PEOPLE OVER SEGMENTS: A DATA-AXIS SCALING LAW FOR PERSON-IDENTITY CONTRASTIVE PPG PRETRAINING

Anonymous authors
Paper under double-blind review

ABSTRACT

Scaling data is the central lever behind foundation models, but a physiological corpus cannot be scaled like a web corpus: every additional segment must be recorded from a person, so growth is bounded not by what can be scraped but by who can be enrolled. A fixed segment budget therefore splits as K distinct people $\times$ M segments per person. Should a collector recruit more people, or record more from each? Both spend the same budget, yet for raw physiological waveforms the question is unsettled. We study this question for person-identity contrastive self-supervised learning on a real-world photoplethysmography (PPG) cohort of 114k people, in over 200 controlled pretraining runs. The number of *distinct people* seen in training is the stronger axis: segments per person help, but more slowly, and the most people-heavy allocation we sampled is best at every budget. More people do not merely lower the pretraining loss; they change *what* the representation encodes, separating held-out people more cleanly by identity, in step with better demographic prediction. A fitted scaling law makes the asymmetry quantitative: the people axis buys down its share of the error nearly twice as fast per decade as the segment axis, an ordering that holds across the fitting choices we test. Solving the law for its optimum yields a budget-adaptive acquisition rule: at small budgets it recruits more people than any allocation we sampled, and as the budget grows it shifts toward more segments per person. Training the rule's recommended points confirms it: each matches or beats the best allocation sampled at its budget. The message is simple: grow the number of distinct people before the volume of signal per person.

1 INTRODUCTION

Wearable devices have made photoplethysmography (PPG) one of the most abundant physiological signals ever recorded, and almost all of it is unlabeled. Self-supervised learning (SSL) on such corpora is accordingly the dominant route to PPG foundation models, which transfer to health tasks from demographic attributes to cardiovascular screening (Abbaspourazad et al., 2024; Pillai et al., 2025; Saha et al., 2025), and the question that follows is how to scale them (Kaplan et al., 2020; Hoffmann et al., 2022). But a physiological corpus does not scale like a web corpus: it cannot be scraped in bulk, because every additional segment has to be recorded from a person—someone new, or someone already enrolled, monitored for longer. Outside a handful of industry cohorts, the corpora available for pretraining show what that constraint produces: MESA contributes about two thousand participants, each recorded for one overnight sleep study (Chen et al., 2015); PulseDB assembles 5.2M segments from only 5.4k clinical patients (Wang et al., 2023), roughly a thousand per person, recorded in the ICU or the operating room. Rich in signal per person and thin in people, such corpora have already picked a side of a trade-off, usually by default rather than by design.

We formalize this trade-off as a budget allocation problem. The cost of a pretraining corpus is, to first order, proportional to the segments it contains, $B = K \cdot M$, for K distinct people at M segments each. B is finite for every collector, whether funding and wear time or training compute sets the limit, and more resources only pose the same choice at a larger B: many people at few segments each, or few people at many? The split cannot be tuned after the fact: it fixes the acquisition protocol before a single model is trained, and choosing wrong buys a needlessly worse representation for the same total

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.

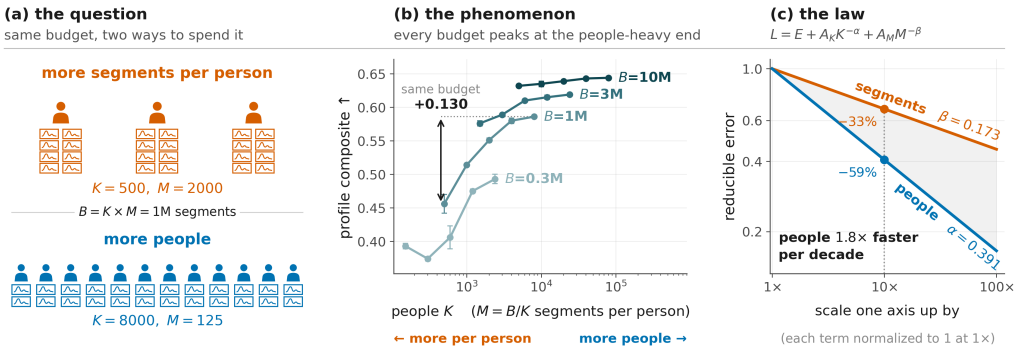

Figure 1: **People over segments.** Under a fixed budget of $B = K \cdot M$ segments (K people, M segments per person), recruiting more people beats recording more from each. **(a)** The choice: 1M segments from 500 people at length, or from 8000 briefly (icons schematic). **(b)** Every fixed-budget diagonal peaks at its people-heavy end: $+0.130$ in profile composite (mean probe skill on age, BMI, and sex) at B=1M. **(c)** The fitted law: per decade, the people term shrinks by 59% and the segment term by 33%; curves are fitted terms, not measurements.

wear time and training compute. Nor is the right choice obvious. More segments from one person buy *longitudinal* coverage (rest, exertion, sleep), so they supply harder positive pairs rather than merely more of them; and they avoid the consent, screening, and retention overheads that come with every new participant, none of which a segment count captures (Truslow et al., 2024). Both arguments pull segment-heavy, and the published corpora follow them. Existing scaling laws do not settle it: they ask how large a corpus should be, not how it is distributed over the people who produced it. The one wearable study we know of that poses the question finds that, at equal total hours, trading subjects for hours per subject changes performance little, so that total hours rather than subject count drive gains (Narayanswamy et al., 2025); but it does so on minute-level derived features under a masked-reconstruction objective, not on raw waveforms under an objective that tells people apart.

We study this question for person-identity contrastive learning, a recipe used in large-scale PPG pretraining, in which two segments from the same person form a positive pair—so identity is the discriminative target, and headcount a design variable rather than an incidental property of the pool. We find that people pay off faster than segments per person and that, at every budget we test, recruiting more people beats recording more from each (Figure 1). We make three contributions. **(a) A decomposition of the acquisition budget, and a design that identifies both axes.** A fixed-budget diagonal at each of four budgets (0.3M to 10M segments) isolates the allocation decision; orthogonal sweeps move one axis at a time so each can be measured alone, in over 200 single-factor runs on a 114k-person cohort, with both sweeps and the 1M diagonal replicated at three model capacities. **(b) K-dominance, and what it changes in the representation.** The people-heavy allocation wins at every budget, by the largest margins at the two smallest, and it does so not only by optimizing the objective better (the pretraining loss already favors it) but by changing what is encoded: it separates held-out subjects more cleanly by identity, and within each budget this separation tracks the profile composite (mean probe skill on age, BMI, and sex). **(c) A scaling law, and the acquisition rule it yields.** Fit as a first-order law, the two axes separate (people exponent $\alpha = 0.391$ vs. segment exponent $\beta = 0.173$): a decade of people removes 59% of its reducible term, a decade of segments 33% of its own. The exponents depend on how the law is estimated, the ordering much less: it holds under every treatment of the noisiest arms, widens when a cross term relaxes separability, and survives a wide range of assumed error floors. We therefore claim the direction rather than an exact exponent. Solving the law for its optimum then converts a segment budget into how many people to recruit and segments to record. We forward-validate this allocator by training its recommended points: each matches or beats the best allocation sampled at its budget. At the scarcest budget the recommended point is more people-heavy than any grid arm and beats all of them, so the rule extrapolates rather than merely restating the grid; at the largest it reaches the same composite while recruiting 72% fewer people, sparing the per-person overheads that B does not count.

2 Related Work

Foundation models for wearable physiological signals. Pretraining on wearable data splits by the level of signal the model consumes. One line works on *derived summaries* (heart rate, heart-rate variability, sleep, activity) aggregated to the minute or the hour: the LSM line (Narayanswamy et al., 2025; Xu et al., 2025) pretrains on 40M hours of wearable data and SensorFM (Narayanswamy et al., 2026) on over a trillion minutes, both by masked reconstruction (He et al., 2022), while WBM (Erturk et al., 2025) applies person-level contrast to hourly behavioral metrics. The other works on *raw waveforms*, where per-beat morphology survives, and it is here that our setting sits. Reconstruction carries over (Lee et al., 2026; Zhou et al., 2026), but the prevailing route is contrastive: Abbaspourazad et al. (2024) take person identity as the source of positives, while open models keep the frame and vary the target, such as signal morphology (Pillai et al., 2025), motif-based relative distances over uncurated field data (Saha et al., 2025), or alignment to synchronized ECG (Nie et al., 2026). Across both levels, how the corpus is distributed over people is usually left implicit. The exception is LSM (Narayanswamy et al., 2025): on per-minute derived features and a reconstruction loss, it compares subjects with hours per subject at equal total hours and finds that "total number of hours rather than number of subjects drives gains". On raw waveforms, with an identity-discriminating objective and a frozen downstream probe, we find a clear people-first ordering.

Contrastive learning and the discriminative target. Within that contrastive route, methods differ in what they discriminate. The standard objectives (Chen et al., 2020; van den Oord et al., 2018; Grill et al., 2020; Bardes et al., 2022) are all built around augmented views of a single sample; person-identity contrast sits at the far end of the spectrum, its positives being different segments of the same person, so the task is to tell people apart rather than views apart. That placement marks the boundary of our claim: where identity is regularized away instead of targeted (Yarici et al., 2025), the people axis need not dominate.

Neural scaling laws and corpus composition. Scaling laws describe how error falls as resources grow and how a fixed budget is best allocated, classically between model size and tokens (Kaplan et al., 2020; Hoffmann et al., 2022). Refinements that look inside the corpus still count samples, not people: how often one may be repeated (Muennighoff et al., 2023; Goyal et al., 2024), how many must be distinct (Hammoud et al., 2024), which are worth keeping at all—a choice that can bend the exponent itself (Sorscher et al., 2022). But a corpus can consist entirely of distinct, non-repeated segments and still come from very few individuals. Neighboring biological domains already report that scale alone underdelivers (DenAdel et al., 2026; Jiang et al., 2025). To our knowledge no prior work fits a scaling law over the two axes, or turns one into an acquisition rule.

3 Method

The recipe below is what every run holds fixed; the scaling methodology that follows (the budget decomposition, the two experimental designs, and the form of the law) is what varies.

