Knowledge Transfer Scaling Laws for 3D Medical Imaging

Ho Hin Lee* Vanderbilt University
Dongna Du Zhejiang University
Chu Wang McGill University
Yuankai Huo Vanderbilt University
Shi Gu Zhejiang University
James C. Gee University of Pennsylvania
Yifan Wu University of Pennsylvania

Abstract

Vision foundation models are increasingly moving beyond 2D to volumetric domains such as 3D medical imaging, where unified pretraining across different imaging modalities (i.e. CT, MRI, and PET) could provide foundational models for diverse clinical tasks. However, training such models requires mixing heterogeneous imaging domains, and current mixture strategies remain largely heuristic. In this work, we observe that different medical imaging domains scale at variable rates during pretraining, and knowledge transfer between domains is strongly asymmetric: training on one domain can substantially improve another, but the reverse may be much weaker. Interestingly, both MAE reconstruction loss and cross-domain transfer follow predictable power-law trends with domain-specific behaviors. Motivated by these findings, we formulate data allocation as a **scaling-law optimization problem**. The derived allocations reveal an interpretable **hub-and-island** structure: highly transferable domains emerge as *hubs* that benefit many others and deserve strategic allocation, while isolated domains act as *islands* requiring direct investment. Empirically, Data-Proportional sampling incurs 4–58% higher reconstruction loss than transfer-aware allocation across the evaluated budgets, with the largest gap at $50K$ volume draws; predictions generalize to a held-out budget with $r{=}0.989$. Downstream validation on disease classification and organ/lesion segmentation further confirms that the derived transfer-aware mixtures provide stronger pretrained representations for clinical 3D medical imaging tasks.

1 Introduction

Clinical imaging spans CT, MRI, and PET, each with distinct acquisition physics, resolution, and clinical signal, yet the fundamental skills required to interpret them (volumetric reasoning, structural context, cross-anatomy knowledge) are shared. A **unified** 3D medical foundation model that learns reusable representations across modalities and anatomies could support diverse clinical tasks, from organ/lesion segmentation to abnormality discovery and disease monitoring from a single pretrained initialization [34, 45, 37, 38]. However, training such a unified model requires large-scale heterogeneous volumetric datasets with diversity, but while architectures and self-supervised objectives have received substantial attention, a more fundamental question remains open: **how much of each imaging domain should be used under a fixed volume-draw budget?**

In natural language processing, scaling laws [22, 17] have enabled principled data mixture optimization by modeling how performance scales with compute and data [48, 18, 35, 43]. However, these formulations assume relatively homogeneous data sources or shared scaling behavior, and have seen

*Corresponding author: ho.hin.lee@vanderbilt.edu

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

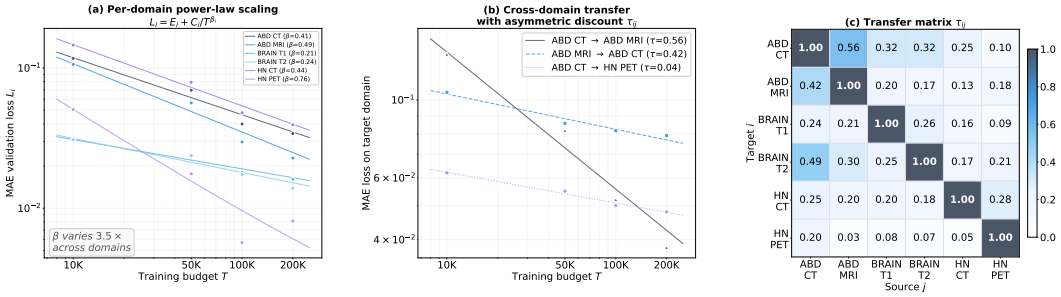

Figure 1: Observations motivating the transfer-aware scaling law. **(a)** Per-domain MAE loss follows $L_i = E_i + C_i/T^{\beta_i}$ with domain-specific exponents varying $3.5\times$ ($\beta = 0.21$ for BRAIN T1 to $\beta = 0.76$ for HN PET), indicating that domains benefit from additional data at fundamentally different rates. **(b)** Cross-domain transfer obeys the same power-law form with a directional discount τ_{ij}: ABD CT transfers to other targets with different efficiencies; the fitted coefficient need not be equal in the reverse direction. **(c)** Full 6×6 transfer matrix τ_{ij} (from source j to target i) estimated from pure-domain proxy runs. The matrix is asymmetric ($\tau_{ij} \neq \tau_{ji}$) and reveals hub-and-island structure: ABD CT (hub) transfers broadly to most domains, while HN PET (island) is largely isolated.

limited adoption in vision settings, particularly in 3D medical imaging. In practice, mixtures are still chosen by heuristics (i.e. uniform or data-proportional sampling) that treat all imaging domains as interchangeable. However, as shown in Figure 1(a), we observe that each imaging domain exhibit substantially different scaling behavior and the corresponding self-supervised masked autoencoding (MAE) reconstruction loss follows a predictable power law $L \approx E + C/T^{\beta}$, where the scaling exponent β controls how rapidly a domain benefits from additional data, but varies by $3.5\times$ across domains. Furthermore, Figure 1(b) demonstrates that cross-domain transfer also follows the power law discounted by an asymmetric coefficient τ_{ij}. Such two regularities **heterogeneous scaling** and **directional transfer**, provide the analytical foundation to predict mixture losses in closed form and optimize allocations numerically.

