Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Token allocation, not resampling, sets the square-canvas penalty in a small from-scratch audio transformer

Weiwei Xiao

xiaoww@ncut.edu.cn

North China University of Technology

Jianmin Liu

North China University of Technology

Research Article

Keywords: audio spectrogram transformer, token allocation, input geometry, from-scratch training, empirical comparison, environmental sound classification.

Posted Date: November 24th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11376285/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Token allocation, not resampling, sets the square-canvas penalty in a small from-scratch audio transformer

Weiwei Xiao[1,*]

Jianmin Liu[1,*]

[1]College of Science, North China University of Technology, Beijing, China

[*]Corresponding authors: xiaoww@ncut.edu.cn, ljm29318636@gmail.com

Abstract

Audio spectrogram transformers are usually applied by resizing the spectrogram onto a square canvas, which changes the signal and the time–frequency layout of the tokens at the same time. This study separates the two for one fixed platform: a 10.84M-parameter hierarchical axial transformer trained from scratch, with the front end, optimizer, augmentation and evaluation rule held constant while the share of roughly 500 tokens placed on the time axis is swept from 9.9% to 99.8% over six configurations, three of them replicated over three seeds. Accuracy rises with the time share and then saturates: the replicated pair of time-heavy configurations is statistically indistinguishable at a statistical resolution of 1.64 percentage points, and the single-seed extreme of the plateau is indistinguishable from them as well, while the frequency-heavy end loses up to 14.95 percentage points. The square-canvas penalty is therefore an allocation effect rather than an interpolation effect. Resampling either axis alone at a matched allocation, 87.5% and 90.0% time share against the reference's 88.7%, costs nothing detectable (1.18 and 0.08 percentage points, both intervals containing zero), whereas reaching the same 50% allocation on the unresampled spectrogram reproduces most of the penalty (6.05 percentage points) and lands within 0.06 percentage points of the curve's interpolation. The penalty is nearly constant across folds on ESC-50 and carries no penalty larger than 3.5 percentage points on a ten-class urban corpus, and holding the budget equal within each pair reproduces its direction. Linear probes on the same data explain the asymmetry:

class information is carried by the spectral signature and the accuracy by how that signature evolves, so coarsening the lattice loses detail inside tokens — larger on the mel axis, and tolerated because the spectral summary survives pooling — and temporal positions, which exist only on the time axis and which no resolution inside a token restores.

Keywords: audio spectrogram transformer, token allocation, input geometry, from-scratch training, empirical comparison, environmental sound classification.

1 Introduction

Vision transformers build their representations from a sequence of tokens[1], and audio models construct those tokens from spectrograms[2]. In the common workflow, a spectrogram is resized onto a square canvas so that a backbone designed for images can be reused unchanged. That workflow bundles two decisions that are rarely separated. The first is what happens to the signal: a 128×501 spectrogram resized to 224×224 is interpolated, and the time axis is compressed by more than a factor of two. The second is what happens to the tokens: at a fixed patch size, the same resize moves most of the sequence off the time axis, so the model attends across a much coarser temporal grid. A pretrained model is largely insulated from this distinction: its weights were fitted to whatever layout the pretraining corpus used, and the final accuracy mixes the layout, the corpus and the fine-tuning budget. In the from-scratch setting on a small target dataset the target dataset supplies the task and all of the training information, so every configuration choice must be justified within that budget. This is the setting studied here. Recent work has already established the direction of the effect. The Audio Spectrogram Transformer (AST) reports that, at a matched patch count and without pretraining, a rectangle that keeps the full frequency axis and segments time outperforms a square patch of the same area (0.154 against 0.143 mean average precision (mAP) on balanced AudioSet, Table 6 of [2]). The Self-Supervised Audio Spectrogram Transformer (SSAST)[3] reports, also for from-scratch training on ESC-50, that a frame-based model keeping the full frequency axis inside each token (128×2 patches) reaches 53.7% against 41.9% for the 16×16 square-patch model at equal patch area, a gap of 11.8 percentage points (Table 4 of [3]). Neither number settles whether time-heavy tokenization is preferable; both fix an accuracy at one allocation on one corpus. If the benefit is a property of the allocation axis rather

than of one preferred layout, then it should have a shape: a region where the allocation no longer matters, a boundary beyond which it degrades, and a measurable relationship between the two. That shape, and the attribution of the effect to allocation rather than to interpolation, are what this study supplies, together with the question those comparisons also leave open: what determines where the effect is present and where it disappears. The comparison is organized around a single fixed platform: a 10.84M-parameter hierarchical axial vision transformer (ViT) trained from scratch, with the token budget, front end, optimizer, augmentation and evaluation rule held constant while the share of tokens on the time axis varies. Six configurations span 9.9% to 99.8%, and the three that carry the reported claims are replicated over three seeds and five predefined folds. Three further experiments attribute the effect: an axis decomposition that resamples one axis at a time at a matched allocation, a control that reaches the same 50% allocation by re-partitioning the native unresampled spectrogram, and a schedule arm that extends the training budget from 40 to 300 epochs. A companion study (Additional file 1) then addresses the two corpus statistics that the cross-corpus comparison leaves confounded, excluding the training pool by a property of the signal and manipulating the number of classes directly. Four tasks in total are used, and they are chosen to span the range the mechanism predicts: fifty-class ESC-50, ten-class UrbanSound8K, six-class marmoset call classification and gibbon call detection, the last two from a different domain and both near ceiling for the fixed recipe. The mechanism that the measurements support is stated in Section 7.2. The comparison yields five observations, four about the allocation axis and one about the training budget.

1. A dose–response characterization of token allocation at a fixed budget. Six configurations of a fixed platform map the share of tokens on the time axis from 9.9% to 99.8%. Accuracy rises with the time share, saturates over the plateau arms, and collapses toward the frequency-heavy end over a range of 14.95 percentage points, with fold-paired and seed-paired intervals throughout.
2. An attribution of the effect to allocation rather than to resampling. Resampling one axis at a time at a matched allocation leaves no detectable penalty, whereas reaching the 50% allocation on the unresampled native spectrogram reproduces most of the penalty measured for the square canvas. The prediction taken from the allocation curve agrees with that measurement to 0.06 percentage points.

3. The scope of that effect. The penalty is no larger than 3.5 percentage points on a ten-class urban corpus, the class count is manipulated directly on a single corpus, the training pool is set aside because the quantity that predicts the penalty is a property of the recording rather than of the training set, and the mechanism's prediction is tested on near-ceiling call-classification tasks from a different domain, where the manipulation is inert (Sections 6 and 6.1).
4. A measured account of why the axes behave differently. Linear probes separate what each axis carries, what the lattice discards along each axis, and what pooling leaves recoverable, and they identify the binding quantity as the number of temporal positions the lattice can relate (Section 7.2).
5. A second design axis, the training budget, varied independently of the allocation, which shows that representation and optimization contribute independent gains (the budget observation).

The comparison holds the architecture fixed and varies only the input-side configuration, so it measures which spectrogram decisions matter and by how much when no pretrained weights are available. How that configuration compares with a convolutional or hybrid architecture trained on the same data is a separate question.

2 Related work

2.1 Audio transformers and the spectrogram as an image

The Audio Spectrogram Transformer applies a vision transformer to log-mel spectrograms and established the practice of reusing image backbones for audio[2]; self-supervised pretraining[3] and parameter-efficient supervised training on AudioSet[4] followed, together with keyword spotting without pretraining[5], along with separable axial processing for the two-dimensional structure of spectrograms[6]. Because the reported gains of these systems combine architecture, corpus and schedule, the from-scratch regime isolates the configuration component.

2.2 Token layout, resolution, and allocation

Multiscale audio spectrogram transformers pool tokens at several granularities, changing the effective resolution seen by each stage[7]; ElasticAST varies input length and resolution at inference time[8]; FlexiAST obtains patch-size flexibility by training with randomly selected patch sizes and resized patch

and positional embeddings[9]; full-frequency temporal patching takes the opposite extreme, keeping the entire frequency axis inside each token and scanning the number of temporal patches[10]; and multi-phase schedules[11] and token merging or pruning after the patch embedding[12, 13] vary the same quantities from the training side. Convolutional stems before attention stabilize image transformers[14] and axial attention factorizes dense two-dimensional attention[15]; frequency-domain global mixing has been explored in vision[16, 17], with graph spectral filtering as one operator family used there[18]. The quantities varied in this literature are patch size, patch count and inference-time resolution; the share of a fixed budget assigned to each axis, swept across the whole range and reported with a decomposition, has not been systematically mapped.