3.1 Pretraining Recipe

We build on the participant-contrastive recipe of Abbaspourazad et al. (2024), in which *person identity*—not augmentation—is the discriminative target: two segments drawn from the same person form a positive pair, all segments from other people in the batch are negatives, and the remaining same-person segments are masked out of the InfoNCE (van den Oord et al., 2018) denominator so a subject is never contrasted against itself. Headcount therefore enters the objective itself. The rest is conventional machinery. An EMA teacher (He et al., 2020; Grill et al., 2020) supplies the targets, and two terms guard against embedding collapse: a KoLeo spreading term (Sablayrolles et al., 2019), and a VICReg variance–covariance regularizer (Bardes et al., 2022) that is an addition to that recipe and measurably raises the embedding's effective rank (Appendix B). Each segment of a pair is z-normalized and augmented with standard time-series transforms (Um et al., 2017), and the encoder is a 1D adaptation of EfficientNet (Tan & Le, 2019). Every downstream evaluation reads its frozen embedding (Appendix A.2).

3.2 Scaling Methodology

Decomposing the data budget. The cost of building a pretraining corpus is, to first order, proportional to the total number of segments collected,

$$B = K \cdot M, \tag{1}$$

where K is the number of distinct people recruited and M the number of segments collected per person. "How much data to collect" therefore splits into a sharper question with a free variable: at a fixed budget B, how much goes to headcount K and how much to segments-per-person M?

Two experimental designs. No one design answers that question on its own, so we probe the (K, M) plane with two.

- **Diagonal grid.** Holding the budget fixed at $K \cdot M = B$ and varying only the allocation gives a controlled qualitative read on which side wins at equal B. Because the two axes move in lockstep, it cannot say how much either contributes on its own.
- **Orthogonal sweeps.** Moving one axis at a time (sweep K at fixed M, and M at fixed K) measures each alone, which is what allows the joint fit to separate the two exponents. Because the budget changes along them, they in turn say nothing about a fixed-budget trade-off.

Together they cover the plane as a cross plus a diagonal. Every comparison is kept single-factor (one fixed evaluation set, nested sampling across budgets, and seed pairing), so that only the allocation under study varies (Appendix A.3).

Form of the law. The sweeps give two curves; a functional form is what turns them into two comparable rates. We model downstream quality on the error side (writing $L(K, M)$ for a loss that decreases as the representation improves) as an additively separable power law,

$$L(K, M) = \underbrace{E}_{\text{floor}} + \underbrace{A_K\, K^{-\alpha}}_{\text{people axis}} + \underbrace{A_M\, M^{-\beta}}_{\text{segments axis}}. \tag{2}$$

The form is standard—a floor plus one reducible term per resource (Kaplan et al., 2020; Hoffmann et al., 2022); our departure is the pair of axes it is applied to. Separability is what makes the comparison possible: α and β are the rates at which each axis shrinks its own error term, so their ordering says which axis pays off faster per decade. We treat Equation 2 as a first-order working model and test separability in Appendix E.5; Section 6 fits it.

4 Experiments

This section fixes the three things every later comparison rests on (the pool the budgets are drawn from, the ruler they are read with, and the setup held constant across them), so that two runs differ only in their allocation (K, M).

4.1 Datasets

Pretraining pool. All pretraining draws from a single large-scale PPG cohort of 114,369 people and 53.6M segments, a segment being a non-overlapping 30 s window of 50 Hz green-channel PPG—so M=125 is about an hour of retained signal per person and M=2000 about seventeen, spread over days to months of free-living wear (Appendix A.1). The cohort is comparable in size to those behind published wearable foundation models (Abbaspourazad et al., 2024; Narayanswamy et al., 2025) and lets K span nearly two orders of magnitude. We hold out a fixed set of 5,000 evaluation people once at the start, leaving a training pool of 109,369 people, and every *arm* of the study (one (K, M) configuration, trained from scratch) is nested-sampled from this pool so that no budget ever sees an evaluation subject.

Downstream tasks. The frozen representation is evaluated on the panel of held-out tasks summarized in Table 1, following the demographic-and-cardiovascular probes standard in this literature. The profile composite runs on all 5,000 held-out people; the clinical endpoints run only on the subsets carrying a reference measurement or a clinical label, one to two orders of magnitude smaller (Appendix A.1).

Table 1: Datasets and their roles. A fixed set of 5,000 people is held out of the source cohort for evaluation, leaving 109,369 to pretrain on. All downstream tasks are evaluated with a linear probe on the frozen representation. OSA: obstructive sleep apnea; AFib: atrial fibrillation.

Stage	Task	Scale	Role
Pretraining	PPG contrastive	114,369 people, 53.6M seg.	Source cohort
Downstream	Profile: age / BMI / sex (reg.)	5,000 people	Primary
	Blood pressure: SBP / DBP (reg.)	200 people (paired cuff)	
	Hypertension (cls.)	510 people (200 + 310)	Auxiliary
	OSA / AFib (cls.)	200 / 34 people	

4.2 Evaluation Metrics

Profile composite. The primary dependent variable is the *profile composite*, the mean skill of the three profile regressions, based on their mean absolute error (MAE),

$$\text{composite} = \frac{1}{3} \sum_{t \in \{\text{age, BMI, sex}\}} \left(1 - \frac{\text{MAE}_t}{\text{MAE}_t^{\text{base}}}\right), \tag{3}$$

where $\text{MAE}_t^{\text{base}}$ is the error of a constant predictor. Each term is therefore a unit-free *skill* rather than an error, capped at 1 but unbounded below: 1 is a perfect predictor, 0 is no better than the baseline, and a probe worse than the baseline scores below zero. Averaging the three puts age, BMI, and sex on one axis and makes the composite comparable across budgets. Why this composite, rather than any of the other tasks, is the right ruler is argued in Section 5.1.

Auxiliary metrics. Two further signals are read alongside the composite. The SSL model's own *valid loss* measures optimization of the pretraining objective rather than transfer, so it is a diagnostic: it falls as either data axis grows and, at a fixed budget, favors people-heavy allocations (Appendix C.1). The remaining downstream tasks are reported in their natural units (blood pressure as MAE, the clinical endpoints as AUC) but have one to two orders of magnitude fewer labeled subjects than the composite, too few to resolve the differences the scaling curves turn on (Appendix C.2). Neither is a scaling target; both are checked only for agreement in direction.

4.3 Implementation Details

The encoder is a 1D-EfficientNet producing a 256-dimensional embedding, trained for 10 epochs at global batch size 4096 with AdamW (Loshchilov & Hutter, 2019) under a cosine schedule (Loshchilov & Hutter, 2017). Every hyperparameter is held fixed across every budget arm; the encoder is then frozen and a linear probe fit on its embeddings, which makes each comparison one of representations rather than of fine-tuning capacity. Every arm is repeated over 3 seeds, with error bars reporting seed standard deviations. Appendix A.2 gives the full specification.

5 Results

Two preliminaries fix the objective and the ruler; the $K \times M$ grid then shows that more people wins at a fixed budget, and identity measures on held-out people show what that changes in the representation.

5.1 Preliminaries: the objective and the ruler

Two choices are fixed before any scaling result. We scale identity-contrastive because it wins by a tier: across 7 SSL objectives on one fixed 1M-segment pool under one frozen linear probe, the two

identity-contrastive rows lead by $0.15+$ over every objective whose target is a weakened, proxied, or absent form of person identity (Appendix B). We read every scaling conclusion off the profile composite (Equation 3), committed to on three *a priori* grounds: $5{,}000$ held-out people, so it resolves small cross-budget differences; a dimensionless skill, so one curve reads across budgets; attribute regressions, free of small-sample class imbalance. Valid loss and the clinical endpoints were set aside on principle rather than on outcome (Appendix C).

5.2 The $K \times M$ grid: more people wins at every budget

The diagonal grid instantiates the collector's decision (at a fixed budget $B = K \cdot M$, many people with few segments each, or few with many) at four budgets $B \in \{0.3\text{M}, 1\text{M}, 3\text{M}, 10\text{M}\}$, 5 allocation points on the $K \cdot M = B$ diagonal within each, $\times$ 3 seeds, for 60 runs. The two ends, `mostpeople` and `mostdata`, hold M=125 and M=2000, so headcount shrinks $16\times$ between them (at 1M, 8000 people $\times$ 125 segments versus 500 people $\times$ 2000 segments).

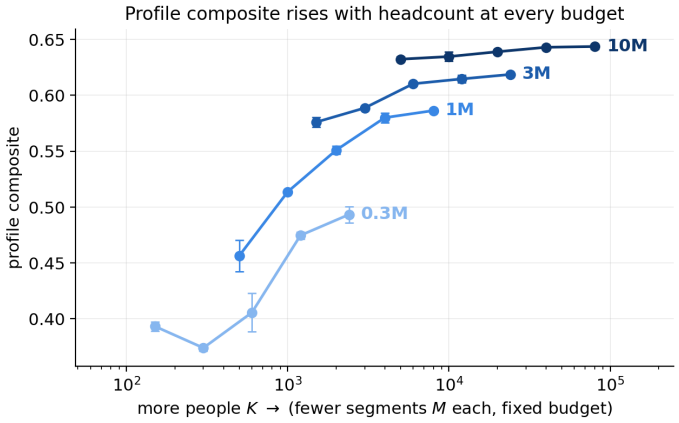

Figure 2: Profile composite along each fixed-budget diagonal, one curve per budget, against headcount K (log-spaced; more people to the right), means over 3 seeds.

The pattern repeats at all four budgets (Figure 2; numbers in Appendix E.1, Table 5): every curve peaks at its people-heavy end and falls away toward the segment-heavy end, so at equal B the most people-heavy allocation is best—K-dominance throughout. The one departure is at 0.3M, where `mostdata` (0.393 ± 0.004) edges above its neighbor `moredata` (0.374 ± 0.003) outside seed noise; both still sit $0.10+$ below that budget's people-heavy end (0.493) and are its lowest-headcount, highest-leverage cells (150 and 300 people), which Section 6 refits with and without.

What the allocation is worth depends on the budget: the `mostpeople`$-$`mostdata` gap is $+0.100/+0.130/+0.043/+0.012$ at $0.3/1/3/10$M. A large budget dilutes even a poor allocation; a scarce one cannot, so the allocation matters most at the budgets where data is hardest to collect.

5.3 What headcount changes: identity structure

Both axes improve the composite on their own (Section 6), yet at a fixed budget people help more; we ask what headcount changes in the representation.

Each frozen encoder is scored on the $5{,}000$ subjects held out of every training pool by leave-one-out person-identification recall at rank 1 (R@1) and by a person-level silhouette coefficient, two views of whether it separates identities it was never trained on. Along every fixed-budget diagonal, moving the budget from segments to people raises both (Figure 3), so the identity resolution is generalized—learned from seeing more distinct people, not memorized from seeing more of each; a by-person t-SNE of held-out subjects shows the same (Appendix D, Figure 7).