We therefore formulate 3D medical data mixture design as a **scaling-law optimization problem** [22, 17, 35, 18]. Rather than directly fitting downstream segmentation performance, we use masked autoencoding (MAE) as a controlled proxy objective that can be applied consistently across heterogeneous 3D imaging domains [14]. Reconstruction loss is a mixture-selection proxy; its usefulness for clinical representations must be established through downstream evaluation. We propose a **transfer-aware scaling law** with two ingredients: *(i)* domain-specific scaling exponents β_i, which capture how efficiently each domain benefits from additional training budget; and *(ii)* an asymmetric transfer matrix τ_{ij}, which measures how effectively training on source domain j contributes to target domain i. Together, these components define a closed-form surrogate for predicting domain-wise MAE loss under arbitrary mixtures. Optimizing this surrogate yields a compute-dependent allocation $\boldsymbol{h}^*(T)$ without exhaustive and compute hungry data mixture ablation experiments.

The resulting allocations reveal interpretable **hub-and-island** structure: domains with broad outgoing transfer act as *hubs*, where compute allocated benefits multiple domains simultaneously, while domains with weak incoming transfer act as *islands* that require direct investment, as knowledge from hubs cannot be transferred to them effectively. Such deterministic knowledge base structure shows that dataset-size weighting alone is insufficient, where data recipe design must account for both domain-specific scaling and cross-domain transfer. Empirically, across six heterogeneous 3D imaging domains spanning CT, MRI, and PET, the estimated transfer matrix reveals strong asymmetry and the derived optimal allocations differ substantially from uniform and data-proportional heuristics. Across the evaluated pretraining budgets, transfer-aware allocation achieves lower MAE loss than the heuristic baselines, with Data-Proportional incurring 4–58% higher loss, and extrapolates to a held-out budget with strong per-domain agreement ($r = 0.989$). We further validate the derived optimal allocation, whose transfer matrices are estimated from principled MAE proxy experiments, on downstream clinical tasks such as abnormal findings classification and lesion segmentation. Our results demonstrate that the proposed transfer-aware data recipe provides a strong prior for supervised 3D medical imaging tasks. In summary, our main contributions are:

- We formulate cross-domain data allocation in 3D medical imaging as a scaling-law optimization problem, using MAE reconstruction loss as a controlled proxy objective across six heterogeneous imaging domains.
- We propose a transfer-aware scaling law that jointly models heterogeneous scaling exponents and an asymmetric transfer matrix, yielding a closed-form surrogate for compute-dependent mixture optimization from small-scale proxy runs.
- We show that the learned allocations reveal interpretable hub-and-island structure, outperform standard heuristic mixtures across budgets, extrapolate from small proxy runs to unseen budgets, and provide a useful prior for downstream tasks.

2 Related Work

3D medical foundation models and self-supervised pretraining. Self-supervised pretraining has become a key strategy for learning transferable representations from unlabeled 3D medical images, where dense annotations are costly. Early work explored restoration, contrastive, and masked prediction objectives for volumetric data [51, 36, 37], while recent foundation-style models scale pretraining to larger and more heterogeneous corpora, including CT volumes, multi-parametric MRI, and diverse multi-domain medical datasets [40, 34, 45, 38]. These studies show that heterogeneous 3D data can support reusable representations across tasks and domains. However, they primarily focus on architectures [19, 13, 12, 25, 50, 27, 26], objectives [38, 24, 44, 49], and evaluation protocols [41, 21, 39, 3, 5], while the data mixture is typically fixed by corpus construction or heuristic sampling rather than optimized.

Scaling laws and data-mixture optimization. Scaling laws provide a quantitative framework for predicting how loss changes with model size, data, and compute [22, 17]. Recent data-centric extensions further show that training mixtures can be estimated from small-scale proxy runs instead of exhaustively swept, including data mixing laws, RegMix, BiMix, and optimal mixture scaling formulations [48, 28, 11, 35]. Related work on multilingual and general-purpose pretraining also studies proportional, temperature-based, or learned reweighting strategies for imbalanced corpora [2, 46, 8, 43]. However, these approaches are largely developed for language, general vision, or native multimodal pretraining, and generally do not model the asymmetric source-to-target transfer that arises across heterogeneous 3D medical domains.

Cross-domain transfer in medical imaging. Medical imaging domains differ substantially in acquisition physics, spatial resolution, anatomical coverage, and clinical signal. Prior multi-domain and multi-modal work has shown that joint training can improve transfer and generalization, including Med3D, Hermes, and recent 3D foundation models [7, 10, 34, 45]. More broadly, transfer effectiveness can be predictable and directional [15]. Transfusion [32] studies feature reuse from natural-image pretraining for 2D medical classification; we instead model directed transfer among 3D medical domains to allocate a fixed pretraining budget (Appendix F). Most medical imaging studies evaluate transfer under a fixed mixture, rather than parameterizing transfer as a quantity for allocation. In contrast, we estimate an asymmetric transfer matrix and use it to optimize compute-aware data allocation across 3D medical domains.

3 Method

We formulate heterogeneous 3D medical data mixture design as a transfer-aware scaling-law optimization problem. Given a fixed self-supervised pretraining budget, our goal is to predict the masked autoencoding (MAE) validation loss induced by a candidate data mixture and choose the mixture that minimizes the expected loss across target domains. Our method is built on two observations. First, different 3D medical domains exhibit different scaling behavior: some domains continue to benefit from additional data, while others saturate more quickly. Second, cross-domain transfer is directional: training on source domain j may benefit target domain i differently from how training on i benefits j. We capture these effects with domain-specific scaling exponents and an asymmetric transfer matrix.

3.1 Problem setup

We consider self-supervised pretraining on K heterogeneous 3D medical imaging domains $\mathcal{D}_1, \ldots, \mathcal{D}_K$. Each domain corresponds to a modality-anatomy pair, such as abdominal CT, abdominal

MRI, brain MRI, or head-and-neck PET (HN-PET). Given a domain weight vector $\boldsymbol{h} = (h_1, \ldots, h_K)$, we define the mixture distribution

$$\mathrm{mix}(\boldsymbol{h}) = \sum_{j=1}^{K} h_j \mathcal{D}_j, \qquad \boldsymbol{h} \in \Delta_\epsilon^{K-1}, \tag{1}$$

where

$$\Delta_\epsilon^{K-1} = \left\{ \boldsymbol{h} \in \mathbb{R}^K : h_j \geq \epsilon,\ \sum_{j=1}^{K} h_j = 1 \right\}. \tag{2}$$

Thus, domain j is sampled with probability h_j. The lower bound ϵ prevents degenerate zero-allocation mixtures and ensures that every domain remains represented during mixture optimization.