2.3 Prior evidence and what was missing

The direction of the effect is established, as the previous subsection records, but three things are absent, each a precondition for using the effect as a design rule: the relation has been sampled only at endpoints or along one direction, so the plateau and the degradation boundary are unlocated; the comparisons change the input and the token layout together, so interpolation has not been separated from allocation; and nothing indicates the conditions under which the relation holds at all. The experiments below address these three questions in order.

3 Platform, protocol, and parameterization

3.1 Platform

All configurations use the same hierarchical axial ViT of 10.84M parameters. The trunk has three stages with embedding widths $[128, 256, 384]$ and depths $[2, 4, 2]$, four, eight and twelve attention heads, an MLP ratio of 4.0, and axial attention factorized over the frequency and time axes[15]. The input stem is a three-layer convolutional patch embedder with asymmetric kernels that favor temporal resolution; convolutional stems of this kind stabilize the optimization of vision transformers and raise their peak accuracy at matched compute[14], and the stem's architecture is identical across every arm of the allocation sweep, its stride following the patch shape so that the measured grids of Table 1 are produced. Layer scale[19] is initialized at 0.1, drop path at 0.2, attention and embedding dropout at 0.1. The front

end is the same everywhere: 16 kHz audio, a 400-sample window with a 160-sample hop, 128 mel channels to 8 kHz, amplitude-to-decibel conversion with an 80 dB floor, and per-clip standardization. Only the canvas, the patch shape and the schedule vary between arms.

3.2 Evaluation protocol

Experiments on ESC-50 use the five predefined folds[20]: each run trains on four and is evaluated once on the held-out fold, so no test information can enter model selection. Unless stated otherwise, runs use 300 epochs of AdamW at a peak learning rate of 5×10^{-4}, batch size 32, mixup with $\alpha = 0.4$, and the final epoch rather than the best-validated one, which keeps selection independent of any validation split. Arms carrying a claim are repeated over seeds 0, 1 and 2; fold-paired contrasts compare configurations on the same folds and seed-paired contrasts on the same (seed, fold) cells.

3.3 Parameterizing the allocation

Let F and T denote the frequency and time extents of the native spectrogram, here $F = 128$ mel channels and $T = 501$ frames. A patch of shape (p_f, p_t) partitions the canvas into a grid of $N_f = \lfloor F/p_f \rfloor$ frequency tokens and $N_t = \lfloor T/p_t \rfloor$ time tokens, giving a budget of $N = N_f N_t$ tokens. Two quantities describe the arm. The first is the token budget N. The second is the share of that budget placed on the time axis,

$$\rho = \frac{N_t}{N_f + N_t} = \frac{1}{1 + (F/T)\, r}, \qquad r = \frac{p_t}{p_f}. \tag{1}$$

Holding N fixed imposes a hyperbola on the patch shape, $p_f p_t = FT/N$, so r and therefore ρ are determined once the budget and the patch area are fixed. Sweeping ρ at a constant N is thus equivalent to sweeping the ratio of temporal to frequency patch extent, and the axis of Figure 2 is unambiguously the allocation rather than the patch size (Figure 1). Two consequences of the parameterization matter for interpreting the results. At $\rho \to 1$ the frequency axis is collapsed into a single token, which is the configuration used by full-frequency temporal patching[10]; at small ρ the time axis is reduced to a handful of tokens, so a clip is summarized by a few temporal positions.

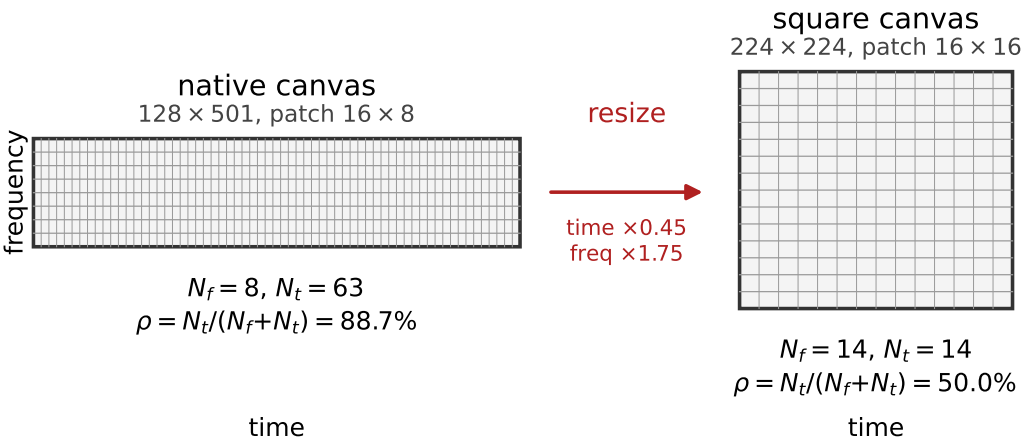


Figure 1: The allocation parameterization. A patch shape partitions the canvas into a token lattice, and ρ is the share of that lattice placed on the time axis. The square canvas on the right is the configuration compared throughout against the native reference.

3.4 Grids are reported as measured

The grid a convolutional stem produces is not always $\lfloor F/p_f \rfloor \times \lfloor T/p_t \rfloor$, because padding can add a token at an axis end; every configuration below is therefore reported with the grid the stem prints at run time, and the budget is computed from that measured grid. Table 1 lists the six arms with their measured grids, budgets and time shares.

Table 1: The six configurations of the allocation sweep. Grids are the measured stem outputs; the budget and time share are derived from them.

Patch (p_f, p_t)	Measured grid	Tokens N	Time tokens	Time share ρ	Seeds
2×64	64×7	448	7	9.9%	1
4×32	32×15	480	15	31.9%	1
8×16	16×31	496	31	66.0%	3
16×8	8×63	504	63	88.7%	3
32×4	4×126	504	126	96.9%	3
128×1	1×501	501	501	99.8%	1

3.5 Controls

The reading of the sweep is protected by two controls rather than by an external baseline: the budget, whose variation inside the sweep is at most 11% and does not drive the effects reported below, and the schedule, treated in Section 5.3.

3.6 Statistical treatment

Each reported contrast is a paired difference over shared experimental units: fold-paired contrasts use the five folds with a t interval at four degrees of freedom, and seed-paired contrasts the fifteen (seed, fold) cells. Where a quantity is a location on the allocation axis rather than a difference of means, a bootstrap over folds records which configuration attains the maximum accuracy, which reports the stability of the optimum without assuming a functional form. Intervals are reported for every contrast, and the dispersion of an effect separately from its mean.