Age, BMI, and sex are stable within a person, and the probe reads them from each person's mean embedding; within each budget, the composite rises in step with identity separation (Figure 3). More segments per person add views of identities already seen, whereas more people add new ones, consistent with headcount being the stronger axis.

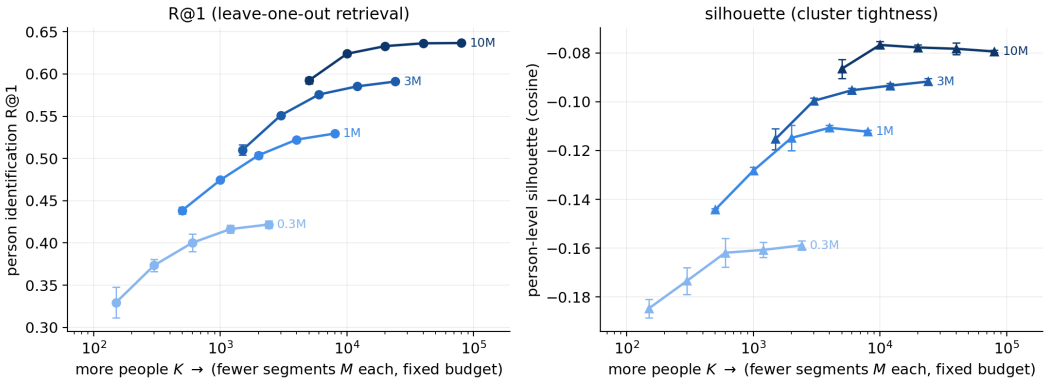

Figure 3: **More people, cleaner identities.** Person-identification R@1 (leave-one-out cosine top-1 retrieval, 32 segments/person; left) and person-level silhouette (right) on the 5,000 held-out subjects. Along every fixed-budget diagonal (shades = budget; the diagonal of Figure 2), headcount raises R@1 monotonically ($+0.09/+0.09/+0.08/+0.04$ at $0.3/1/3/10$M) and lifts silhouette; R@1 tracks the composite within each budget (Spearman $\rho \geq 0.9$). Means over 3 seeds, error bars $\pm$s.d.; silhouette is negative throughout (expected at 5,000 identities), read only for its trend.

The model side is consistent. At three capacities of the same backbone (0.49M / 3.12M / 22.8M parameters), the 1M diagonal favors people-heavy allocations at every size. Along the orthogonal sweeps, extra capacity is wasted while data is scarce on either axis; once it pays off, the large model's advantage holds as K grows but narrows slightly as M grows, as a diversity signal predicts, though the effect is modest (Appendix D.2). On the other tasks, blood pressure moves the same way on its 200-person cohort but with large seed variance, and the clinical classifiers show no consistent trend in K; their cohorts are too small to resolve one (Appendix C.2).

6 A Scaling Law for People vs. Segments

We now turn the grid's qualitative finding into a quantitative rule: fit a law with independent exponents, check how well it extrapolates, then read off a budget allocator.

6.1 Fitting the law: $\alpha > \beta$

The diagonal grid cannot separate the two axes; the orthogonal sweeps can, and we run them under the grid's exact protocol (same evaluation set, seeds, and data-pool split):

- ***K*-sweep**: fix M=250, sweep $K \in \{1000, 2500, 5000, 10000, 25000, 50000\}$;
- ***M*-sweep**: fix K=10000, sweep $M \in \{10, 50, 125, 250, 500, 1000\}$.

The two sweeps add 36 runs (3 seeds each); with the grid's 60, these 96 runs realize the 30 distinct (K, M) points the law is fit on, covering the plane as a "cross plus diagonal". The three sources merge safely: the one configuration reached by two independent paths agrees to 0.0013 in composite, well inside seed noise (Appendix E.2).

Fitting the separable power law of Equation 2 on the error side ($L = 1 -$ composite), jointly across all observations, gives

$$L(K, M) = E + A_K\, K^{-\alpha} + A_M\, M^{-\beta}, \qquad \alpha = 0.391 \pm 0.016 \;>\; \beta = 0.173 \pm 0.037, \quad (4)$$

where α (people) and β (segments) are the rates at which each axis shrinks its own error term, and the fit is stable (weighted $R^2 = 0.963$, every restart reaching the same basin). So $\alpha > \beta$ means people pay off faster: a decade of people cuts its own term by $1 - 10^{-\alpha} \approx 59\%$, against $1 - 10^{-\beta} \approx 33\%$ for a decade of segments (Figure 4c). One point resists the separable form: at M=10 the fit overpredicts composite by ≈ 0.10, the only residual far outside seed noise (Figure 4b). Keeping it is the conservative choice: it is what pulls β up, so dropping it would widen the gap we report.

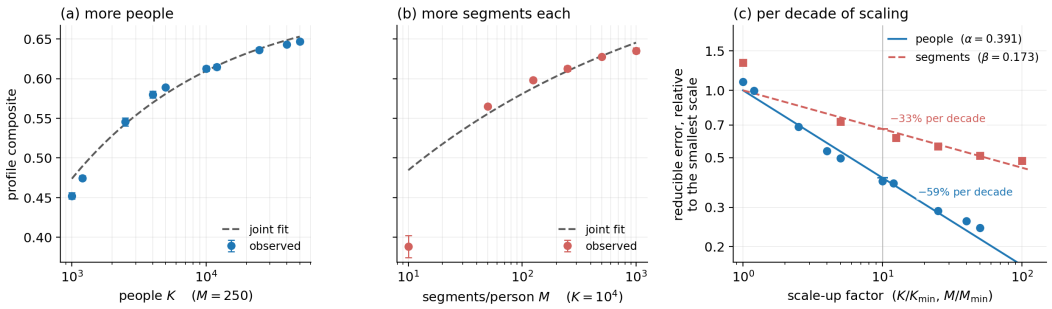

Figure 4: The fitted law along each sweep, and the comparison its exponents make. **(a, b)** Composite against K at fixed M=250 and against M at fixed K=10000, 3-seed points with the joint fit of Equation 4; in these raw units the two sweeps look alike, which is why the asymmetry cannot be read off them. **(c)** The same data with the floor and the other axis's constant term removed and each axis normalized by the fit at its smallest measured scale, so that only the slopes compare: per decade, the people term shrinks by 59% and the segment term by 33%.

The exponents themselves are less certain than their ordering, because floor and exponent trade off against each other: a curve falling steeply toward a high floor is hard to tell from one falling gently toward a low floor, and a single sweep cannot break the tie. Fit each alone, each with its own floor, and all the curvature is absorbed into the exponent, giving $\hat{\alpha} \approx \hat{\beta} \approx 0.65$ on both. What breaks the tie is the joint fit: all points share one irreducible error E, and each sweep's own asymptote is E plus the other axis's fixed term, so the two curves constrain each other; only then do α and β separate (Appendix E.3). Such dependence on the estimation procedure is not peculiar to our setting, as re-examinations of the Chinchilla fit found the same (Besiroglu et al., 2024; Li et al., 2025); we therefore claim K-dominance rather than a particular value of α. That ordering is what proves durable: it survives every profiled floor $E \lesssim 0.27$ (the free fit lands at $\hat{E} = 0.13$), a budget-only model collapsing both axes into KM fits much worse, and relaxing separability with a cross term only widens the gap (Appendix E.5).

6.2 Robustness and Held-Out Extrapolation

Edge-policy ablation. The 0.3M budget's two most segment-heavy arms (`moredata`, 300 people; `mostdata`, 150 people) are noisy and high-leverage: trained on so few people, they sit at the far left of the $\log K$ axis and pull hard on the slope. Refitting under three policies for them (keep, downweight, drop) never flips the direction (Appendix E.4, Table 6): removing the noisy arms steepens both exponents together, and under all three the scale-invariant ratio $(1-10^{-\alpha})/(1-10^{-\beta})$ stays in $[1.6, 2.0]$. We report downweight as the point estimate with {keep, drop} as bounds.

Held-out budget extrapolation. A law must extrapolate, not just fit, so we hold out a whole budget and predict it from the others (Hoffmann et al., 2022): 3M to test interpolation, 10M to test extrapolation. Trained on the remaining budgets alone (Figure 10), interpolation is reliable: holding out 3M recovers the arm ordering exactly ($\rho = 1.00$, best arm correct, MAE 0.0075). Holding out 10M, past the largest budget measured, turns optimistic by about 0.01 composite. Restoring the two sweeps to the training set puts the separable and interaction models on one footing (Table 8), and there the cross term removes that optimism almost entirely: the bias is a limit of the separable working model—the budget-level term it omits (Appendix E.5)—not evidence against the direction.

6.3 The Acquisition Formula: A Budget-Adaptive Allocator

Minimizing L at a fixed budget $B = K \cdot M$ gives the optimal split

$$K^*(B) = \left[\frac{\alpha A_K}{\beta A_M} B^{\beta}\right]^{1/(\alpha+\beta)}, \qquad M^* = \frac{B}{K^*}, \qquad \text{so} \quad \frac{M^*}{K^*} \propto B^{(\alpha-\beta)/(\alpha+\beta)}. \tag{5}$$

Because $\alpha > \beta$ that exponent is positive (here 0.387), so M^*/K^* rises with the budget: the optimal split shifts toward more segments per person as the budget grows.

Solving the law is not the same as trusting it: the optimum lands at splits never trained. We therefore train all four recommended points and read off their composite (Table 2). At every budget the recommendation matches or beats the best grid arm, and at the scarcest it does more than match: at 0.3M it places $K^* = 7{,}769$, more people-heavy than any grid arm (`mostpeople` has $K = 2{,}400$), and gains $+0.036$ composite over it—well outside the seed band (± 0.017). The allocator is thus not a restatement of the grid: here it extrapolates to an unsampled split that beats every allocation the grid sampled.

Table 2: The acquisition formula and its forward validation. Each recommended point (K^*, M^*) is trained (3 seeds, grid protocol) and compared with the best grid arm at that budget (`mostpeople`, M=125); Δ is the composite difference.

	formula's prescription		profile composite		
budget B	people K^*	segments/person M^*	at (K^*, M^*)	best grid arm	Δ
0.3M	7,769	39	0.529	0.493	+0.036
1M	11,238	89	0.588	0.586	+0.002
3M	15,740	191	0.618	0.619	−0.001
10M	22,769	439	0.640	0.644	−0.004

At the large end the prescription moves the other way: at 10M it matches `mostpeople`'s composite with 72% fewer people ($K^* = 22{,}769$ against 80,000). This does not contradict "more people wins": the grid's people-heavy end wins because its segment-heavy arms are starved of people, while past about 20,000 people extra headcount buys little at this budget (the 57,000 more in `mostpeople` add $+0.004$). The law predicts a slight loss past K^* where the grid still shows a slight gain, consistent with its weaker fit at the largest budget (Section 6.2); either way, do not under-recruit, and weigh headcount beyond K^* against its per-person cost (Appendix E.7).