Let T denote the total pretraining budget, measured in volume draws. Training with mixture $\boldsymbol{h}$ allocates $h_j T$ expected samples to domain j. We train a 3D masked autoencoder with parameters θ using samples from $\mathrm{mix}(\boldsymbol{h})$. Let $\ell(x, \theta)$ denote the reconstruction loss for a 3D volume x. For a target domain $\mathcal{D}_i$, the target-domain MAE loss is

$$L_i(\theta) = \mathbb{E}_{x \sim \mathcal{D}_i} \left[\ell(x, \theta) \right]. \tag{3}$$

After training on mixture $\boldsymbol{h}$ for budget T, the optimization procedure returns parameters $\theta^*(\boldsymbol{h}, T)$. The loss induced by this training run on target domain i is

$$L_i(\boldsymbol{h}, T) = L_i\left(\theta^*(\boldsymbol{h}, T)\right). \tag{4}$$

Our goal is to predict $L_i(\boldsymbol{h}, T)$ for each target domain without exhaustively training models under many candidate mixtures. We therefore learn a surrogate $\widehat{L}_i(\boldsymbol{h}, T)$ for each target domain and choose the mixture that minimizes the average predicted loss:

$$\boldsymbol{h}^*(T) \in \operatorname*{arg\,min}_{\boldsymbol{h} \in \Delta_\epsilon^{K-1}} \frac{1}{K} \sum_{i=1}^{K} \widehat{L}_i(\boldsymbol{h}, T). \tag{5}$$

This objective treats all target domains equally. A weighted version is introduced in Section 3.3.

3.2 Transfer-aware scaling law

We now derive the predictive law used in Eq. (5). The derivation has three components: domain-specific self-scaling, directed cross-domain transfer, and a mixture-level effective budget.

Domain-specific self-scaling. We first model how each domain scales when trained in isolation. When the model is trained only on samples from $\mathcal{D}_i$, we model the validation MAE loss on the same domain using a power law:

$$L_i(T) = E_i + \frac{C_i}{T^{\beta_i}}, \tag{6}$$

where $E_i \geq 0$ is the asymptotic loss floor, $C_i > 0$ controls the reducible loss, and $\beta_i > 0$ is the scaling exponent for target domain i. The exponent β_i controls the marginal benefit of additional effective data. A larger β_i gives faster decay of the reducible loss within the fitted budget range; β_i alone does not measure anatomical regularity or domain difficulty. This is an empirical surrogate, with units and assumptions detailed in Appendix C. Unlike mixture laws that assume a shared exponent across domains, we allow β_i to vary with the target domain. This is important for 3D medical imaging because the domains differ not only in dataset size, but also in modality, anatomical coverage, spatial resolution, and signal characteristics. The self-scaling law therefore captures the intrinsic data-scaling behavior of each target domain before considering transfer.

Directed cross-domain transfer. Self-domain scaling alone only accounts for samples drawn from the same domain as the target. However, pretraining on one 3D medical domain can improve representations for another domain. We model this effect through a directed transfer matrix $\boldsymbol{\tau} \in [0, 1]^{K \times K}$. The entry τ_{ij} measures how effectively training on source domain $\mathcal{D}_j$ contributes to the target loss on $\mathcal{D}_i$. We set $\tau_{ii} = 1$ by definition. For $i \neq j$, τ_{ij} converts source-domain samples into target-domain effective samples. When the model is trained only on source domain $\mathcal{D}_j$, we predict

the loss on target domain $\mathcal{D}_i$ by reusing the target-domain scaling parameters and discounting the budget:

$$L_{i\leftarrow j}(T) = E_i + \frac{C_i}{(\tau_{ij}T)^{\beta_i}}. \tag{7}$$

This gives τ_{ij} a direct interpretation: T samples from source domain j are equivalent, for target domain i, to $\tau_{ij}T$ samples from the target domain itself. A value close to one indicates strong transfer, while a value near zero indicates weak transfer. As the matrix is directed, so we do not require $\tau_{ij} = \tau_{ji}$. Such directionality is central to our formulation. For example, a source domain with strong anatomical structure may provide useful representations for several targets, while a specialized target domain may not provide the same benefit in reverse. By estimating τ_{ij} for each ordered pair, the model can distinguish transferable sources from domains that require direct allocation.

Transfer-aware mixture law. For a general mixture $\boldsymbol{h}$, source domain j receives h_jT samples. Its contribution to target domain i is discounted by τ_{ij} (the transfer coefficient *from* source j *to* target i), yielding $\tau_{ij}h_jT$ effective target-domain samples. Under a first-order additive transfer approximation, the effective budget reaching target domain i is

$$T_{\mathrm{eff},i}(\boldsymbol{h}, T) = \sum_{j=1}^{K} \tau_{ij}h_jT. \tag{8}$$

Substituting this effective budget into the self-domain scaling law in Eq. (6) gives our transfer-aware scaling law:

$$\widehat{L}_i(\boldsymbol{h}, T) = E_i + \frac{C_i}{\left(\sum_{j=1}^{K} \tau_{ij}h_jT\right)^{\beta_i}}. \tag{9}$$

Equation (9) is the central surrogate used for mixture allocation. It is closed-form, differentiable with respect to $\boldsymbol{h}$, and can be evaluated for arbitrary budgets once its parameters are estimated. The law separates two effects that are entangled in direct mixture sweeps: the target-domain scaling rate, controlled by β_i, and the source-to-target transfer structure, controlled by τ_{ij}. This separation allows us to ask whether a domain should receive budget because it is a high-value target, because it is a transferable source, or because it cannot be reached effectively through other sources.