4 Token allocation at a fixed budget

The sweep holds the platform, the front end and the schedule fixed and varies only the split of the token budget between the two axes. Accuracy increases with the time share across the whole frequency-heavy half of the range and then stops changing (Figure 2, Table 2). Between a 9.9% and a 31.9% time share the loss is $4.75\,\mathrm{pp}\ [-5.91, -3.59]$; between 31.9% and 66.0% it is a further $7.70\,\mathrm{pp}\ [-9.15, -6.25]$; and between 66.0% and 88.7% it is $2.50\,\mathrm{pp}\ [-4.13, -0.87]$. Every one of those intervals excludes zero. Beyond 88.7%, the curve is flat: the step to a 96.9% time share is $+0.20\,\mathrm{pp}\ [-1.77, +2.17]$ and the step from there to 99.8% is $+0.25\,\mathrm{pp}\ [-1.86, +2.36]$, so neither contrast is distinguishable from zero. The full range from the best to the worst configuration is $14.95\,\mathrm{pp}$. Replicating the 88.7%, 66.0% and 96.9% configurations over three seeds gives fifteen paired (seed, fold) cells per arm. The 96.9% arm is indistinguishable from the reference at $+0.10\,\mathrm{pp}\ [-1.07, +1.27]$ and is the more stable of the two across seeds (per-seed means 76.25%, 76.90%, 76.60% against 76.45%, 78.00%, 75.00%), while the 66.0% arm lies below the reference by $2.83\,\mathrm{pp}\ [-3.90, -1.76]$ with all three per-seed differences negative. A bootstrap over folds selects the reference arm in 56.4% of resamples, the 99.8% arm in 23.3% and the 96.9% arm in 20.3%, and never a configuration below the plateau. The plateau arms 88.7% and 96.9% therefore share the optimum, while the 66.0% arm sits below the reference with an interval that excludes zero and the frequency-heavy half degrades monotonically and significantly. Two controls guard the reading of the curve. The first concerns budget rather than allocation: the arms of Table 2 span 448 to 504 tokens, an 11% spread, which is too small to account for a range of $14.95\,\mathrm{pp}$; it does not follow that the budget is unimportant in general. The second concerns the

extreme: the 99.8% configuration collapses the frequency axis into a single token, the layout used by full-frequency temporal patching[10]. It is neither better nor worse than the reference, its interval centred within 0.45 pp of the reference mean, so the configuration reported as best on AudioSet-18k and Speech Commands v2, where from-scratch training under that layout raises AST from 11.25 to 18.32 mAP and from 85.27% to 95.94% accuracy, lies on the plateau of this curve rather than at a distinguishable optimum of its own.

Table 2: Allocation sweep on ESC-50. Intervals are fold-paired t intervals at seed 0, computed from the seed-0 per-fold standard deviation shown in the SD column.

Time share	Tokens	Accuracy (%)	SD	95% CI	Seeds	loss vs. previous
9.9%	448	61.50	2.03	[58.98, 64.02]	1	—
31.9%	480	66.25	2.39	[63.28, 69.22]	1	−4.75 [−5.91, −3.59]
66.0%	496	73.95	3.07	[70.14, 77.76]	3	−7.70 [−9.15, −6.25]
88.7%	504	76.45	3.18	[72.50, 80.40]	3	−2.50 [−4.13, −0.87]
96.9%	504	76.25	2.34	[73.35, 79.15]	3	+0.20 [−1.77, +2.17]
99.8%	501	76.00	3.01	[72.27, 79.73]	1	+0.25 [−1.86, +2.36]
99.8% vs 88.7%	—	—	—	—	—	−0.45 [−3.38, +2.48]

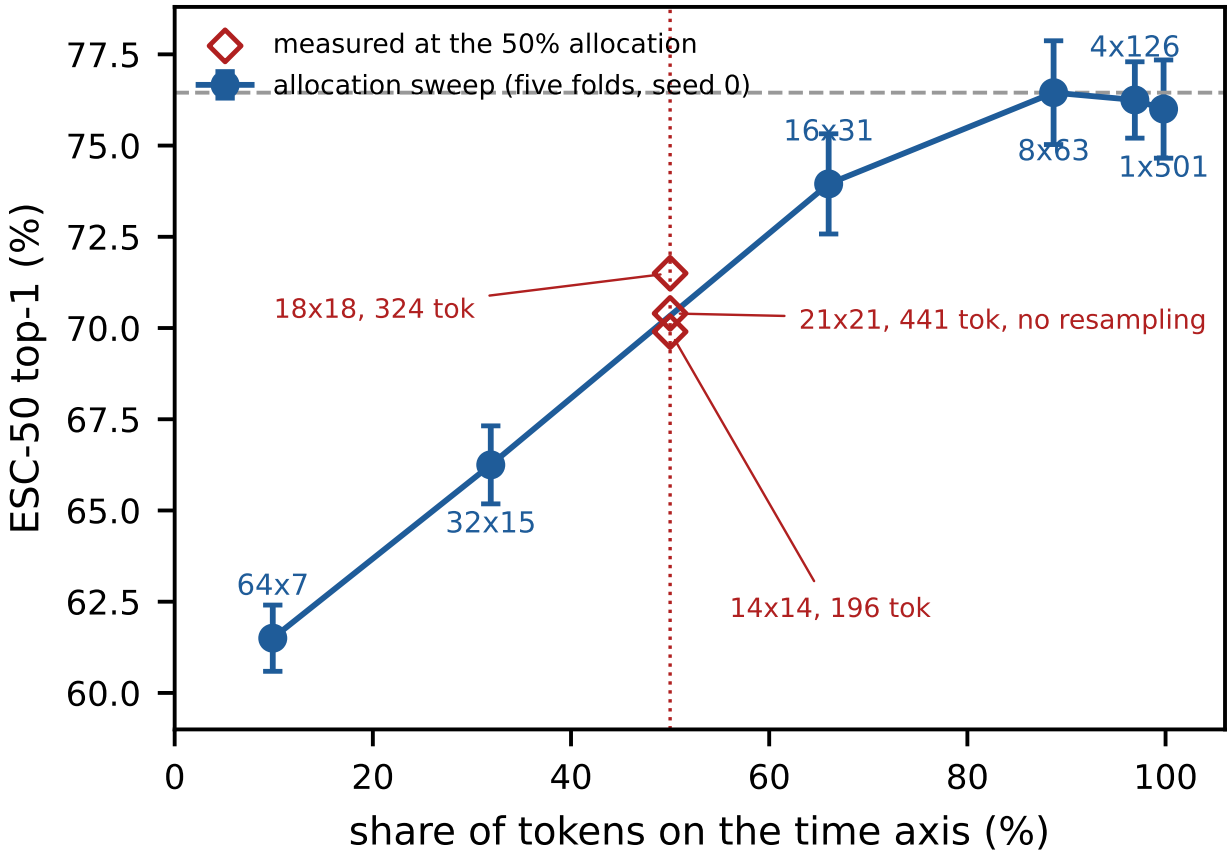


Figure 2: Accuracy against the share of tokens placed on the time axis. The budget is fixed at roughly 500 tokens; filled circles are the six sweep configurations, and the open diamonds are three independent configurations measured at the 50% allocation, none of which participates in the curve.

5 Where the penalty comes from

The sweep establishes that the allocation matters and that the extreme allocation used by the square canvas is suboptimal. Moving from a native spectrogram onto a square canvas changes two things at once: the signal is interpolated on both axes, and the token grid moves from an 88.7% to a 50% time share. The experiments in this section separate them.

5.1 Resampling one axis at a time

Two further configurations resample exactly one axis while the front end, the schedule and the evaluation rule stay fixed: the first reproduces the temporal compression of the square canvas without touching the frequency axis (128×224), the second stretches the frequency axis to 224 rows with the 501 temporal frames intact. Both use a patch shape giving a token budget comparable to the reference (448 and 441 tokens) and a per-token temporal window close to it, so the comparison changes the resampling rather than the attention grid. Table 3 reports the measured grids, the resampling factors, and the receptive field of one token in the coordinates of the original spectrogram, which makes the matching explicit: the time-only arm covers 8.95 original frames per token against 8.0 for the reference, and the frequency-only arm exactly 8.0. The outcome, over three seeds and five folds per condition, is that neither axis is responsible for the penalty on its own. Compressing the time axis by a factor of 2.24 changes accuracy by $-1.18\,\mathrm{pp}$ $[-2.58, +0.22]$, an interval that includes zero, and the effect is not consistent across seeds (-2.30, -2.35 and $+1.10\,\mathrm{pp}$). Stretching the frequency axis by a factor of 1.75 costs $0.08\,\mathrm{pp}$ $[-0.82, +0.65]$, so interpolating the mel axis of a spectrogram onto a taller canvas is, on its own, close to free. The square canvas, which does both and additionally re-partitions the tokens, costs $7.20\,\mathrm{pp}$ over three seeds (the seed-0 value used elsewhere is $6.55\,\mathrm{pp}$) $[-8.70, -5.70]$. The two single-axis penalties sum to $1.26\,\mathrm{pp}$, roughly one sixth of the square-canvas penalty. Figure 3(a) shows the four conditions together.