7 Discussion and Conclusions

Three lines of evidence—a fixed-budget grid, orthogonal sweeps that separate the two scaling exponents, and identity measures on held-out people—agree that identity-contrastive PPG pretraining gains more from people than from segments per person. Solving the law for its optimum yields the paper's practical contribution: an allocator that converts a total segment budget into how many people to enroll and how many segments to record from each. Training its recommended points confirms it: at every budget the recommendation matches or beats the best allocation we sampled. The rule is budget-adaptive rather than uniformly "more people": *how* people-heavy is set by the budget.

Limitations and future work. The claim is scoped to identity-contrastive pretraining, where headcount enters the objective itself; recipes that regularize identity away rather than target it need not favor the people axis. Three limitations bound it further. The first is the single acquisition cohort, which binds in both directions: whether the ordering survives across devices and populations is open, so we frame the finding as a regularity of this setting rather than a universal law, and the cohort caps what we could probe—its most people-heavy arm nearly exhausts the people available, so the flat composite at the top of that axis says headcount stops paying within the range one cohort reaches, not that the people axis is spent. The second is that K is a count, not a guarantee of coverage: we say nothing about *which* people are recruited, and physiological corpora skew demographically (Bondarenko et al., 2025), so headcount buys diversity only if the added people differ. The third is the ruler: every scaling conclusion is read off demographic prediction, and the clinical endpoints are too small to confirm or refute it. Three directions follow: (i) external validation across cohorts, devices, and modalities such as ECG; (ii) replacing headcount with a measured effective diversity, to ask whether the law fits it more tightly than K; and (iii) whether generative modeling can expand the people axis at lower cost. Our result bears on this last direction: synthetic data that adds segments to existing people would feed the weaker of the two axes, and any such approach must also reckon with what training on generated data does to a model (Shumailov et al., 2024). Running through all three is one hypothesis: that the diversity of people, not the volume of signal per person, is the axis along which physiological representations are best grown.

AI USE STATEMENT

The research idea, the framing of the acquisition question, the decomposition of the data budget into people and segments, the choice of experimental design, and the paper's argument, interpretation of results, and conclusions are the authors' own. Generative AI tools were used as an execution aid under author direction and review. Specifically, we used them to help implement the pretraining and evaluation code, to clean and reformat data, to run the numerical fit of the scaling law under the functional form and fitting protocol specified by the authors, to produce figures, to search for and summarize related literature, and to draft and edit portions of the manuscript. The authors verified the output of each of these steps: code was reviewed and its results cross-checked against independent computations, the fitted exponents were subjected to the robustness and held-out extrapolation checks reported in this paper, and every citation was checked against its primary source for correct authorship, venue, and the claim it is used to support. We did not use generative AI tools to generate synthetic data, and this paper makes no formal mathematical claims requiring proof. We have reviewed all AI-assisted work and take full responsibility for the final content of this paper, including all text, claims, and figures produced with the aid of these tools.

REFERENCES

Salar Abbaspourazad, Oussama Elachqar, Andrew C. Miller, Saba Emrani, Udhyakumar Nallasamy, and Ian Shapiro. Large-scale training of foundation models for wearable biosignals. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=pC3WJHf51j.

Adrien Bardes, Jean Ponce, and Yann LeCun. VICReg: Variance-invariance-covariance regularization for self-supervised learning. In *The Tenth International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=xm6YD62D1Ub.

Tamay Besiroglu, Ege Erdil, Matthew Barnett, and Josh You. Chinchilla scaling: A replication attempt. *arXiv preprint arXiv:2404.10102*, 2024. URL https://arxiv.org/abs/2404.10102.

Maksym Bondarenko, Carlo Menon, and Mohamed Elgendi. Demographic bias in public remote photoplethysmography datasets. *npj Digital Medicine*, 8:593, 2025. doi: 10.1038/s41746-025-01973-9.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 1597–1607. PMLR, 2020. URL https://proceedings.mlr.press/v119/chen20j.html.

Xiaoli Chen, Rui Wang, Phyllis Zee, Pamela L. Lutsey, Sogol Javaheri, Carmela Alcántara, Chandra L. Jackson, Michelle A. Williams, and Susan Redline. Racial/ethnic differences in sleep disturbances: The Multi-Ethnic Study of Atherosclerosis (MESA). *Sleep*, 38(6):877–888, 2015. doi: 10.5665/sleep.4732.

Alan DenAdel, Madeline Hughes, Akshaya Thoutam, Anay Gupta, Andrew W. Navia, Nicolo Fusi, Srivatsan Raghavan, Peter S. Winter, Ava P. Amini, and Lorin Crawford. Evaluating the role of pretraining dataset size and diversity on single-cell foundation model performance. *Nature Methods*, 23:1447–1457, 2026. doi: 10.1038/s41592-026-03120-y. URL https://www.nature.com/articles/s41592-026-03120-y.

Eray Erturk, Fahad Kamran, Salar Abbaspourazad, Sean Jewell, Harsh Sharma, Yujie Li, Sinead Williamson, Nicholas J. Foti, and Joseph Futoma. Beyond sensor data: Foundation models of behavioral data from wearables improve health predictions. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 15516–15541. PMLR, 2025. URL https://proceedings.mlr.press/v267/erturk25a.html.

Sachin Goyal, Pratyush Maini, Zachary C. Lipton, Aditi Raghunathan, and J. Zico Kolter. Scaling laws for data filtering—data curation cannot be compute agnostic. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22702–22711,

2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Goyal_Scaling_Laws_for_Data_Filtering--_Data_Curation_cannot_be_Compute_CVPR_2024_paper.html.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. Bootstrap your own latent: A new approach to self-supervised learning. In *Advances in Neural Information Processing Systems*, volume 33, pp. 21271–21284, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/f3ada80d5c4ee70142b17b8192b2958e-Abstract.html.

Hasan Abed Al Kader Hammoud, Tuhin Das, Fabio Pizzati, Philip H. S. Torr, Adel Bibi, and Bernard Ghanem. On pretraining data diversity for self-supervised learning. In *Computer Vision – ECCV 2024*, pp. 54–71. Springer Nature Switzerland, 2024. doi: 10.1007/978-3-031-72992-8_4.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9729–9738, 2020. URL https://openaccess.thecvf.com/content_CVPR_2020/html/He_Momentum_Contrast_for_Unsupervised_Visual_Representation_Learning_CVPR_2020_paper.html.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16000–16009, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/He_Masked_Autoencoders_Are_Scalable_Vision_Learners_CVPR_2022_paper.html.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Thomas Hennigan, Eric Noland, Katherine Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karén Simonyan, Erich Elsen, Oriol Vinyals, Jack W. Rae, and Laurent Sifre. An empirical analysis of compute-optimal large language model training. In *Advances in Neural Information Processing Systems*, volume 35, pp. 30016–30030, 2022. doi: 10.52202/068431-2176. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/c1e2faff6f588870935f114ebe04a3e5-Abstract-Conference.html.

Linxing Preston Jiang, Shirui Chen, Emmanuel Tanumihardja, Xiaochuang Han, Weijia Shi, Eric Shea-Brown, and Rajesh P. N. Rao. Data heterogeneity limits the scaling effect of pretraining in neural data transformers. In *COLM 2025 Workshop on Large Language Modeling for Scientific Discovery (LM4Sci)*, 2025. URL https://openreview.net/forum?id=WiECjcxNw4.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020. URL https://arxiv.org/abs/2001.08361.

Simon A. Lee, Cyrus Tanade, Hao Zhou, Juhyeon Lee, Megha Thukral, Md Sazzad Hissain Khan, Baiying Lu, Migyeong Gwak, et al. HiMAE: Hierarchical masked autoencoders discover resolution-specific structure in wearable time series. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=iPAy5VpGQa.

Margaret Li, Sneha Kudugunta, and Luke Zettlemoyer. (Mis)Fitting: A survey of scaling laws. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=xI71dsS3o4.

Ilya Loshchilov and Frank Hutter. SGDR: Stochastic gradient descent with warm restarts. In *The Fifth International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=Skq89Scxx.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *The Seventh International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Niklas Muennighoff, Alexander M. Rush, Boaz Barak, Teven Le Scao, Nouamane Tazi, Aleksandra Piktus, Sampo Pyysalo, Thomas Wolf, and Colin A. Raffel. Scaling data-constrained language models. In *Advances in Neural Information Processing Systems*, volume 36, pp. 50358–50376, 2023. doi: 10.52202/075280-2191. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/9d89448b63ce1e2e8dc7af72c984c196-Abstract-Conference.html.

Girish Narayanswamy, Xin Liu, Kumar Ayush, Yuzhe Yang, Xuhai Xu, Shun Liao, Jake Garrison, Shyam Tailor, Jake Sunshine, Yun Liu, Tim Althoff, Shrikanth Narayanan, Pushmeet Kohli, Jiening Zhan, Mark Malhotra, Shwetak Patel, Samy Abdel-Ghaffar, and Daniel McDuff. Scaling wearable foundation models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=yb4QE6b22f.

Girish Narayanswamy, Maxwell A. Xu, A. Ali Heydari, Samy Abdel-Ghaffar, Marius Guerard, Kara Vaillancourt, Zhihan Zhang, Jake Garrison, Levi Albuquerque, Dimitris Spathis, Hong Yu, Hamid Palangi, Xuhai Xu, David G. T. Barrett, Joseph Breda, Jed McGiffin, Yubin Kim, Yuwei Zhang, Naghmeh Rezaei, Samuel Solomon, Karan Ahuja, Tim Althoff, Jake Sunshine, Ming-Zher Poh, Benjamin Yetton, Ari Winbush, Nicholas B. Allen, James M. Rehg, Isaac Galatzer-Levy, Yun Liu, John Hernandez, Anupam Pathak, Conor Heneghan, Yuzhe Yang, Ahmed A. Metwally, Pushmeet Kohli, Mark Malhotra, Shwetak Patel, Xin Liu, and Daniel McDuff. Towards a general intelligence and interface for wearable health data. *arXiv preprint arXiv:2605.22759*, 2026. URL https://arxiv.org/abs/2605.22759.

Guangkun Nie, Xiaocheng Fang, Gongzheng Tang, Yujie Xiao, Jun Li, Bo Liu, Hongyan Li, and Shenda Hong. AnyPPG: An ECG-guided PPG foundation model trained on over 100,000 hours of recordings for holistic health profiling. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2026. doi: 10.1145/3770855.3818911.