Special cases. Our formulation recovers simpler data-mixture laws as special cases. If cross-domain transfer is disabled by setting $\tau_{ij} = \delta_{ij}$, where δ_{ij} is the Kronecker delta ($\delta_{ij} = 1$ if $i = j$ and 0 otherwise), then each target receives effective budget only from its own domain:

$$\widehat{L}_i(\boldsymbol{h}, T) = E_i + \frac{C_i}{(h_iT)^{\beta_i}}. \tag{10}$$

We refer to this as a transfer-naive mixture law. If we further constrain all domains to share the same exponent, $\beta_i = \beta$, the model reduces to a shared-exponent mixture law. Our formulation therefore generalizes these alternatives by combining domain-specific scaling and directed cross-domain transfer.

3.3 Fitting and optimal mixture estimation

Once the functional form in Eq. (9) is specified, we estimate its parameters from small-scale pure-domain pretraining runs. We then optimize the fitted surrogate over the mixture simplex. The fitting procedure has two stages.

Fitting self-domain scaling parameters. For each domain i, we train using only samples from $\mathcal{D}_i$ at several small-scale data budgets and record the validation MAE loss across random seeds. We fit (E_i, C_i, β_i) by minimizing:

$$\min_{E_i, C_i, \beta_i} \ \mathbb{E}_{d,\,e}\left[\left(L_i^{(d,e)} - E_i - \frac{C_i}{d^{\beta_i}}\right)^2\right], \tag{11}$$

where the expectation averages uniformly over budgets d and seeds e. We solve this bounded nonlinear least-squares problem using trust-region reflective optimization [6], with constraints $E_i \geq 0$, $C_i > 0$, and $\beta_i \in [0.01, 2.0]$.

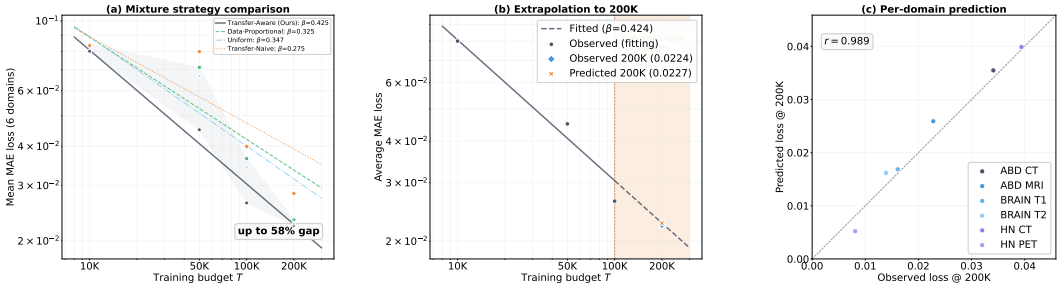

Figure 2: Scaling law validation and extrapolation. **(a)** Transfer-Aware allocation achieves the steepest scaling ($\beta = 0.425$) and lowest MAE loss across all training budgets among the original heuristic baselines. The gap uses $(L_{\text{DP}} - L_{\text{TA}})/L_{\text{TA}}$ and peaks at 58% at $50K$. **(b)** The scaling law fitted on $\{10K, 50K, 100K\}$ predicts 0.0227 at the held-out $200K$ budget, versus 0.0224 observed. **(c)** Per-domain predicted vs. observed loss at $B = 200K$ shows strong agreement across all six domains ($r = 0.989$).

Fitting directed transfer coefficients. For each directed pair (i, j) with $i \neq j$, we train on source domain $\mathcal{D}_j$ and evaluate on target domain $\mathcal{D}_i$. Holding the target-domain scaling parameters (E_i, C_i, β_i) fixed, we estimate the transfer coefficient τ_{ij} (*from* source j *to* target i) by minimizing the expected squared residual over training budgets and random seeds:

$$\hat{\tau}_{ij} = \underset{0<\tau\leq 1}{\arg\min}\ \mathbb{E}_{d,\,e}\left[\left(L_{i\leftarrow j}^{(d,e)} - E_i - \frac{C_i}{(\tau d)^{\beta_i}}\right)^2\right], \tag{12}$$

where d indexes the training budget and e indexes random seeds at that budget. The expectation weights each budget level equally regardless of the number of seeds evaluated, so that budgets with more seeds do not dominate the fit. This two-stage procedure separates intrinsic target-domain scaling from cross-domain transfer and avoids jointly fitting all $K^2 + 2K$ parameters. Each off-diagonal transfer coefficient is fit independently using the same bounded trust-region reflective solver.

Parameter count and estimation cost. The model requires $3K$ self-domain parameters $\{E_i, C_i, \beta_i\}_{i=1}^K$ and $K(K-1)$ off-diagonal transfer coefficients $\{\tau_{ij}\}_{i\neq j}$. For K domains, calibration covers K self-domain and $K(K-1)$ directed source-to-target evaluation configurations, for K^2 configurations in total. These small-scale configurations estimate the scaling-law surrogate. Once fitted, the surrogate can be optimized for new budgets without additional mixture-training sweeps.