Table 3: Axis decomposition. The effective window is the receptive field of one token expressed in original spectrogram coordinates, obtained from the resampling factors rather than assumed.

Condition	Canvas	Patch	Grid	T'/T	F'/F	Eff. time window	Eff. freq. window
Native reference	128×501	16×8	8×63	1.000	1.000	8.00 frames	16.00 bins
Time only	128×224	16×4	8×56	0.447	1.000	8.95 frames	16.00 bins
Frequency only	224×501	32×8	7×63	1.000	1.750	8.00 frames	18.29 bins
Square canvas	224×224	16×16	14×14	0.447	1.750	35.79 frames	9.14 bins

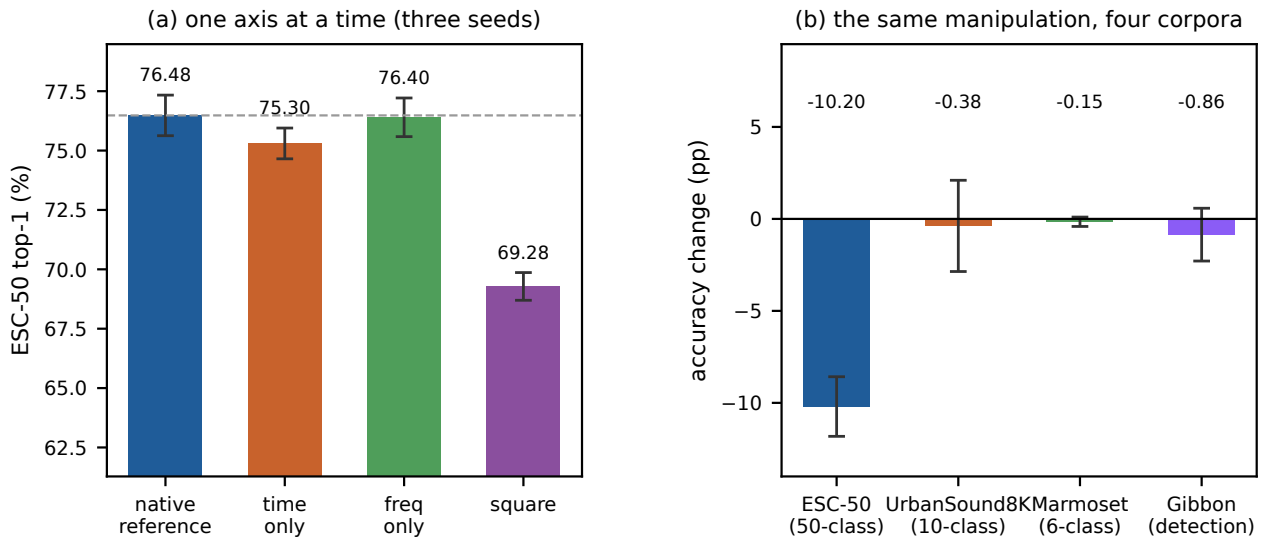


Figure 3: (a) The four conditions of the axis decomposition. (b) The allocation penalty of the same platform on the four corpora and tasks, each with its paired interval, against the extent to which the square canvas moves the time axis.

5.2 The 50% allocation, measured

The axis decomposition leaves the allocation of the square canvas explained by the curve rather than measured on it. The curve interpolates to 70.34% at a 50% time share, between the 31.9% and 66.0% measurements, and the square canvas, whose 14×14 grid places exactly half of its tokens on the time axis, is measured at 69.90%. Three configurations close that gap by measurement: the square canvas at 196 tokens; the same canvas with a 12×12 patch (18×18 grid, 324 tokens); and a 6×24 patch on the native spectrogram (21×21 grid, 441 tokens, half of them on time, signal untouched). All three land within 1.6 percentage points of one another. The unresampled configuration reaches 70.40% against the curve's 70.34% interpolation, a discrepancy of 0.06 pp, and it loses 6.05 pp $[-7.43, -4.67]$ relative to the native reference. The square canvas at the larger budget reaches 71.50%, which is 1.60 pp $[-0.48, +3.68]$ above the 196-token square canvas, so at a fixed allocation the budget contributes a small and not significant amount. The difference between the two 50% configurations that differ in canvas, -1.10 pp $[-3.47, +1.27]$, is not distinguishable from zero. Taken together: reaching the 50% allocation on an unresampled spectrogram reproduces most of the square-canvas penalty, while the resampling that accompanies the square canvas contributes no detectable share of it. Figure 2 marks the three measurements; none of them enters the curve.

5.3 The training budget is not the source

The allocation effect could in principle be a symptom of an under-trained model: a coarse token grid may cost little when the backbone has already extracted what it can, and more when it has not. Table 4 tests that reading by holding the allocation fixed and varying only the schedule, from 40 to 300 epochs, under the same final-epoch selection rule. The two configurations do not converge to the same accuracy: the native advantage is $+8.05\,\mathrm{pp}$ at 40 epochs, $+8.20\,\mathrm{pp}$ at 100 and narrows only to $+6.55\,\mathrm{pp}$ at 300, while the longer schedule adds $15.50\,\mathrm{pp}$ on the square canvas and $14.00\,\mathrm{pp}$ on the native one. Representation and optimization therefore contribute largely independent gains, and the allocation effect is not a short-schedule artifact — which matters for Section 6, where the second corpus is trained for 100 epochs and where the effect is absent.

Table 4: Accuracy against the training budget for the two canvas configurations under the identical protocol.

Model and budget	Square canvas (%)	Native canvas (%)	Difference	Gain from 40 epochs
ViT, 40 epochs	54.40	62.45	$+8.05$	—
ViT, 100 epochs	62.60	70.80	$+8.20$	$+8.20$ / $+8.35$
ViT, 300 epochs	69.90	76.45	$+6.55$	$+15.50$ / $+14.00$

6 Across corpora and tasks

A configuration rule is only useful if its scope is known, so the allocation sweep was repeated on a second corpus with a different task structure. UrbanSound8K comprises 8 732 urban recordings in ten classes with ten predefined folds[21]. Its native spectrogram is 128×401, four seconds at the same front end, and four patch shapes mirror the ESC-50 sweep at a matched budget: 16×8, 8×16, 4×32 and 2×64 give measured grids of 8×50, 16×25, 32×12 and 64×6 tokens at time shares of 86.2%, 61.0%, 27.3% and 8.6%. Because the corpus is larger, the schedule was set to 100 epochs, which corresponds to roughly 24 500 optimizer steps against the 15 000 of the ESC-50 runs. Figure 4 places the two curves on one axis by plotting each corpus against its own reference. The contrast is between a slope and a flat line. On ESC-50 the four configurations span $14.95\,\mathrm{pp}$, falling monotonically as the time share decreases; on UrbanSound8K the same four configurations span $0.67\,\mathrm{pp}$ over all ten official folds, and none of the three paired contrasts differs from zero ($-0.26\,\mathrm{pp}$ $[-2.10, +1.59]$ at a 61.0% time share, $-0.38\,\mathrm{pp}$ $[-2.86, +2.10]$ at 27.3%, and $+0.29\,\mathrm{pp}$ $[-2.45, +3.04]$ at 8.6%), the 27.3% value being the endpoint contrast that Table 5 reports. The most frequency-heavy configuration, which reduces the time axis to