Arvind Pillai, Dimitris Spathis, Fahim Kawsar, and Mohammad Malekzadeh. PaPaGei: Open foundation models for optical physiological signals. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=kYwTmlq6Vn.

Alexandre Sablayrolles, Matthijs Douze, Cordelia Schmid, and Hervé Jégou. Spreading vectors for similarity search. In *The Seventh International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=SkGuG2R5tm.

Mithun Saha, Maxwell A. Xu, Wanting Mao, Sameer Neupane, James M. Rehg, and Santosh Kumar. Pulse-PPG: An open-source field-trained PPG foundation model for wearable applications across lab and field settings. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 9(3):126:1–126:35, 2025. doi: 10.1145/3749494.

Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross Anderson, and Yarin Gal. AI models collapse when trained on recursively generated data. *Nature*, 631(8022):755–759, 2024. doi: 10.1038/s41586-024-07566-y.

Ben Sorscher, Robert Geirhos, Shashank Shekhar, Surya Ganguli, and Ari Morcos. Beyond neural scaling laws: Beating power law scaling via data pruning. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/7b75da9b61eda40fa35453ee5d077df6-Abstract-Conference.html.

Mingxing Tan and Quoc Le. EfficientNet: Rethinking model scaling for convolutional neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 6105–6114. PMLR, 2019. URL https://proceedings.mlr.press/v97/tan19a.html.

James Truslow, Angela Spillane, Huiming Lin, Katherine Cyr, Adeeti Ullal, Edith Arnold, Ron Huang, Laura Rhodes, Jennifer Block, Jamie Stark, James Kretlow, Alexis L. Beatty, Andreas Werdich, Deepali Bankar, Matt Bianchi, Ian Shapiro, Jaime Villalpando, Sharon Ravindran, Irida Mance, Adam Phillips, John Earl, Rahul C. Deo, Sumbul A. Desai, and Calum A. MacRae. Understanding activity and physiology at scale: The Apple Heart & Movement Study. *npj Digital Medicine*, 7(1): 242, 2024. doi: 10.1038/s41746-024-01187-5.

Terry T. Um, Franz M. J. Pfister, Daniel Pichler, Satoshi Endo, Muriel Lang, Sandra Hirche, Urban Fietzek, and Dana Kulić. Data augmentation of wearable sensor data for Parkinson's disease monitoring using convolutional neural networks. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, pp. 216–220, 2017. doi: 10.1145/3136755.3136817.

Aäron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. URL https://arxiv.org/abs/1807.03748.

Weinan Wang, Pedram Mohseni, Kevin L. Kilgore, and Laleh Najafizadeh. PulseDB: A large, cleaned dataset based on MIMIC-III and VitalDB for benchmarking cuff-less blood pressure estimation methods. *Frontiers in Digital Health*, 4:1090854, 2023. doi: 10.3389/fdgth.2022.1090854.

Maxwell A. Xu, Girish Narayanswamy, Kumar Ayush, Dimitris Spathis, Shun Liao, Shyam A. Tailor, Ahmed Metwally, A. Ali Heydari, Yuwei Zhang, Jake Garrison, Samy Abdel-Ghaffar, Xuhai Xu, Ken Gu, Jacob Sunshine, Ming-Zher Poh, Yun Liu, Tim Althoff, Shrikanth Narayanan, Pushmeet Kohli, Mark Malhotra, Shwetak Patel, Yuzhe Yang, James M. Rehg, Xin Liu, and Daniel McDuff. LSM-2: Learning from incomplete wearable sensor data. *arXiv preprint arXiv:2506.05321*, 2025. URL https://arxiv.org/abs/2506.05321.

Yavuz Yarici, Kiran Kokilepersaud, Mohit Prabhushankar, and Ghassan AlRegib. Subject invariant contrastive learning for human activity recognition. In *2025 IEEE 35th International Workshop on Machine Learning for Signal Processing (MLSP)*, pp. 1–6, Istanbul, Turkey, 2025. IEEE. doi: 10.1109/MLSP62443.2025.11204249.

Hao Zhou, Simon A. Lee, Cyrus Tanade, Keum San Chun, Juhyeon Lee, Migyeong Gwak, Megha Thukral, Justin Sung, Eugene Hwang, Mehrab Bin Morshed, Li Zhu, Viswam Nathan, Md Mahbubur Rahman, Subramaniam Venkatraman, and Sharanya Arcot Desai. Physiology-aware masked cross-modal reconstruction for biosignal representation learning. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2605.00973.

APPENDIX

Appendix A describes the data, the implementation of a pretraining run, and the controls that make runs comparable. Appendices B and C justify the two choices fixed before the scaling study: the pretraining objective and the evaluation metric. Appendix D gives further evidence for the identity analysis of Section 5.3, and Appendix E details how the scaling law is fit, tested for robustness, and validated.

A SETUP: WHAT A RUN IS, AND WHY RUNS ARE COMPARABLE

No comparison across budgets means anything unless every run is produced the same way and scored the same way. This appendix records both.

A.1 THE PRETRAINING COHORT AND THE DOWNSTREAM PANEL

Signal and cohort. All pretraining data comes from a single wrist-worn consumer smartwatch, collected under a large-scale health research program operated by the device manufacturer, whose name is withheld for anonymous review. The green-channel PPG stream is sampled at $50\,\text{Hz}$ and cut into non-overlapping $30\,\text{s}$ windows, giving the 1500-sample segments the encoder consumes (Appendix A.2). The full pool is $114{,}369$ participants and 53.6M segments. Because the recordings come from free-living wear rather than a clinical session, a participant's segments are spread over days to months, so segments-per-person M buys *longitudinal* coverage of one person and headcount K buys *cross-sectional* coverage of the population—the two axes this paper puts against each other.

Ethics and governance. The study protocol was approved by the institutional ethics committee of an academic medical center, and data was de-identified before it reached the modeling pipeline. The committee's name and the approval identifier are withheld for anonymous review and will be stated in the camera-ready version.

Downstream panel. The $5{,}000$ evaluation participants of Table 1 are drawn from the same cohort and removed from every pretraining pool. Profile targets (age, BMI, sex) are self-reported at enrollment; the blood-pressure, hypertension, OSA, and AFib subsets are the participants for whom a corresponding reference measurement or clinical label is available, which is why they are one to two orders of magnitude smaller than the profile set—the size asymmetry that Appendix C.2 shows is decisive for which task can carry a scaling curve. The two blood-pressure subsets differ in what the label requires. The SBP/DBP regressions are point-to-point, each PPG recording paired with a cuff reading from the same session, which 200 evaluation participants have. The hypertension classifier needs only a person-level label: for those same 200 participants it is the standard threshold on that cuff reading (SBP ≥ 140 or DBP $\geq 90\,\text{mmHg}$), and a further 310 participants contribute a known hypertension status with no time-aligned reading, for 510 in total. The blood-pressure cohort is thus nested inside the hypertension cohort, and the two endpoints are not independent readouts. The profile composite treats sex as a regression target so that all three profile terms share the one normalized-MAE skill scale of Equation 3; as for age and BMI, its baseline is a constant predictor, so a composite of 0 still means no better than a constant guess.

A.2 WHAT A RUN IS: RECIPE, BACKBONE, AND PROBE

All runs share the recipe and backbone of Section 3.1 and the setup of Section 4.3; only the data allocation (K, M), and (where a run is an ablation) the single factor under study, differs.

Backbone. The encoder is a 1D-EfficientNet operating on a single-channel PPG segment of length 1500. It is a stem convolution (kernel 7) followed by seven stages of mobile inverted-bottleneck blocks (MBConv, expansion 6 except the first stage, depthwise kernels 3 or 5, squeeze-and-excitation ratio 0.25, SiLU activations), then a 1×1 head convolution and global average pooling to a 256-dimensional embedding. The base size has 16 blocks and 3.12M parameters. The three capacities of Appendix D.2 are compound-scaled from this base by $(\text{width}, \text{depth})$ multipliers—small $(0.5, 0.5)$, base $(1.0, 1.0)$, large $(2.0, 1.5)$—giving 0.49M / 3.12M / 22.8M parameters; channel widths are rounded to multiples

of 8 in the standard EfficientNet manner. During pretraining the 256-d embedding is passed through a projection head ($256 \to 1024 \to 128$, batch-norm, SiLU, dropout 0.1) on which the contrastive loss is computed; every downstream probe reads the frozen 256-d encoder embedding directly.

Augmentations. Each segment of a positive pair is z-normalized and then passed through four stochastic PPG augmentations, each applied independently with its own probability: random cut-out (p=0.2, zeroing a 20% contiguous span), magnitude warping (p=0.15, spline with σ=0.2, 4 knots), additive Gaussian noise (p=0.15, σ=0.1), and time warping (p=0.05, σ=0.1).

Pretraining hyperparameters. All arms train for 10 epochs at global batch size 4096 over 4 NVIDIA A800 (80 GB) GPUs under DDP (per-GPU batch 1024, InfoNCE denominator all-gathered across ranks), with AdamW (learning rate 5×10^{-4} cosine-decayed to 1×10^{-5}, weight decay 10^{-2}), InfoNCE temperature 0.07, KoLeo weight 0.1, VICReg variance–covariance regularization, and an EMA teacher (momentum 0.99). These are held fixed across every budget arm so that only the allocation (K, M) varies. Unless stated otherwise runs use seed 42, and every scaling arm is repeated over 3 seeds (42/43/44) with error bars reporting seed standard deviations.

Downstream probe. The encoder is frozen and a linear probe is fit on its embeddings: ridge regression for continuous targets (profile age/BMI/sex, blood pressure) and balanced logistic regression for the clinical classifiers. Profile features are mean-pooled per person before probing. The 5,000 evaluation people are held out of every pretraining pool (Appendix A.3), so no probe is ever fit on a training subject.

A.3 Why runs are comparable: single-factor control

All cross-budget / cross-allocation comparisons in the paper are kept single-factor to rule out comparison noise, via four aligned choices:

- **Fixed evaluation set**: the 5,000 evaluation people are removed from the data pool once at the start; no budget's training pool contains them, giving zero evaluation leakage. All budgets share the same evaluation set, so cross-budget comparisons introduce no "different evaluation set" noise.
- **Nested sampling**: the people of a smaller budget are a subset of those of a larger budget, so cross-budget comparisons introduce no "different batch of people" noise—budget is the only variable that changes.
- **Seed pairing**: same seeds, same data-pool split, same recipe; only the single factor under study (K, M, budget, or algorithm) varies, everything else aligned; error bars come from 3 seeds.
- **Matched compute along a fixed-budget diagonal**: the epoch count is fixed, so two allocations of the same budget $B = K \cdot M$ take the same number of gradient steps, and the K-versus-M contrast is not a contrast in compute. Headcount does also enter the difficulty of the in-batch task—segments are drawn at random, so an arm with few people fills its 4096 batch slots from few identities (at K=150, at most 149 distinct people to be contrasted against). That is by design rather than a confound, identity being the discriminative target, and the held-out identity measures of Section 5.3 are scored on the trained encoder, where batch composition no longer applies.

