Chung Y-A, Glass J, SSAST: Self-supervised audio spectrogram transformer, Proc. AAAI Conf. Artif. Intell. 36, 10699–10709 (2022).

[4] Koutini K, Schlüter J, Eghbal-zadeh H, Widmer G, Efficient training of audio transformers with patchout, in Proc. Interspeech 2022 (ISCA, 2022), pp. 2753–2757.

[5] Berg A, O'Connor M, Cruz MT, Keyword transformer: A self-attention model for keyword spotting, in Proc. Interspeech 2021 (ISCA, 2021), pp. 4249–4253.

[6] Ristea NC, Ionescu RT, Khan FS, SepTr: Separable transformer for audio spectrogram processing, in Proc. Interspeech 2022 (ISCA, 2022), pp. 4103–4107.
[7] Ghosh S, Seth A, Umesh S, Manocha D, MAST: Multiscale audio spectrogram transformers, in Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP 2023) (IEEE, 2023), pp. 1–5.
[8] Feng J, Erol MH, Chung JS, Senocak A, ElasticAST: An audio spectrogram transformer for all length and resolutions, in Proc. Interspeech 2024 (ISCA, 2024), pp. 4743–4747.
[9] Feng J, Erol MH, Chung JS, Senocak A, FlexiAST: Flexibility is what AST needs, in Proc. Interspeech 2023 (ISCA, 2023), pp. 2828–2832. arXiv:2307.09286.
[10] Makineni A, Geng B, Tian Q, Full-frequency temporal patching and structured masking for enhanced audio classification, arXiv:2508.21243 [cs.SD] (2025).
[11] Feng J, Erol MH, Chung JS, Senocak A, From coarse to fine: Efficient training for audio spectrogram transformers, in Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP 2024) (IEEE, 2024), pp. 1416–1420.
[12] Behera SR, Dhiman A, Gowda K, Narayani AS, FastAST: Accelerating audio spectrogram transformer via token merging and cross-model knowledge distillation, in Proc. Interspeech 2024 (ISCA, 2024), pp. 4733–4737.
[13] Lee T, Lee H, Token pruning in audio transformers: Optimizing performance and decoding patch importance, in Frontiers in Artificial Intelligence and Applications, ECAI 2025 (IOS Press, 2025). arXiv:2504.01690.
[14] Xiao T, Singh M, Mintun E, Darrell T, Dollár P, Girshick R, Early convolutions help transformers see better, Adv. Neural Inf. Process. Syst. 34 (2021). arXiv:2106.14881.
[15] Ho J, Kalchbrenner N, Weissenborn D, Salimans T, Axial attention in multidimensional transformers, arXiv:1912.12180 [cs.CV] (2019).
[16] Rao Y, Zhao W, Zhu Z, Lu J, Zhou J, Global filter networks for image classification, Adv. Neural Inf. Process. Syst. 34 (2021). arXiv:2107.00645.
[17] Huang Z, Zhang Z, Lan C, Zha Z-J, Lu Y, Guo B, Adaptive frequency filters as efficient global token mixers, in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV 2023) (IEEE, 2023), pp. 6026–6036.
[18] Defferrard M, Bresson X, Vandergheynst P, Convolutional neural networks on graphs with fast localized spectral filtering, Adv. Neural Inf. Process. Syst. 29 (2016). arXiv:1606.09375.
[19] Touvron H, Cord M, Sablayrolles A, Synnaeve G, Jégou H, Going deeper with image transformers, in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV 2021) (IEEE, 2021), pp. 32–42.
[20] Piczak KJ, ESC: Dataset for environmental sound classification, in Proc. 23rd ACM Int. Conf. Multimedia (ACM, 2015), pp. 1015–1018.
[21] Salamon J, Jacoby C, Bello JP, A dataset and taxonomy for urban sound research, in Proc. 22nd ACM Int. Conf. Multimedia (ACM, 2014), pp. 1041–1044.
[22] Lamothe C, Obliger M, Best P, Trapeau R, Ravel S, Artières T, Marxer R, Belin P, A large annotated dataset of the common marmoset vocal repertoire, Sci. Data 12, 782 (2025).
[23] Lamothe C, Obliger M, Best P, Trapeau R, Ravel S, Artières T, Marxer R, Belin P, MarmAudio: A large annotated dataset of the common marmoset vocal repertoire [data set], Zenodo (2025).
[24] Warden P, Speech commands: A dataset for limited-vocabulary speech recognition, arXiv:1804.03209 [cs.CL] (2018).