six tokens and refines the mel axis to 64, is in fact the highest-scoring of the four on this corpus, which indicates insensitivity to the allocation on this corpus rather than a preference for frequency-heavy layouts. The two corpora also differ in the structure of the effect rather than only in its size. On ESC-50 every one of the five folds loses between 8.5 and 11.75 percentage points, and the fold-to-fold standard deviation of the penalty ($1.30\,\mathrm{pp}$) is smaller than that of the reference accuracy itself ($3.18\,\mathrm{pp}$), so the loss does not depend on how hard a particular fold is. On UrbanSound8K the same manipulation produces -6.98, -1.91, $+0.11$, $+3.94$, $+1.28$, $+2.31$, -5.61, $+1.36$, $+0.86$ and $+0.84$ percentage points over the ten folds: a range of 10.9 points and a sign that changes. A manipulation that costs the same on every fold of one corpus and a different amount, of either sign, on the folds of another is not a fixed property of the architecture but of the task the model faces on each fold. Table 5 collects both corpora with the two bioacoustic tasks. The mechanism of Section 7.2 accounts for both patterns at once. The penalty is paid when the average spectrum does not separate the classes and the model must compare local spectra across time; it vanishes when the average spectrum already separates them, because then temporal positions carry nothing the frequency axis has not delivered. The two corpora sit at different points on that axis, and the difference is visible in the stationary content each carries: a classifier that sees only the time-averaged spectrum reaches 46.5% on ten-class UrbanSound8K against 28.7% on fifty-class ESC-50. UrbanSound8K's classes come from a global taxonomy of urban sounds, and its average spectrum carries far more of the class information, which is why a coarse temporal lattice costs it little; the fold-to-fold spread of that corpus's penalty is reported below but is not attributed here to any particular pairing of classes. ESC-50's fifty classes include many that share a spectral profile and differ in how it evolves, so the temporal axis is needed on every fold. The companion study tests that reading directly by reducing the class count of one corpus with everything else fixed (Section 6.1). The cross-corpus difference is therefore not a limit on the rule but a statement of when the rule applies. The rule concerns how a fixed token budget should be split between the two axes, and it has teeth exactly when the task requires relating spectral content across time. The screening measurements support that reading and rule out the alternatives: the corpora do not differ in the temporal modulation they carry (the 90% modulation bandwidth is 110 cycles per clip for UrbanSound8K against 103 for ESC-50, and the fraction of temporal modulation energy above the frequency-heavy arm's Nyquist is 35.3% against 35.1%, so a transient-richness account predicts the wrong direction), and they do not differ in

overall difficulty (73.08% against 76.45%). Figure 3(b) shows the four penalties with their intervals side by side, against the four conditions of the axis decomposition in panel (a). The mechanism of Section 7.2 predicts where the penalty should disappear, and that prediction can be tested on tasks from a different domain altogether. If the penalty is paid only when the average spectrum fails to separate the classes, then a task whose classes are spectrally distinct should show no penalty at all, whatever the corpus and whatever the signal. Two primate vocalization tasks are exactly that regime, and they were run on the same platform with a one-shot evaluation on a held-out split, though not with the same front end or the same selection rule as the sweep, since their recorded runs use a shorter window and select on validation: marmoset call classification over the five classes present in the evaluation split of the six-class task, drawn from the MarmAudio dataset[22, 23], which the recipe solves at $99.26 \pm 0.01\%$ over three seeds, and gibbon call detection under a source-audited split, which it solves at $F_1 = 0.976 \pm 0.001$. On both tasks the allocation manipulation is inert. Repeating the sweep's endpoint contrast there gives $-0.15\,\mathrm{pp}$ $[-0.41, +0.10]$ on the marmoset and $-0.86\,\mathrm{pp}$ $[-2.29, +0.58]$ on the gibbon; the square-canvas comparison, which leaves the native time share at 75.8% on these recordings, gives $-0.44\,\mathrm{pp}$ (gibbon) and $-0.08\,\mathrm{pp}$ (marmoset) under the same convention. All four intervals contain zero, and the manipulation that costs $10.20\,\mathrm{pp}$ on ESC-50 costs nothing on either task. Read together with the two corpora above, the four tasks order themselves by one quantity, the share of the class information that the average spectrum already carries: 28.7% on ESC-50 and 46.5% on UrbanSound8K under the protocol of this study, and near ceiling on the two call tasks, where the model's accuracy leaves almost no room for any manipulation to show. The allocation rule therefore has a domain rather than a corpus list: it binds when a task needs local spectra to be related across time, and it is inert when the spectrum alone suffices. The companion study manipulates the class count of a single corpus (Section 6.1), and it is why the rule is reported with its scope attached. Speech Commands v2[24] was screened during development under a protocol that differs from the controlled one.

6.1 Class count and training pool

Section 6 reports that two candidate explanations remain confounded there: the corpus on which the penalty is measured has five times as many classes and about a fifth of the training pool. The

Table 5: The allocation penalty across corpora and tasks. Each row compares a time-heavy reference with the frequency-heavy arm of the same pair at a comparable token budget, within 5%; the equal-budget pairs are in Additional file 1.

Corpus (classes)	Native canvas	Budget	ρ (ref $\rightarrow$ freq.-heavy)	Penalty	95% CI
ESC-50 (50)	128×501	504 / 480	$88.7\% \rightarrow 31.9\%$	-10.20	$[-11.82, -8.58]$
UrbanSound8K (10)	128×401	400 / 384	$86.2\% \rightarrow 27.3\%$	-0.38	$[-2.86, +2.10]$
Marmoset (6)	128×201	200 / 192	$75.8\% \rightarrow 15.8\%$	-0.15	$[-0.41, +0.10]$
Gibbon (detect.)	128×201	200 / 192	$75.8\% \rightarrow 15.8\%$	-0.86	$[-2.29, +0.58]$

All four rows report the endpoint contrast of a sweep. ESC-50 and Marmoset use five and three folds and UrbanSound8K all ten official folds, whose first five give -0.71 pp $[-5.80, +4.37]$ for the same contrast; the gibbon and marmoset conditions are scored over three seeds, and the gibbon as positive-class F_1 rather than top-1 accuracy.

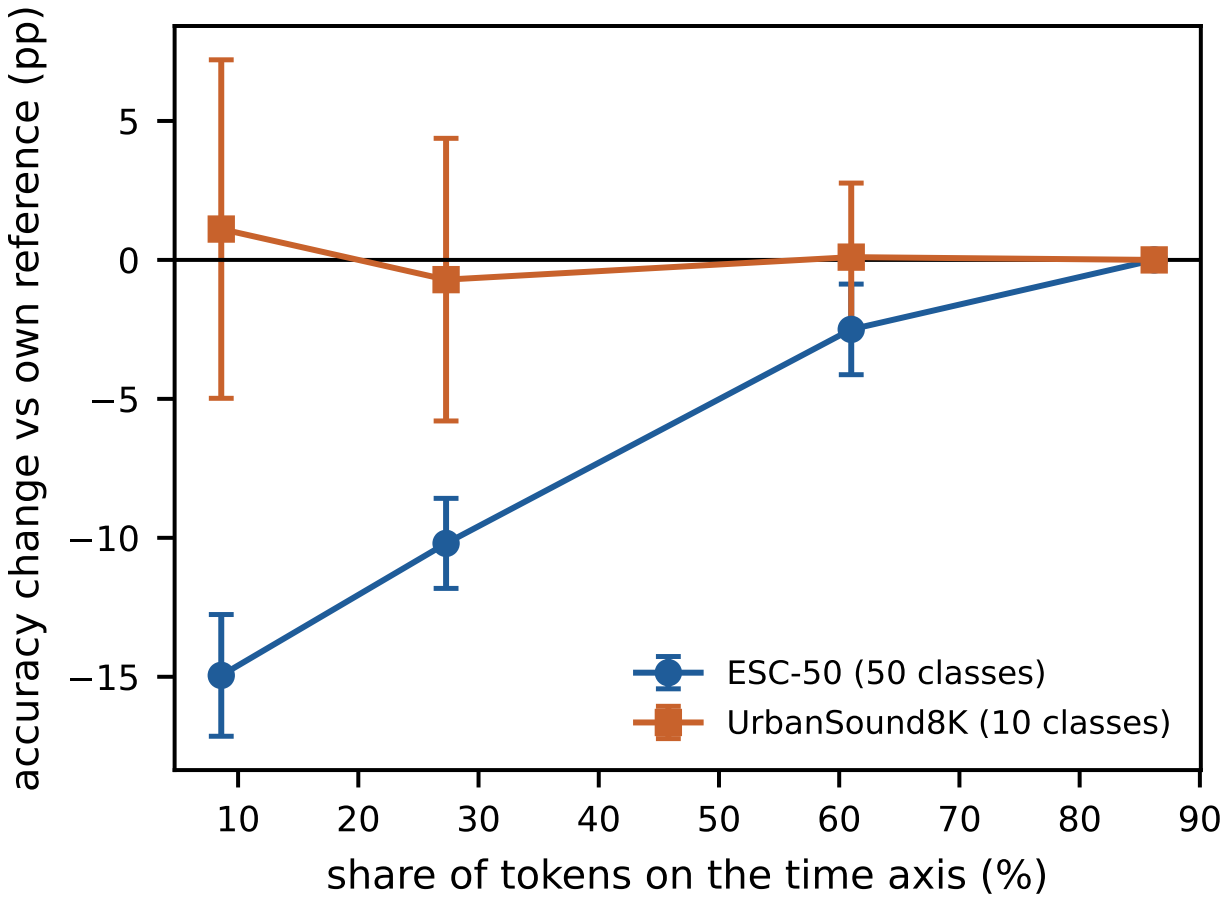


Figure 4: The same allocation manipulation on two corpora. Each corpus is plotted against its own reference so that the shapes can be compared; four configurations span the same allocation range at a matched token budget.