This guarantees that the fixed-budget diagonal, the three-source merge of the law fit, and the objective-selection comparison are all single-factor controlled.

B The objective we scale

The scaling study fixes one pretraining objective before varying anything else. We choose it by comparing 7 representative SSL objectives under a shared protocol (Table 3): all are pretrained on a fixed 1M-segment pool (K=8000 people, M=125 segments each) and evaluated by the same frozen linear probe, so the comparison is single-factor in the sense of Appendix A.3.

Table 3: Objective selection under a shared protocol. pid = person-identity discrimination; dropping the VICReg term from the top row recovers the base recipe. Overall AUC is a cross-task weighted classification AUC over the clinical endpoints. All objectives are trained from scratch on the same pool; the PaPaGei row uses its sVRI-based objective, not the released model.

objective	discrimination	composite	overall AUC	SER
identity-contrastive (base recipe)	pid + VICReg	0.587	0.902	109
w/o VICReg (Abbaspourazad et al., 2024)	pid	0.580	0.896	89
Masked autoencoder (He et al., 2022)	none	0.406	0.874	—
SimCLR (Chen et al., 2020)	weakened pid	0.387	0.815	63
BYOL (Grill et al., 2020)	no negatives	0.384	0.825	1.97
PaPaGei (Pillai et al., 2025)	sVRI-bin for pid	0.322	0.857	4.42
VICReg-only (Bardes et al., 2022)	none (pure reg.)	0.208	0.741	128

The table falls into two tiers. The two identity-contrastive rows sit at 0.58–0.59; every other objective sits at 0.41 or below. Below that boundary the ordering does not follow how far the target is from identity: masked reconstruction, which has no identity target at all, is the strongest of the lower tier (0.406), ahead of the weakened (SimCLR 0.387, BYOL 0.384) and proxied (PaPaGei 0.322) forms, with VICReg-only last (0.208). What the comparison establishes is therefore the tier boundary, not an ordering within the lower tier. SER (smooth effective rank) is a collapse diagnostic, not a quality metric: VICReg-only has the highest SER yet the worst composite, so keeping feature dimensions "alive" is not the same as making them transferable. Adding the VICReg term to the base recipe is what raises the embedding's effective rank (SER 89 → 109; Section 3.1) at no cost on the composite (0.580 → 0.587).

C THE RULER WE MEASURE ON

Section 5.1 fixes the profile composite as the sole scaling ruler on *a priori* grounds, and two things could unseat it: a better ruler, or a rival task that reads the same conclusion differently. The natural better ruler is the model's own pretraining loss—it moves the same way, but measures the wrong thing (Appendix C.1). The rival tasks are too small to resolve it (Appendix C.2).

C.1 THE VALID-LOSS DIAGNOSTIC

The SSL model's own valid loss measures optimization of the pretraining objective, not downstream transfer. Because the law guides acquisition, it is fit on the downstream value of the representation (Section 6); valid loss serves as a consistency check.

Valid loss responds to both budget and allocation. Along the K-sweep (fixed M=250) and the M-sweep (fixed K=10000), valid loss falls monotonically as either axis grows (K-sweep: 5.06 → 3.41 as K goes 1000 → 50000; M-sweep: 6.26 → 3.63 as M goes 10 → 1000). At a fixed budget it favors people-heavy splits, in the same direction as the composite: from `mostdata` to `mostpeople` it falls by 1.08/1.23/0.93/0.50 at 0.3M / 1M / 3M / 10M (Figure 5a). The conclusion "recruit more people" is thus not a preference of one particular metric.

The agreement holds across all budgets. Pooling all 31 scaling points (the grid and both orthogonal sweeps), valid loss and composite move together almost perfectly (Spearman $\rho = -0.98$; Figure 5b): a lower valid loss predicts a higher composite regardless of how the budget was allocated. The composite is therefore not an idiosyncratic ruler—only one we can also read at the resolution the scaling curves need.

C.2 WHY NO OTHER DOWNSTREAM TASK CAN CARRY A SCALING CURVE

The rest of the downstream panel is too small to carry a scaling curve: blood pressure moves with the composite, and the clinical classifiers show no consistent trend.

Figure 6 reads every downstream task along the K-sweep. Blood-pressure MAE on the 200-person BP cohort falls with K in the same direction as the composite (DBP monotonically, 6.36 → 5.39;

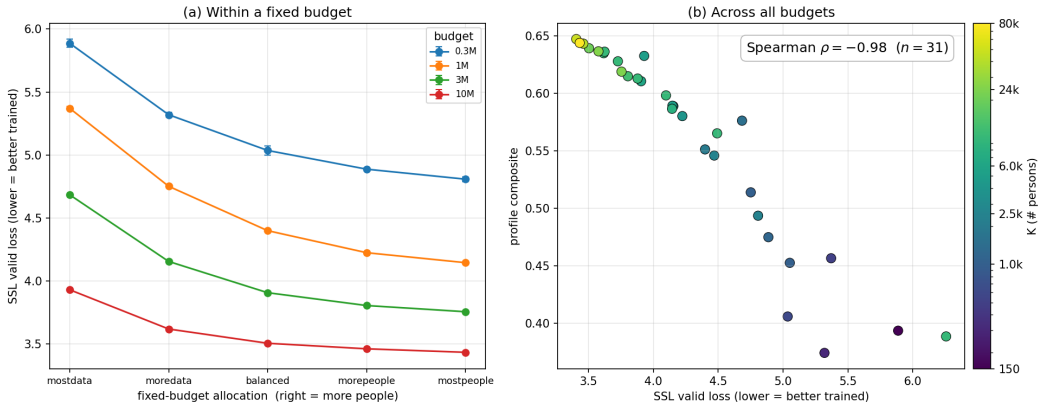

Figure 5: SSL valid loss agrees with the profile composite. **(a)** Valid loss along the fixed-budget diagonal ($K \cdot M$ constant), one curve per budget: within every budget it falls as the allocation shifts toward more people, the same ordering the composite shows in Figure 2. **(b)** Composite against valid loss across all 31 scaling points (grid + both orthogonal sweeps; color = headcount K): with the budget not held fixed, the two still trace a near-monotone band (Spearman $\rho = -0.98$). The replicate configuration enters once per source here, whereas the law's fit merges its two training paths into a single observation—hence 31 points against the 30 of Section 6.1.

SBP $8.54 \rightarrow 7.59$, with a plateau over $K \in [5000, 25000]$), so the benefit of headcount leaves a trace on an independent continuous task and is not an artifact of the composite metric; both targets, however, carry sizeable seed variance. The clinical classifiers, with only tens to a few hundred labeled subjects, show no consistent trend in K and are too small to resolve one either way.

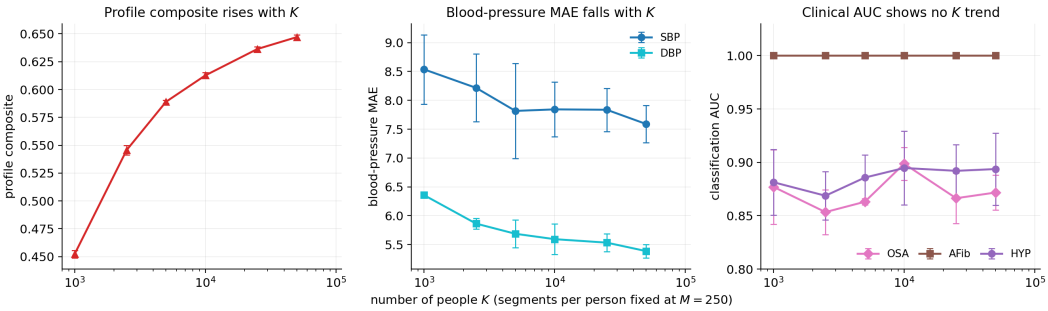

Figure 6: Downstream tasks along the K-sweep (fixed M=250, 3-seed mean $\pm$ standard deviation). **Left:** the profile composite rises monotonically with K and is the only downstream signal stable enough to read a clean trend from. **Middle:** blood-pressure MAE falls with K in the same direction, but both SBP and DBP carry large seed variance. **Right:** the small-sample classifiers show no K trend—AFib is saturated at AUC 1.0, and OSA and hypertension (HYP) jitter inside overlapping error bars.

Table 4 makes the same cut along the allocation rather than along K, reading out the two extreme allocations of the largest budget on every downstream task. Only the profile composite, on 5,000 held-out people, is stable enough to carry the scaling curves.

Table 4: Endpoint check at the largest budget (10M): the most people-heavy vs. the most segment-heavy allocation on every downstream task (3-seed means; composite and AUC higher is better, BP MAE lower is better). The people-heavy allocation is numerically ahead on the composite, both blood-pressure targets, and OSA, and behind on hypertension (0.851 vs. 0.859); AFib is saturated. At 10M every allocation gap is small (Section 5.2), and none of the clinical endpoints has the labeled sample size to resolve it. HYP: hypertension.

allocation	composite	SBP MAE	DBP MAE	HYP AUC	AFib AUC	OSA AUC
`mostpeople`	0.644	8.00	5.35	0.851	1.000	0.876
`mostdata`	0.632	8.05	5.47	0.859	1.000	0.867

D IDENTITY STRUCTURE: WHAT HEADCOUNT CHANGES

Section 5.3 shows with the held-out identity measures of Figure 3 that headcount changes what the representation encodes about identity. Two further measurements agree—one qualitative, one from the model side.

D.1 IDENTITY STRUCTURE, SEEN DIRECTLY

Figure 7 is the qualitative companion to the R@1 and silhouette trends of Figure 3.