Estimating the optimal mixture. With all knowledge transfer scaling-law parameters τ_{ij}'s estimated, we solve for the data mixture weights $\boldsymbol{h} = (h_1, \ldots, h_K)$ that minimize the predicted loss across all target domains. Recall that h_j specifies the fraction of the total training budget T allocated to domain j, with $\sum_j h_j = 1$ and $h_j \geq \epsilon$. Each target domain i receives an effective budget computed by Eq. (8) with cross-domain transfer matrix τ_{ij}, and the surrogate $\widehat{L}_i(\boldsymbol{h}, T)$ predicts the resulting MAE loss. The optimal mixture minimizes the weighted sum of predicted losses across targets:

$$\boldsymbol{h}^*(T) \in \underset{\boldsymbol{h}\in\Delta_\epsilon^{K-1}}{\arg\min} \sum_{i=1}^{K} w_i \widehat{L}_i(\boldsymbol{h}, T), \tag{13}$$

where $w_i \geq 0$ controls the relative importance of target domain i in reconstruction loss. Setting $w_i = 1/K$ weights all domains equally, which is our default objective. Unequal weights allow practitioners to prioritize specific clinical domains if desired. As $\widehat{L}_i(\boldsymbol{h}, T)$ is closed-form and differentiable in $\boldsymbol{h}$, we optimize Eq. (13) using gradient-based constrained optimization. We use SLSQP [23], as it natively enforces both the equality constraint $\sum_j h_j = 1$ and inequality constraints $h_j \geq \epsilon$ that standard unconstrained solvers (e.g., L-BFGS) cannot handle directly [35]. The resulting allocation is budget-dependent: as T changes, the marginal value of assigning samples to each source domain shifts through both the target-domain exponent β_i and the transfer matrix $\boldsymbol{\tau}$, yielding a compute-dependent allocation rule rather than a single static mixture.

Table 1: Effective budget decomposition under Transfer-Aware allocation at $T{=}100K$. The effective budget reaching target domain i is $T_{\text{eff},i} = \sum_{j=1}^{K} \tau_{ij} \cdot h_j \cdot T$, where each source domain j contributes $\tau_{ij} \cdot h_j \cdot T$ effective samples. Self-contributions ($\tau_{ii}{=}1$) are shown in **bold**. ABD-CT (hub, $h^*{=}23.7\%$) provides the largest cross-domain contribution to three targets. HN-PET (island, $h^*{=}56.3\%$) receives almost no incoming transfer and relies on its large direct allocation. Floor domains ($h^*{=}5\%$) achieve 3.7–6.1$\times$ amplification through incoming transfer.

Target i	**Source contribution** $\tau_{ij} \cdot h_j \cdot T$						$T_{\text{eff},i}$	Amplif.
	ABD-CT (h=23.7%)	ABD-MRI (h=5%)	BRAIN-T1 (h=5%)	BRAIN-T2 (h=5%)	HN-CT (h=5%)	HN-PET (h=56.3%)		
ABD-CT (hub)	**23,700**	2,800	1,200	2,450	1,250	2,252	33,652	1.4×
ABD-MRI	11,139	**5,000**	800	1,600	700	1,689	20,928	4.2×
BRAIN-T1	7,584	1,000	**5,000**	1,300	1,950	3,378	20,212	4.0×
BRAIN-T2	8,295	1,300	1,100	**5,000**	950	1,689	18,334	3.7×
HN-CT	5,925	700	1,950	950	**5,000**	15,764	30,289	6.1×
HN-PET (island)	948	150	300	150	1,400	**56,300**	59,248	1.1×

Table 2: Pretraining MAE reconstruction loss (↓). Left: average loss across six domains at different budgets. Right: per-domain loss at $B = 100K$. Gray values denote $(L_{\text{baseline}} - L_{\text{TA}})/L_{\text{TA}}$, the relative increase over Transfer-Aware.

Datamix Strategies	**Mean MAE across budgets**				**Per-domain MAE at** $B = 100K$						
	$10K$	$50K$	$100K$	$200K$	ABD-CT	ABD-MRI	Brain-T1	Brain-T2	HN-CT	HN-PET	Mean
Uniform	0.0853 (+6%)	0.0671 (+49%)	0.0345 (+31%)	0.0228 (+2%)	0.0499	0.0427	0.0202	0.0196	0.0641	0.0106	0.0345
Data-Prop.	0.0843 (+5%)	0.0712 (+58%)	0.0365 (+38%)	0.0233 (+4%)	0.0525	0.0449	0.0213	0.0201	0.0631	0.0169	0.0365
Transfer-Naïve	0.0835 (+4%)	0.0799 (+77%)	0.0399 (+51%)	0.0283 (+26%)	0.0594	0.0476	0.0234	0.0217	0.0692	0.0183	0.0399
Transfer-Aware	**0.0801**	**0.0451**	**0.0264**	**0.0224**	**0.0399**	**0.0297**	**0.0180**	**0.0174**	**0.0479**	**0.0057**	**0.0264**

4 Experimental Design

We evaluate whether transfer-aware scaling laws select better 3D medical pretraining mixtures than standard allocation heuristics, and whether the resulting pretrained encoder transfers to downstream supervised tasks. Our study spans six heterogeneous 3D medical domains: abdominal CT (ABD-CT), abdominal MRI (ABD-MRI), brain MRI-T1 (BRAIN-T1), brain MRI-T2 (BRAIN-T2), head-and-neck CT (HN-CT), and head-and-neck PET (HN-PET). All volumes are resampled to $1.5\,\text{mm}$ isotropic resolution and pretrained using a 3D masked autoencoder (MAE) proxy on 96^3 patches. Preprocessing, model architecture, and optimization details are provided in Appendix A.

Mixture strategies. We compare *Transfer-Aware* with three core baselines; Appendix E adds RegMix and Data Mixing Laws. *Uniform* assigns equal mass to every domain, $h_i = 1/K$. *Data-Proportional* assigns mass proportional to the number of selected pretraining volumes in each domain, reflecting the common default when no scaling-law information is available. *Transfer-Naïve* uses the fitted per-domain scaling laws but disables cross-domain transfer by setting $\tau_{ij} = \delta_{ij}$, isolating the contribution of the directed transfer matrix: it shares the same self-domain calibration as Transfer-Aware but cannot exploit source-to-target transfer. All strategies use the same final pretraining draw budget T; the cost of selecting their mixtures is separate.