companion study addresses both without leaving the single-corpus controlled setting. The pool statistic is answerable from the data rather than by training, because it is a property of the training set while the quantity that predicts the penalty is a property of the signal: a classifier that sees only the time-averaged spectrum recovers 28.7% of the fifty-class problem on ESC-50 and 46.5% of the ten-class problem on UrbanSound8K, both are properties of the recording rather than of the training set, so neither requires training to measure. The class statistic is manipulated directly, by reducing the class count from fifty to

ten at matched folds and a matched recipe. The two conditions cannot be matched exactly on pool size, because ten classes supply 320 training clips at these folds while fifty supply 1600; the fifty-class arms were therefore subsampled to 300 clips, 6 per class, against the 320 that ten classes supply. The penalty survives the reduction: the matched pair at roughly 500 tokens gives $-4.58\,\mathrm{pp}\ [-8.33, -0.84]$ at fifty classes and $-4.58\,\mathrm{pp}\ [-6.38, -2.79]$ at ten, both intervals excluding zero, and the pair at roughly 250 tokens moves from $-1.17\,\mathrm{pp}\ [-4.81, +2.47]$ to $-3.75\,\mathrm{pp}\ [-9.13, +1.63]$. The class count therefore does not reproduce the collapse, and is excluded as its source, although the stationary-content probe itself was not re-measured on the reduced condition. The study also supplies the contrast the main sweep does not contain. Every arm of the sweep holds the budget at roughly 500 tokens, so the sweep compares allocations at one budget. The companion study repeats the allocation contrast with the two arms of each pair holding the same budget, once at the sweep's own budget and once at half of it, so that the budget is fixed inside a pair and only its size differs between the pairs. The direction of the effect is the same in both, and it opposes the direction found with unequal budgets. At 504 and 512 tokens the frequency-heavy arm is $2.75\,\mathrm{pp}$ below the time-heavy arm, with all three paired folds negative but a 95% interval of $[-6.04, +0.54]$ that still includes zero; at 252 and 256 tokens it is $1.83\,\mathrm{pp}$ below, with a minimum detectable difference of $10.9\,\mathrm{pp}$ at 80% power, so this pair bounds nothing beyond the sign of the effect. The pilot that preceded these pairs, whose arms differ in budget by a factor of two, put the frequency-heavy arm $4.08\,\mathrm{pp}$ ahead. The two matched pairs are tabulated in Additional file 1. The direction of the penalty therefore follows the allocation rather than the budget, once the two arms of a comparison are given the same budget; its size is not resolved by either pair, whose intervals both contain zero.

7 Discussion

7.1 Two cascaded sampling operations

The curve saturates rather than peaking: between an 88.7% and a 99.8% time share, three configurations that impose very different layouts, including one that collapses the entire frequency axis into a single token, are mutually indistinguishable, while below roughly two thirds of the budget on the time axis accuracy falls monotonically. The input path performs two sampling operations whose limits are usually

conflated, and separating them accounts for the decomposition measured in Section 5. Consider an axis of n samples with power spectral density $S(\nu)$ in cycles per clip, mapped by bilinear resampling to m samples. For $m \geq n$ the operation is an interpolation: the result is determined by the original, the map is injective and no modulation content is destroyed. For $m < n$, with $r = n/m > 1$, samples are drawn from a triangular kernel $h[k] = (1/r)\,\mathrm{tri}(k/r)$ whose transfer function

$$H(\omega) = \frac{1}{r^2}\left(\frac{\sin(r\omega/2)}{\sin(\omega/2)}\right)^2 \tag{2}$$

is the Fejér kernel of half-width r, normalized to unit gain at zero frequency, with its first null exactly at the new Nyquist frequency $\omega = 2\pi/r$. Content above it is attenuated and folded into the baseband, from which it cannot be recovered: the map is not injective. The square canvas applies both directions at once: the frequency axis is upsampled ($r_f = 0.571 < 1$), so no spectral content is destroyed, while the time axis is compressed ($r_t = 2.237 > 1$), giving a new Nyquist of $T'/2 = 112$ cycles per clip and a surviving fraction of temporal modulation energy of

$$\kappa_{\text{resize}} = \frac{\int_0^{T'/2} S_t(\nu)\,d\nu}{\int_0^{\infty} S_t(\nu)\,d\nu} = 0.909, \tag{3}$$

evaluated on the measured spectrum of ESC-50. A resize-only account therefore predicts a penalty bounded by the 9.1% of temporal modulation energy the operation removes, and the axis decomposition is consistent with that bound ($1.18\,\mathrm{pp}$ for time alone, an interval containing zero, and $0.08\,\mathrm{pp}$ for frequency alone). The token lattice is a second and much tighter sampling stage: a grid of N_t time and N_f frequency tokens represents modulation up to $N_t/2$ and $N_f/2$ cycles per clip respectively, so

$$B_t \leq \frac{N_t}{2}, \qquad B_f \leq \frac{N_f}{2}, \qquad B_t B_f \leq \frac{N}{4} \quad \text{at a fixed budget } N = N_f N_t. \tag{4}$$

The inequality is symmetric and cannot explain an asymmetric outcome; what breaks the symmetry is the relation between each lattice and the signal it samples. The square canvas forms a 14×14 grid whose temporal lattice Nyquist is $14/2 = 7$ cycles per clip against the 112 of the resize it accompanies:

$$\frac{T'/2}{N_t/2} = \frac{T'}{N_t} = \frac{224}{14} = 16. \tag{5}$$

The lattice is sixteen times coarser than the operation it is bundled with: it retains $\kappa_{\text{lattice}} = 0.649$ of the temporal modulation energy, the same fraction as the 15-token arm, against 0.794 for the 63-token reference and 0.553 for the coarsest arm, and across the six configurations of the sweep that fraction tracks accuracy at $r = +0.857$ (descriptive, six points), while the same quantity on the mel axis is negatively correlated with accuracy ($r = -0.618$; Figure 5 plots both relations, and Figure 6 gives the spectra they are computed from). That contrast is the asymmetry the next subsection addresses.