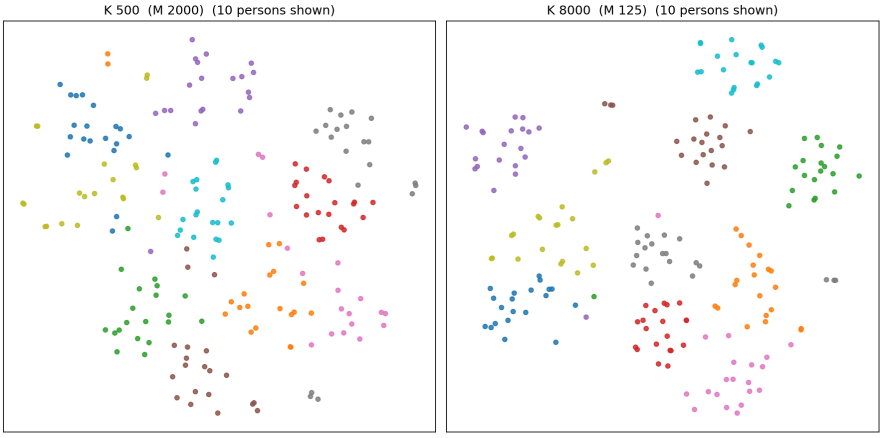

Figure 7: By-person t-SNE of a fixed set of held-out subjects, embedded by two encoders trained on the same 1M-segment budget at opposite allocations: `mostdata` (K=500, M=2000, left) vs. `mostpeople` (K=8000, M=125, right), same seed. Colors are subject identities. With the budget held fixed, more people yields tighter, better-separated per-person clusters on subjects held out of both training pools.

D.2 PERSON×MODEL CAPACITY INTERACTION

If headcount supplies identity diversity rather than more signal per person (Section 5.3), a larger model's advantage should be sustained as K grows but not as M grows: more people are more identities a larger model can exploit, whereas more segments-per-person are only more views of the same identities. We test this at three capacities of the same 1D-EfficientNet backbone, scaled only in depth and width (small 0.49M, base 3.12M, large 22.8M), along the three cuts that touch the claim directly—the two orthogonal sweeps and the fixed-budget diagonal (Figure 8).

The allocation result is the robust part. With $K \cdot M = 1$M held constant, all three capacities decline together as the split skews toward segments (a 0.10–0.15 drop from `mostpeople` to `mostdata`; Figure 8a), so people-dominance is not an artifact of one backbone but holds across a nearly 50× range of parameter counts.

The two sweeps bear on the interaction, and their evidence is weaker. At the low end they look alike: extra capacity is wasted while data is scarce along either axis (K=1000 on the K-sweep, M=10

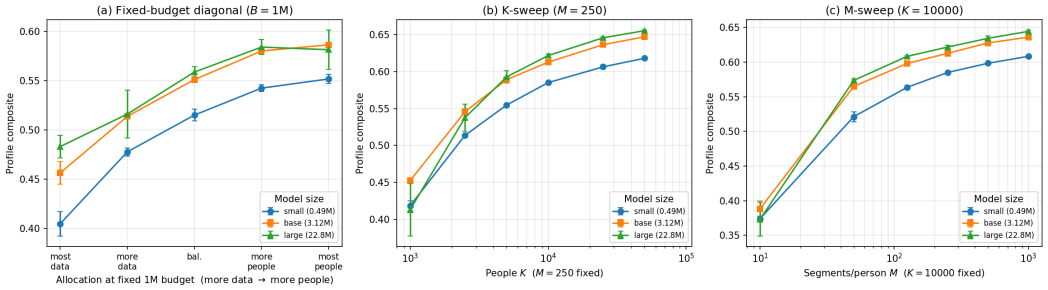

Figure 8: Person×model interaction at three capacities (3 seeds each); moving right means more of the scarce resource. **(a) Fixed-budget diagonal** ($K \cdot M = 1$M): all three capacities rise together toward the people-heavy end. **(b) K-sweep** (M=250 fixed): once $K \geq 5000$ the curves settle into size order and the gap holds (large − small ≈ 0.038). **(c) M-sweep** (K=10000 fixed): past M=10 the gap does not grow but narrows (large − small $= 0.052 \rightarrow 0.036$ from M=50 to M=1000).

on the M-sweep), and the three capacities fall into size order only from K=5000 and M=50 on. Above it they part modestly. Along K the large−small gap, once open, holds at ≈ 0.038 out to K=50000 (Figure 8b); along M it does not grow but narrows, from 0.052 at M=50 to 0.036 at M=1000 (Figure 8c). More people thus preserve what a larger model gains, whereas more segments per person slightly erode it—the direction a diversity signal predicts, though the sweeps alone do not establish a clean split between a complementary and an orthogonal axis.

E THE SCALING LAW: IDENTIFICATION, ROBUSTNESS, AND VALIDITY

This appendix gives the grid measurements the law is fit on, and reports how it is estimated, why $\alpha > \beta$ is identified rather than an artifact of the estimator, how far it survives the noisiest arms and a relaxation of its own functional form, whether it predicts an unseen budget, and where the acquisition formula's recommendation sits relative to the region actually measured.

E.1 THE FIXED-BUDGET GRID IN NUMBERS

Table 5 gives the measured values behind Figure 2: the profile composite at every budget × allocation cell of the diagonal grid, which together with the two orthogonal sweeps is what the law below is fit on.

Table 5: Profile composite by budget × allocation (means ± std, 3 seeds). Allocations run from `mostpeople` (left) to `mostdata` (right); M is segments-per-person and the headcount is $K = B/M$.

budget B	mostpeople M=125	morepeople M=250	balanced M=500	moredata M=1000	mostdata M=2000
0.3M	0.493 ± 0.007	0.475 ± 0.003	0.406 ± 0.017	0.374 ± 0.003	0.393 ± 0.004
1M	0.586 ± 0.002	0.580 ± 0.004	0.551 ± 0.003	0.514 ± 0.001	0.456 ± 0.014
3M	0.619 ± 0.001	0.615 ± 0.003	0.610 ± 0.001	0.589 ± 0.001	0.576 ± 0.004
10M	0.644 ± 0.002	0.643 ± 0.000	0.639 ± 0.000	0.635 ± 0.004	0.632 ± 0.002

E.2 FITTING PROCEDURE AND THE THREE-SOURCE MERGE

Fitting procedure. The separable law $L(K, M) = E + A_K K^{-\alpha} + A_M M^{-\beta}$ is fit by weighted nonlinear least squares (SciPy `curve_fit`, trust-region with box constraints). Each aggregated (K, M) point is weighted by $1/\sigma^2$, where σ is the 3-seed standard deviation floored at the median multi-seed standard deviation to keep zero-variance points from dominating, and the two high-leverage 0.3M edge arms have their σ inflated a further 5× under the main policy (see below). Inputs are rescaled (k=$K/10^4$, m=$M/10^2$) so the coefficients are $O(1)$ and the Jacobian is well-conditioned. The floor E is left free (bounded only below the best observed error $L_{\min} \approx 0.354$,

which is never binding): it settles at $\hat{E} = 0.13$. Because floor and exponents trade off against each other in any additive power law, E is profiled rather than trusted; that check is reported with the other tests of the functional form (Appendix E.5). The fit is run from 32 random restarts and the minimum weighted-SSR solution is kept; the reported solution is reached by 100% of restarts (best-basin rate).

Three-source merge check (calibration gate). Before merging grid diagonal + K-sweep + M-sweep, a consistency check is run, whose core is the exact replicate point (K=10000, M=1000): it is simultaneously the grid's 10M `moredata` arm and the M-sweep's M=1000 arm. The same configuration via two independent training paths, after merging, gives composite $= 0.635 \pm 0.004$ ($n_{\text{seed}} = 6$), with between-source gap $|\Delta| = 0.0013 \ll$ seed noise $\Rightarrow$ the three sources lie on the same surface and merging is safe. The M=250 and K=10000 shared lines are also consistent at the crossover.

E.3 Joint fit vs. 1-D exponents: where the asymmetry lives

Fitting each sweep alone gives nearly equal exponents, which appears to deny the asymmetry of Section 6.1. It does not (Figure 9); the two estimates answer different questions:

- **Joint planar fit** (both lines share one irreducible error E; each line's own asymptote adds the other axis's fixed term): $\alpha = 0.391 > \beta = 0.173$. The asymmetry lives in the exponents, which is where it is scale-invariant: a decade of people removes $1-10^{-\alpha} \approx 59\%$ of its own reducible term against 33% for a decade of segments, so people buy down their share of the error nearly twice as fast per decade. (We deliberately do not compare the coefficients A_K, A_M directly: because the two terms carry different units, their ratio depends on how K and M are rescaled—it is 8× in raw units but below 1 in the fit's internal k=$K/10^4$, m=$M/10^2$ coordinates—so it is not a meaningful measure of dominance.)
- **Independent 1-D power laws** (each line with its own floating E): $\hat{\alpha} = 0.647 \pm 0.029$ (K-sweep), $\hat{\beta} = 0.650 \pm 0.054$ (M-sweep), nearly equal. No contradiction: the 1-D fit assigns all curvature to the exponent; in the joint fit E is shared and each line's asymptote is tied to the other axis, and then the two exponents separate ($\alpha = 0.391 > \beta = 0.173$). What we claim is this ordering under the joint fit, not a particular value of α.

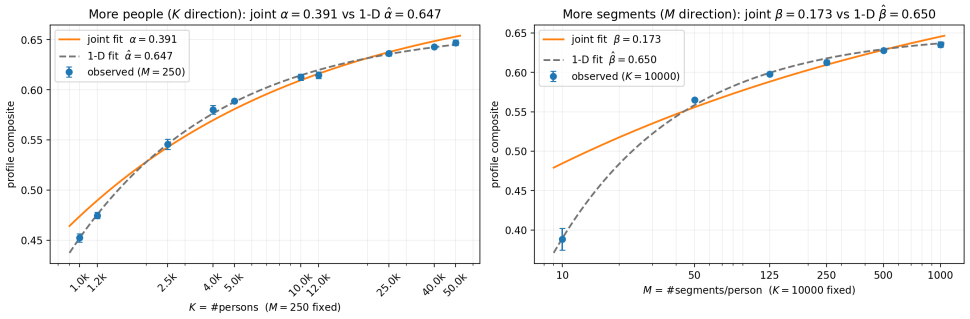

Figure 9: Joint fit (solid, one shared irreducible error E) vs. 1-D independent fits (dashed, each its own floor) along the two sweeps. A single line cannot separate floor from exponent, and the 1-D exponents come out nearly equal ($\hat{\alpha} = 0.647$, $\hat{\beta} = 0.650$); the joint fit, sharing E, separates them, $\alpha = 0.391 > \beta = 0.173$.