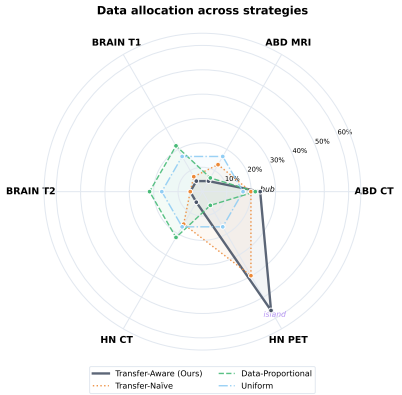

Figure 3: Data allocation across strategies. Transfer-Aware concentrates budget on the hub (ABD-CT) and island (HN-PET), while heuristics spread budget uniformly or by dataset size.

Scaling-law estimation and optimization. We calibrate the law using a 12M-parameter proxy over $K^2 = 36$ source-target evaluation configurations: 6 self-domain and 30 directed cross-domain configurations. Budgets $T \in \{10K, 50K, 100K\}$ are used for the extrapolation fit; $200K$ and mixed-domain runs evaluate its predictions. The selected mixture is then used to pretrain the approximately 121M-parameter target model for downstream

Table 3: Downstream performance across pretraining strategies; mixture-pretrained rows use $T = 100K$. Classification entries are accuracy (AUC), both in percent; segmentation entries are Dice (%). The complete mixture comparison is in Appendix E.

	Classification (Accuracy (AUC) %)			Segmentation (Dice %)					
	OASIS-1	SynBrain-PET	OrganMNIST3D	AMOS		BraTS		HECKTOR	BHSD
Pretraining	Brain MRI	Brain PET	CT	ABD-CT	ABD-MRI	Brain-T1	Brain-T2	HN-CT	Brain CT
No mixture									
From scratch	65.1 (73.8)	75.4 (95.8)	84.6 (98.8)	76.0	62.2	45.2	57.8	62.4	53.4
Single-Modality only	67.4 (76.0)	67.1 (92.8)	88.6 (99.3)	79.3	68.3	51.3	62.1	66.7	55.6
Data mixture ($T = 100$K)									
Data-Proportional	67.5 (76.0)	71.9 (93.9)	89.7 (99.4)	82.0	68.8	52.7	63.0	68.7	59.0
Transfer-Aware (Ours)	**69.7 (78.4)**	**78.1 (95.9)**	**90.3 (99.4)**	**83.1**	**69.4**	**55.1**	**65.4**	**71.1**	**61.2**

transfer (Appendix A.3). We fit self-domain parameters (E_i, C_i, β_i) and directed transfer coefficients τ_{ij} as described in Section 3.3, then optimize the mixture over the constrained simplex with floor ϵ=0.05 to obtain $\boldsymbol{h}^*$. Unless otherwise stated, all target domains are weighted equally in the objective. To interpret the resulting allocation structure, we decompose each mixture into target-specific effective budgets $T_{\text{eff},i} = \sum_j \tau_{ij} h_j T$, which measure how much direct and transferred pretraining signal each target receives (Section 5).

Downstream validation. Beyond the MAE proxy objective, we test whether the derived mixture structure transfers to supervised clinical tasks by finetuning the pretrained ViT encoders on both dense prediction and image-level recognition benchmarks. For segmentation, we evaluate on AMOS 2022 for 16-class abdominal organ segmentation (reporting CT and MRI subsets separately) [20], BraTS 2023 Adult Glioma for multi-label brain tumor segmentation [1], HECKTOR 2025 for head-and-neck tumor segmentation and BHSD for brain hemorrhage segmentation [31, 42]. For classification, we use three benchmarks that probe complementary aspects of 3D representation transfer: OASIS-1 for structural brain MRI-T1 classification related to aging and cognitive impairment [30], SynBrain-PET for brain PET classification testing transfer to functional/metabolic imaging, and OrganMNIST3D for 3D abdominal organ recognition from CT volumes [47]. Across all tasks, segmentation decoders, skip connections, and classification heads are randomly initialized, and all finetuning protocols are held fixed across pretraining strategies. We report Dice (%) for segmentation and accuracy with AUC for classification.

5 Scaling Law Analysis

The learned transfer structure is asymmetric. Figure 1(c) shows the estimated directed transfer matrix τ_{ij} across the six 3D medical domains. Transfer is highly non-uniform and asymmetric: some sources provide broad benefit across targets, while others receive limited external support. In particular, ABD-CT acts as a high-yield source, contributing substantially to multiple targets, whereas HN-PET behaves as a low-transfer island and relies primarily on direct allocation. This structure is reflected in the effective-budget decomposition in Table 1, and a non-uniform optimized mixture visualized in Figure 3. At $T = 100K$, ABD-CT receives 23.7% of the direct budget but provides the largest cross-domain contribution to several targets, including ABD-MRI, BRAIN-T1, and BRAIN-T2. In contrast, HN-PET receives 56.3% of the direct budget because its incoming transfer from other domains is weak. These results show that the optimized mixture is not simply proportional to dataset size or uniformly spread across domains; it reflects a learned source–target transfer structure.