7.2 The mechanism

The measurements above leave one question: why does a lattice that keeps many temporal positions tolerate an arbitrarily coarse frequency axis, while a lattice that keeps the frequency axis and spends its budget on time does not? This section answers it with one argument in four parts, each part supported by a measurement on the same corpus. Table 6 collects the probe accuracies the argument uses. (1) The two axes are not two instances of the same decision. The frequency axis is an analysis convention and the analysis is redundant: adjacent mel channels overlap, so neighboring coefficients carry strongly dependent content. The time axis is the physical axis on which the recording unfolds and carries the order of events. Coarsening the mel axis replaces the spectrum by a coarser description of the same quantity; coarsening time removes the positions at which the spectrum is observed. The probes give that distinction a magnitude. An element of a token correlates with the token mean that pooled it at between 0.82 and 0.94 on both axes, over every lattice measured, so on either axis the pooling is a faithful summary of what it pooled, and the difference between the axes lies downstream of pooling, in what the tokens let the network do. The probe used here is not the one quoted in Section 6.1: it is fitted on a different representation of the same clips, so the two sets of percentages are not comparable and serve different purposes. A classifier that sees only the spectrum, averaged over time, reaches 33.25% under the companion probe; one that sees only the time evolution, averaged over frequency, reaches 10.45% under the same probe; the trained model reaches 76.45%. The companion probe differs from the one quoted in Section 6.1 in its classifier and in the mel pipeline that feeds it, and not in the clips it is fitted on: on one common feature matrix the classifier moves the result by 1.00 and 0.55 percentage points; the remaining part of the gap comes from the two mel pipelines and is not separately quantified here. The two probes agree on the ordering the argument needs, spectral content far above temporal, but their

absolute values are not interchangeable. Class identity is therefore carried by the spectral signature and the accuracy by how that signature evolves, and the second component is relational: it exists only in a representation that keeps many temporal positions. (2) The two consequences of a coarse lattice are not of the same size. Coarsening the lattice produces two effects at once. One is a loss of resolution in what a token holds: the token mean becomes a less complete description of the content it pooled. This loss is real, it is present on both axes, and the probes show that per token it is actually the larger one on the mel axis, where the within-token dispersion adds $+10.80$ percentage points of probe accuracy on the 16×32 lattice against $+6.80$ for time ($+6.35$ against $+4.55$ on 4×63). The other effect is a loss of temporal positions, and it exists only on the time axis. The two effects are not comparable in consequence, because what they remove from the model is different in kind. The resolution loss removes detail that the spectral summary tolerates, which is why the frequency-heavy arm compresses the time axis to 15 tokens or fewer without paying the full cost everywhere: on the urban corpus, whose classes the average spectrum already separates, that arm lands within 3.5 percentage points of its own reference. The position loss removes the relations themselves, and no resolution inside a token restores a relation between positions: spending the budget on time instead costs up to 14.95 percentage points. The two effects also separate by budget. Within the sweep, whose arms differ in budget by at most 11%, the budget does not move accuracy; and the companion study repeats the allocation contrast with both arms of a pair holding the same budget, at the sweep's budget and at half of it, recovering the same direction in both (Additional file 1). (3) The plateau is the point at which the relations are resolved. The mechanism predicts a curve that rises while temporal positions are scarce and then flattens once they suffice, with the remaining budget free to move to the other axis at a cost limited by the redundancy of the mel analysis. That is the measured curve: the 88.7%, 96.9% and 99.8% time shares are mutually indistinguishable, including the configuration that collapses the mel axis into a single token, while below roughly two thirds of the budget on the time axis accuracy falls monotonically, and within the budget range the sweep covers that fall is not recovered by adding tokens at a fixed allocation. The saturation is therefore not a separate empirical fact but the signature of the mechanism: the constraint has a floor set by the corpus, and above the floor the allocation is free. The argument covers the settings it was measured in: a small transformer trained from scratch on environmental recordings whose classes are distinguished by how their spectra evolve. It makes two predictions for other settings. A corpus

whose classes are separated by spectral content that survives temporal pooling should move the plateau toward the frequency side; content is a sustained spectral signature should be insensitive to the temporal share, as the urban corpus examined above is. Additional file 1 reports the probe records behind each statement in this section, together with the companion study.

Table 6: 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; the absolute values carry a run-to-run spread of up to about 0.5 pp from the randomly initialized classifier, and the order of the rows is stable. Values are from the companion probe, described in Additional file 1, Section 2.1.

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 ($N_f \times N_t$)	252	21.50
$+$ within-token dispersion along time	$+252$	25.65
$+$ within-token dispersion along frequency	$+252$	27.45
Token means of the 16×32 lattice ($N_f \times N_t$)	512	28.70
$+$ within-token dispersion along time	$+512$	35.40
$+$ within-token dispersion along frequency	$+512$	39.45

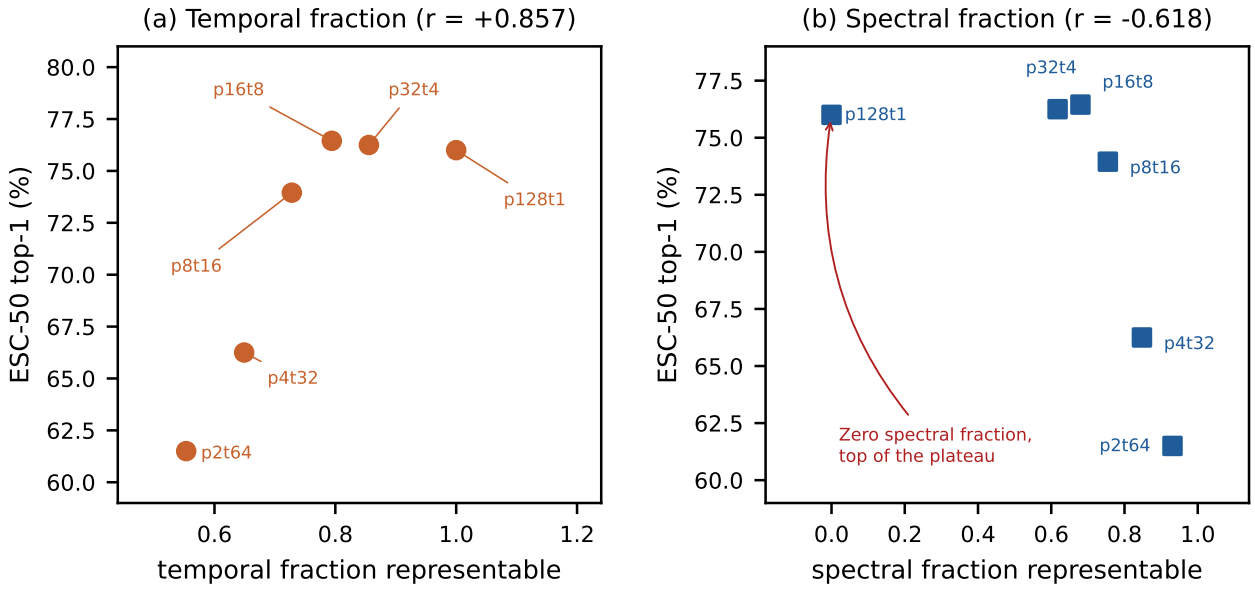


Figure 5: Accuracy against the fraction of modulation energy each configuration can represent. The temporal fraction tracks accuracy while the spectral fraction does not, because the convolutional stem encodes sub-token spectral structure into channels before the lattice pools it; the arm whose spectral fraction is zero nevertheless sits at the top of the plateau.

7.3 The plateau in statistical terms

This subsection reports 90% intervals, while intervals elsewhere in the paper are the usual 95% intervals of a difference test; equivalence testing conventionally reports a $1 - 2\alpha$ interval. The 96.9% and 88.7%