E.4 Edge-policy ablation

Under keep / downweight / drop of the two high-leverage 0.3M edge arms (`moredata`, 300 people; `mostdata`, 150 people), $\alpha > \beta$ holds without exception (Table 6), best-basin is 100% throughout, weighted $R^2 > 0.95$ throughout. We summarize each policy by the scale-invariant per-decade ratio $(1-10^{-\alpha})/(1-10^{-\beta})$—how much faster people buy down reducible error than segments—rather than the raw coefficient ratio A_K/A_M, which is not unit-invariant (Appendix E.3). Removing the noisy arms steepens both exponents together, so the per-decade ratio stays near 1.6–2.0× throughout. The main text reports the downweight point estimate, with {keep, drop} as robustness bounds.

Table 6: Edge-policy ablation: α (people) $>$ β (segments) under all three policies. "K/M per decade" is the scale-invariant ratio $(1-10^{-\alpha})/(1-10^{-\beta})$; R^2 is weighted; basin is the multi-restart best-basin rate.

policy	α	β	K/M per decade	R^2	basin
keep (full weight)	0.274 ± 0.009	0.118 ± 0.037	$1.97\times$	0.958	100%
downweight ($\sigma \times 5$, main)	$\mathbf{0.391 \pm 0.016}$	$\mathbf{0.173 \pm 0.037}$	$1.81\times$	0.963	100%
drop (remove both arms)	0.484 ± 0.022	0.228 ± 0.037	$1.65\times$	0.970	100%

E.5 Separability, the cross term, and floor robustness

The separable law of Equation 2 is a first-order working model (Section 3.2). Here we test its two load-bearing assumptions—that K and M are genuinely distinct axes, and that they combine additively—and show that relaxing the second only strengthens people-dominance.

People and segments are genuinely distinct axes. A skeptic could ask whether K and M matter only through their product $B = KM$, in which case the whole people-vs-segments framing collapses. They do not. We fit three models to the same 30 aggregated points under the same weighting: a *budget-only* model $L = E + A_B(KM)^{-\delta}$ (3 parameters), the *separable* model (5), and an *interaction* model that adds a cross term,

$$L(K, M) = E + A_K K^{-\alpha} + A_M M^{-\beta} + A_{KM}(KM)^{-\gamma}. \tag{6}$$

The separable model clearly beats budget-only ($\Delta\text{BIC} = 687$; leave-one-out CV MAE 0.019 vs. 0.027), so headcount and segments-per-person are not interchangeable through their product—the distinction the paper rests on is real (Table 7).

The cross term is budget-level, and it sharpens people-dominance. The interaction model in turn beats the separable one ($\Delta\text{BIC} = 178$), so the separable form misses some structure. The added term is best read as a budget-level correction rather than a K–M interaction in the usual sense: $A_{KM}(KM)^{-\gamma}$ depends on K and M only through their product, so it is constant along any fixed-budget diagonal and cannot favor one allocation over another. With $A_{KM} = +0.012 > 0$ and a steep $\gamma = 1.16$, it is an extra reducible error that is largest at small budgets and decays quickly as the budget grows. Once that budget-level curvature is absorbed separately, the segments axis has almost no slope of its own left: β collapses from 0.173 (separable) to 0.052 (interaction), while the people exponent is unchanged ($\alpha = 0.391 \to 0.395$). Much of what the separable fit attributes to segments-per-person is thus captured by budget growth alone; the standalone value of recording more per person is small. The people-vs-segments gap therefore widens under the better-fitting model ($\alpha - \beta$: $0.22 \to 0.34$), and the separable fit we report in the main text is the conservative reading of the ordering. Whether the cross term is real signal rather than the flexibility of two extra parameters is tested out-of-budget in Appendix E.6.

Table 7: Model selection on the 30 scaling points (weighted fit; weighted R^2 uses the same $w = 1/\sigma^2$ weighting as the main-text fit, so the separable row matches the $R^2 = 0.963$ of Section 6.1; BIC uses the weighted SSR as $-2\log\mathcal{L}$; CV MAE is leave-one-point-out, in composite units). Budget-only is rejected (people and segments are distinct axes), and the budget-level cross term is favored over pure separability.

model	params	weighted R^2	BIC	CV MAE
budget-only $E + A_B(KM)^{-\delta}$	3	0.856	947.3	0.0268
separable (Eq. 2)	5	0.963	260.5	0.0191
interaction (Eq. 6)	7	**0.991**	**82.2**	**0.0108**

The direction holds under both models. Solving each fitted model for the budget-optimal split points the same way: at every budget both the separable and the interaction model place the optimum at a more people-heavy split than the balanced grid arm, and M^*/K^* rises monotonically with B in both. The interaction model is in fact more people-heavy (e.g. at 0.3M it recommends $K^* \approx 25{,}000$ vs. the separable model's 7,769). People-first allocation thus does not depend on the separability assumption. We forward-validate only the separable recommendation (Table 2); since even the

conservative separable split already matches or beats the best measured arm at every budget, the more people-heavy interaction split is not needed to make the case.

Floor robustness. Because floor and exponents are correlated in any additive power law, we check that $\alpha > \beta$ is not manufactured by where the floor sits. Fixing E and refitting the remaining parameters across $E \in [0.20, 0.354]$, the ordering $\alpha > \beta$ holds for all $E \lesssim 0.27$, breaking down only as E approaches the best observed error $L_{\min} \approx 0.354$ and the reducible error vanishes. The free-fit value $\hat{E} = 0.13$ lies well below that boundary, and the interaction model drives the floor to ≈ 0 without disturbing $\alpha > \beta$—both sit in the region where the ordering holds. The ordering therefore holds for every floor up to about 0.27 and is not an artifact of a hand-set bound.

E.6 Held-out budget extrapolation

A law must extrapolate, not merely fit, so we hold out an entire budget (all five of its allocation arms) and predict it from the rest, for one interpolated budget (3M) and one extrapolated budget (10M). Two protocols differ in what they train on. *Budgets-only* excludes the two orthogonal sweeps, asking what the diagonal grid alone predicts about an unseen budget; *budgets-plus-sweeps* restores them, the only setting in which the separable and interaction models can be compared on one footing. Because the training sets differ, the same held-out budget receives different numbers under the two, and the protocols are reported separately.

Budgets-only: what the grid alone predicts. Holding out 3M gives MAE $= 0.0075$, $\rho = 1.00$, recommended arm hit; holding out 10M gives MAE $= 0.0119$, $\rho = 0.70$, and a systematic optimistic bias of about 0.01 (Figure 10). Interpolation is reliable; extrapolation past the largest measured budget is best read as an optimistic upper bound.

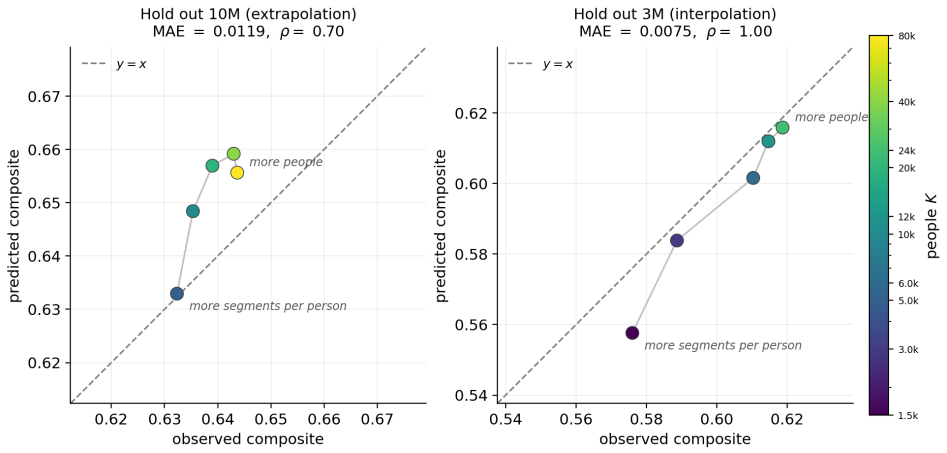

Figure 10: Leave-one-budget-out under the budgets-only protocol. Each point is one allocation arm of the held-out budget (color = headcount K), predicted from the other budgets vs. measured; points on $y{=}x$ are predicted exactly. **Left:** extrapolating past the largest budget (hold out 10M) sits above $y{=}x$—a mild optimistic bias. **Right:** interpolation (hold out 3M) is near-perfect ($\rho = 1.00$, arm order exact).

Budgets plus sweeps: the cross term is signal, not overfitting. With 7 parameters on 30 points, the interaction model's CV edge could be mere flexibility. Under this protocol, far stricter than the leave-one-point-out CV of Table 7, it instead extrapolates better (Table 8): holding out 10M it reaches MAE 0.0036 and $\rho = 0.90$ versus the separable model's 0.0104 and 0.30, and it removes the separable model's $+0.010$ optimistic bias almost entirely ($+0.0001$). An overfit term would degrade out-of-budget; this one corrects the separable form's extrapolation error, consistent with a real budget-level effect rather than noise.

Table 8: Leave-one-budget-out extrapolation under the budgets-plus-sweeps protocol: a held-out budget's five arms predicted from the remaining budgets together with both orthogonal sweeps. Bias > 0 means optimistic (predicted composite too high).

held-out	model	MAE	bias	Spearman ρ	best arm
3M (interp.)	separable	0.0066	-0.0059	0.90	no
	interaction	**0.0045**	-0.0006	**1.00**	yes
10M (extrap.)	separable	0.0104	$+0.0102$	0.30	no
	interaction	**0.0036**	$+$**0.0001**	**0.90**	no

E.7 Validity envelope of the acquisition formula

The allocator of Section 6.3 is the law solved for its own optimum, so it can in principle name a (K^*, M^*) the experiments never visited. It does not stray far. Every recommended coordinate falls inside a measured range: K^* runs from 7,769 to 22,769, interior to the K-sweep's $K \in [1000, 50000]$, and M^* runs from 39 to 439, interior to the M-sweep's $M \in [10, 1000]$. Each axis is therefore interpolated, never extrapolated.

What is extrapolated is their combination at a fixed budget. Each diagonal was measured over $M \in [125, 2000]$, so its people-heavy end sits at $K = B/125$: that is K=2400 at 0.3M and K=8000 at 1M, both below the recommended K^* of 7,769 and 11,238. At these two budgets the allocator therefore names a split more people-heavy than any arm measured at that budget, supported by the two sweeps rather than by the diagonal itself. At 3M and 10M the diagonal reaches K=24,000 and K=80,000, so the recommendation falls inside it and no such gap arises. The direction is safe either way—every measured diagonal peaks at its people-heavy end (Section 5.2). We resolve the small-budget gap directly by training the recommended points (Table 2): at 0.3M and 1M the split more people-heavy than any measured arm is not merely inferred from the sweeps but trained, and it matches or beats the best arm at that budget, by $+0.036$ composite at 0.3M.