Transfer-aware allocation improves MAE pretraining. Table 2 compares Transfer-Aware allocation with Uniform, Data-Proportional, and Transfer-Naïve baselines under the same total budget. Transfer-Aware achieves the lowest mean MAE loss at every budget shown in Figure 2(a). The advantage is largest in the intermediate, data-bound regime: at $B = 50K$, Transfer-Aware obtains a mean loss of 0.0451, compared with 0.0671 for Uniform, 0.0712 for Data-Proportional, and 0.0799 for Transfer-Naïve, so Data-Proportional is 58% higher (a 36.7% reduction using Data-Proportional as the denominator). At $B = 100K$, Transfer-Aware reaches 0.0264, with Uniform, Data-Proportional, and Transfer-Naïve incurring 31%, 38%, and 51% higher losses, respectively. The comparison with Transfer-Naïve is especially important because it uses the same per-domain scaling calibration but disables cross-domain transfer by setting $\tau_{ij} = \delta_{ij}$. The large gap between Transfer-Aware and

Transfer-Naïve therefore isolates the benefit of modeling directed transfer, rather than only fitting separate per-domain scaling curves. The per-domain results at $B = 100K$ further show that the gain is broad-based: Transfer-Aware achieves the lowest MAE loss on all six domains. The largest reduction occurs on HN-PET, where loss decreases from 0.0169 under Data-Proportional to 0.0057 under Transfer-Aware, consistent with the learned allocation assigning more direct budget to domains that receive little incoming transfer. Appendix D reports seed and bootstrap uncertainty; Appendix E shows that RegMix is better at $50K$ and ties at $200K$.

The scaling law extrapolates to a held-out budget. To test predictive extrapolation beyond the calibration regime, we fit the law using budgets $B \in \{10K, 50K, 100K\}$ and predict performance at the held-out budget $B = 200K$. Figure 2(b) shows that the predicted average loss closely matches the observed value: 0.0227 predicted versus 0.0224 observed. Figure 2(c) further shows strong per-domain agreement, with correlation $r = 0.989$ across the six domains. This supports the use of small proxy runs to estimate mixture parameters for larger pretraining budgets. At $B = 200K$, the relative advantage of Transfer-Aware narrows compared with the intermediate-budget regime. This suggests that the proxy model may be approaching either the irreducible loss floor or a capacity-limited regime, where additional data provides diminishing returns regardless of allocation. We therefore interpret the strongest gains as evidence for transfer-aware allocation in the data-bound regime, while larger-model validation remains an important direction for future work.

6 Downstream Finetuning Results

Transfer-aware pretraining improves classification. Table 3 summarizes the main downstream comparison. On OASIS-1 brain MRI, Transfer-Aware improves accuracy from 67.5% under Data-Proportional to 69.7%, and AUC from 76.0% to 78.4%. On SynBrain-PET, the gain is larger: Transfer-Aware reaches 78.1% accuracy and 95.9% AUC, compared with 71.9% accuracy and 93.9% AUC for Data-Proportional. This result is consistent with the scaling-law analysis, where PET behaves as a weak-transfer domain that benefits from targeted direct allocation. On OrganMNIST3D, Transfer-Aware improves accuracy from 89.7% to 90.3% and matches Data-Proportional AUC at 99.4%. These results indicate that the optimized mixture improves global volume-level representations, not only voxel-level reconstruction.

Transfer-aware pretraining improves segmentation. The same trend holds for dense prediction. On AMOS, Transfer-Aware achieves 83.1 Dice on abdominal CT and 69.4 Dice on abdominal MRI, outperforming Data-Proportional by 1.1 and 0.6 points, respectively. The gains are larger on brain and head-and-neck tasks. On BraTS, Transfer-Aware improves over Data-Proportional from 52.7 to 55.1 Dice on Brain T1 and from 63.0 to 65.4 Dice on Brain T2. Transfer-Aware improves HECKTOR from 68.7 to 71.1 Dice and BHSD brain hemorrhage segmentation from 59.0 to 61.2 Dice. Across all six segmentation tasks, Transfer-Aware yields the best performance among the compared pretraining strategies.

Proxy improvements translate to downstream transfer. Overall, the downstream results mirror the proxy-loss ordering observed in Table 2: Transfer-Aware consistently outperforms Data-Proportional, which in turn generally improves over training from scratch. This agreement suggests that the MAE proxy is not merely optimizing reconstruction loss in isolation, but provides a useful signal for selecting pretraining mixtures that transfer to supervised classification and segmentation. Across the full comparison (Appendix E), Transfer-Aware improves mean Dice over RegMix by 1.57 points and has the highest mean classification accuracy, although RegMix leads on OrganMNIST3D accuracy and mean AUC.

7 Discussion

Our transfer-aware scaling law demonstrates that heterogeneous 3D medical data mixtures can be optimized using the fitted surrogate without exhaustive mixture training. The framework outperforms the original heuristic baselines across the tested budgets, predicts 0.0227 versus 0.0224 observed loss at a held-out budget (r=0.989), and transfers to downstream clinical tasks. The learned allocations reveal interpretable hub-and-island structure that provides actionable guidance for practitioners training unified 3D medical foundation models.

Limitations. Two modeling assumptions merit further examination. First, our mixture law assumes transfer contributions compose linearly ($T_{\text{eff},i} = \sum_j \tau_{ij} h_j T$), which cannot capture synergistic or interfering interactions between source domains. The empirical results support the usefulness of this approximation in our six-domain setting, but nonlinear interaction terms may become important as domain count or modality diversity increases. Second, the transfer matrix $\boldsymbol{\tau}$ is estimated once from small-scale proxy runs and held constant across budgets. Our extrapolation results support its predictive use within the $10K$–$200K$ range, but transfer structure could evolve at substantially larger scales as models shift from learning broadly transferable low-level features to more domain-specific representations. Modeling $\tau_{ij}(T)$ as a budget-dependent function is a natural extension for future work. Validation is limited to six domains with 150–500 selected volumes each; cross-capacity invariance of the transfer matrix and anatomical fidelity of reconstructions remain unestablished (Appendix F). A broader impact discussion is provided in Appendix H.