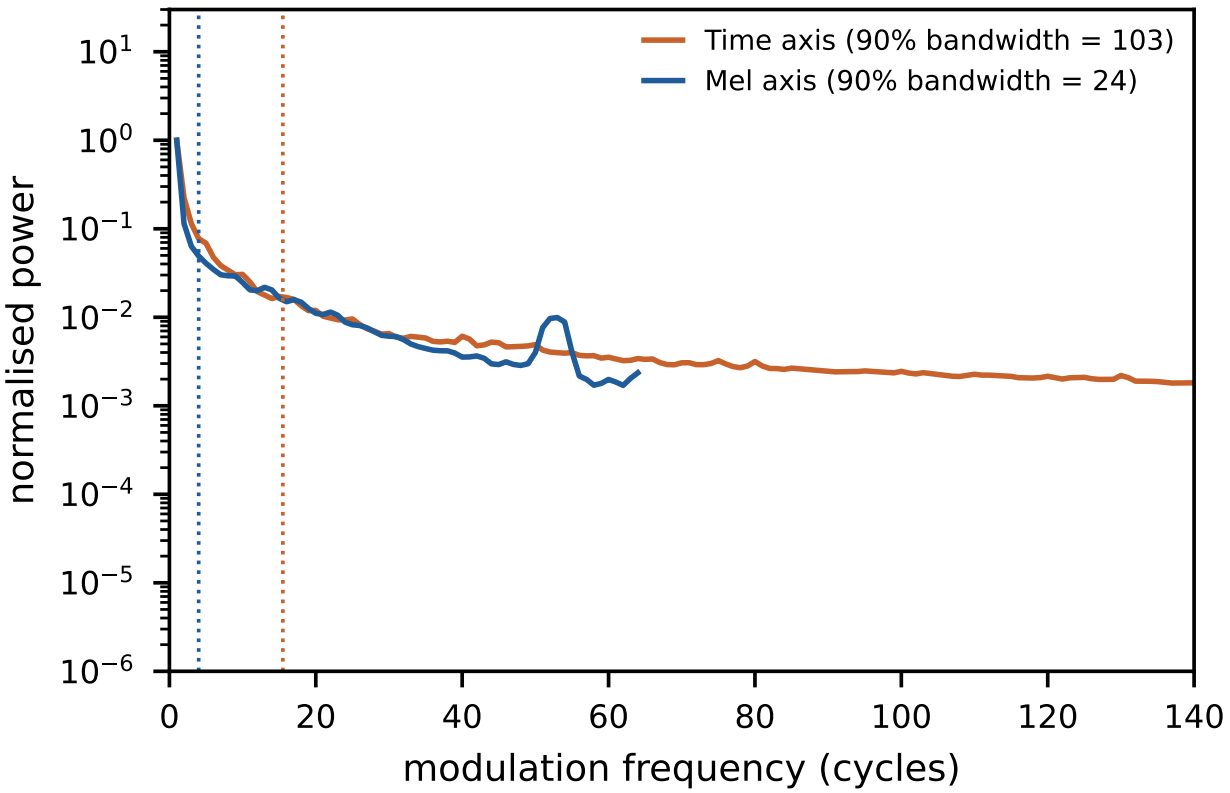


Figure 6: Modulation spectra of the ESC-50 mel spectrograms along time and across the mel axis, with the Nyquist limits of two configurations marked.

arms are not merely indistinguishable; that pair is equivalent within a margin. Testing equivalence rather than difference, the paired difference between the 96.9% and 88.7% arms is $+0.10$ pp over fifteen (seed, fold) cells with a 90% interval of $[-0.86, +1.06]$, inside a margin of ± 2 pp. The outermost plateau pair, 99.8% against 88.7%, has a wider interval and does not meet that margin, so the three time-heavy configurations are reported as indistinguishable rather than as equivalent. The null of a difference of at least Δ is rejected when

$$\hat{\Delta} \pm t_{1-\alpha,\nu} \frac{s}{\sqrt{n}} \subset (-\Delta, +\Delta), \tag{6}$$

which at $\alpha = 0.05$ and $\nu = 14$ uses $t_{0.95,14} = 1.761$ and a margin of 2 pp. The same test on the 66.0% arm gives a 90% interval of $[-3.71, -1.95]$, which is not contained in that margin, so the knee is a difference rather than an underpowered null. The design supports that reading. With fifteen paired cells and an observed standard deviation of 2.11 pp, the smallest difference detectable at 80% power is

$$\delta = \left(t_{1-\alpha/2,\nu} + t_{1-\beta,\nu}\right) \frac{s}{\sqrt{n}} = (2.145 + 0.868) \frac{2.11}{\sqrt{15}} = 1.64\,\text{pp}, \tag{7}$$

so the equivalence margin of $\pm 2\,\mathrm{pp}$ is close to the resolution the experiment actually has. The same reasoning motivates the paired units used throughout: for the contrasts that carry a claim the fold-to-fold variance component exceeds the seed-to-seed one, so comparing configurations on the same folds and seeds removes the dominant nuisance term. The present results locate the published observations on a common axis: the configuration used by full-frequency temporal patching[10] sits on the plateau measured here rather than at a distinguishable optimum, and the endpoint comparisons of AST and SSAST are two points on the rising part of the same curve. What the sweep adds is the shape between and beyond those points. Three statements follow for practice. The cost of a square canvas is not the interpolation but the reallocation it bundles with it, and the reallocation can be avoided without touching the input. A layout that keeps the native spectrogram and refines the frequency axis, pushing the time share down, is the one to avoid. And the rule binds only where the task requires relating local spectra across time, which is measurable before the model is trained: the probe on the time-averaged spectrum says which regime a corpus is in. The claims hold for the platform and corpora they were measured on. The allocation curve is measured on ESC-50, where three of its six arms are replicated over three seeds and five folds and the remaining three at one seed and five paired folds; the plateau and the degradation boundary are carried by the replicated arms. The second corpus used a shorter schedule because its pool is roughly five times larger, and on ESC-50 the penalty is present at 40, 100 and 300 epochs alike, which fixes the schedule as an independent axis rather than a confound. The platform is one small model trained from scratch: a different stem or depth, or pretrained weights, would move the location of the plateau and the size of the penalty, and comparisons against convolutional or hybrid architectures lie outside the design. The mechanism gives a testable prediction: a corpus whose classes are separated by spectral content that survives temporal pooling should move the plateau toward the frequency side. A second direction is to make the allocation adaptive at inference time, in the manner of the resolution-flexible audio transformers[8], since the measurements here identify which direction of adaptation is cheap.

8 Conclusions

This study compared token allocations for a small from-scratch audio transformer on one fixed 10.84M-parameter platform. Sweeping the share of roughly 500 tokens placed on the time axis from 9.9%

to 99.8% showed that accuracy rises with the time share, saturates across the plateau arms, and falls monotonically toward the frequency-heavy end over a range of 14.95 percentage points. The penalty of a square canvas is an allocation effect rather than an interpolation effect: resampling either axis alone cost nothing detectable, while reaching the same 50% allocation on the unresampled spectrogram reproduced most of the penalty and landed within $0.06\,\mathrm{pp}$ of the curve's interpolation. Extending the schedule from 40 to 300 epochs improved both canvas configurations by 14.00 to 15.50 percentage points without closing the gap. Across corpora the penalty was nearly constant across folds on ESC-50 and no larger than 3.5 percentage points on a ten-class urban corpus; the probes locate the asymmetry in the number of temporal positions the lattice can relate rather than in the information each axis carries. The statistical checks and the companion study are in Additional file 1.

Supplementary information

- Additional file 1. File name: supplementary.pdf. File format: PDF.
- Title of data: Statistical robustness and the companion study.
- Description of data: the nuisance-variation decomposition, the resolution of the two cross-corpus measurements, the permutation tests, the probe accuracies used by the mechanism argument, the matched-budget pairs, and the design of the companion study.
- Additional file 2. File name: code_attachment.zip. File format: ZIP. Title of data: code and records for reproduction. Description of data: the analysis code, the per-run training records, and a script that recomputes the tabulated results.

Declarations

Availability of data and materials. The ESC-50, UrbanSound8K and MarmAudio datasets analyzed in this study are publicly available and are cited in the reference list. The gibbon recordings are held by the Chinese Academy of Forestry; the subset, segmentation and partition used here were prepared by the authors, and the split has not been audited for physical-recording independence. The probe measurements and the design of the companion study are documented in Additional file 1.

Ethics approval and consent to participate. Not applicable: the study uses only publicly available, previously published audio datasets and recordings supplied by their owner, and involves no human or animal subjects.

Consent for publication. Not applicable.

Code availability. The analysis code, the per-run records, and a script that recomputes the tabulated values are provided as Additional file 2. The training code depends on modules that are not redistributed; every reported number is derived from the archived records.

Competing interests. The authors declare that they have no competing interests.

Funding. This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Authors' contributions. **Weiwei Xiao:** Conceptualization, Methodology, Supervision, Writing – review and editing. **Jianmin Liu:** Software, Validation, Formal analysis, Investigation, Data curation, Visualization, Writing – original draft. Both authors read and approved the final manuscript.

Use of AI. The authors used DeepSeek, a large language model, to assist with code development, the analysis pipeline, and manuscript preparation and language polishing. All numerical results reported here were verified against the archived per-run training records, and the authors take full responsibility for the content of the paper.

Acknowledgements. The authors thank the providers of the ESC-50, UrbanSound8K and MarmAudio datasets for making their recordings publicly available, and the Chinese Academy of Forestry for access to the gibbon recordings used in the cross-corpus comparison.

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).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- supplementary.pdf