8 Conclusion

We introduced a transfer-aware scaling law for 3D medical data mixture optimization. By modeling both domain-specific scaling behavior and directed cross-domain transfer, our framework converts heterogeneous pretraining mixture design from a heuristic choice into an explicit compute-aware optimization problem. Across six diverse CT, MRI, and PET domains, the learned allocation improves MAE pretraining loss, extrapolates accurately to a held-out budget, and yields pretrained encoders that transfer better to downstream classification and segmentation tasks. These results highlight data allocation as a key design axis for scalable 3D medical foundation models and suggest that proxy-scale scaling laws can guide more efficient use of heterogeneous medical imaging data.

References

[1] Adewole, M., Rudie, J.D., Gbdamosi, A., Toyobo, O., Raymond, C., Zhang, D., Omidiji, O., Akinola, R., Suwaid, M.A., Emegoakor, A., et al.: The brain tumor segmentation (brats) challenge 2023: Glioma segmentation in sub-saharan africa patient population (brats-africa). ArXiv pp. arXiv–2305 (2023)

[2] Arivazhagan, N., Bapna, A., Firat, O., Lepikhin, D., Johnson, M., Krikun, M., Chen, M.X., Cao, Y., Foster, G., Cherry, C., et al.: Massively multilingual neural machine translation in the wild: Findings and challenges. arXiv preprint arXiv:1907.05019 (2019)

[3] Bai, Y., Ying, J., Cao, Y., Lv, X., He, Y., Wang, X., Yu, J., Zeng, K., Xiao, Y., Lyu, H., et al.: Benchmarking foundation models with language-model-as-an-examiner. Advances in Neural Information Processing Systems **36**, 78142–78167 (2023)

[4] Barnard, L., Botha, H., Corriveau-Lecavalier, N., Graff-Radford, J., Dicks, E., Gogineni, V., Zhang, G., Burkett, B.J., Johnson, D.R., Huls, S.J., et al.: An fdg-pet–based machine learning framework to support neurologic decision-making in alzheimer disease and related disorders. Neurology **105**(2), e213831 (2025)

[5] Bassi, P.R., Li, W., Tang, Y., Isensee, F., Wang, Z., Chen, J., Chou, Y.C., Roy, S., Kirchhoff, Y., Rokuss, M., et al.: Touchstone benchmark: Are we on the right way for evaluating ai algorithms for medical segmentation? Advances in Neural Information Processing Systems **37**, 15184–15201 (2024)

[6] Branch, M.A., Coleman, T.F., Li, Y.: A subspace, interior, and conjugate gradient method for large-scale bound-constrained minimization problems. SIAM Journal on Scientific Computing **21**(1), 1–23 (1999)

[7] Chen, S., Ma, K., Zheng, Y.: Med3d: Transfer learning for 3d medical image analysis. arXiv preprint arXiv:1904.00625 (2019)

[8] Chung, H.W., Constant, N., Garcia, X., Roberts, A., Tay, Y., Narang, S., Firat, O.: Unimax: Fairer and more effective language sampling for large-scale multilingual pretraining. arXiv preprint arXiv:2304.09151 (2023)

[9] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)

[10] Gao, Y.: Training like a medical resident: Context-prior learning toward universal medical image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11194–11204 (2024)

[11] Ge, C., Ma, Z., Chen, D., Li, Y., Ding, B.: Bimix: A bivariate data mixing law for language model pretraining. arXiv preprint arXiv:2405.14908 (2024)

[12] Hatamizadeh, A., Nath, V., Tang, Y., Yang, D., Roth, H.R., Xu, D.: Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. In: International MICCAI Brainlesion Workshop. pp. 272–284. Springer (2022)

[13] Hatamizadeh, A., Tang, Y., Nath, V., Yang, D., Myronenko, A., Landman, B., Roth, H.R., Xu, D.: Unetr: Transformers for 3d medical image segmentation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 574–584 (2022)

[14] He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16000–16009 (2022)

[15] Hernandez, D., Kaplan, J., Henighan, T., McCandlish, S.: Scaling laws for transfer. arXiv preprint arXiv:2102.01293 (2021)

[16] Hestness, J., Narang, S., Ardalani, N., Diamos, G., Jun, H., Kianinejad, H., Patwary, M.M.A., Yang, Y., Zhou, Y.: Deep learning scaling is predictable, empirically. arXiv preprint arXiv:1712.00409 (2017), `https://arxiv.org/abs/1712.00409`

[17] Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., Casas, D., Hendricks, L.A., Welbl, J., Clark, A., et al.: Training compute-optimal large language models. arXiv preprint arXiv:2203.15556 **10** (2022)

[18] Ilse, M., Sharma, H., Schwaighofer, A., Bond-Taylor, S., Pérez-García, F., Melnichenko, O., Sykes, A.M.G., Horst, K.K., Khandelwal, A., Reynolds, M., et al.: Data scaling laws for radiology foundation models. arXiv preprint arXiv:2509.12818 (2025)

[19] Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H.: nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. Nature methods **18**(2), 203–211 (2021)

[20] Ji, Y., Bai, H., Yang, J., Ge, C., Zhu, Y., Zhang, R., Li, Z., Zhang, L., Ma, W., Wan, X., et al.: Amos: A large-scale abdominal multi-organ benchmark for versatile medical image segmentation. arXiv preprint arXiv:2206.08023 (2022)

[21] Jin, R., Xu, Z., Zhong, Y., Yao, Q., Dou, Q., Zhou, S.K., Li, X.: Fairmedfm: fairness benchmarking for medical imaging foundation models. Advances in Neural Information Processing Systems **37**, 111318–111357 (2024)

[22] Kaplan, J., McCandlish, S., Henighan, T., Brown, T.B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., Amodei, D.: Scaling laws for neural language models. arXiv preprint arXiv:2001.08361 (2020)

[23] Kraft, D.: A software package for sequential quadratic programming. Forschungsbericht-Deutsche Forschungs- und Versuchsanstalt fur Luft- und Raumfahrt (1988)

[24] Krishnan, R., Rajpurkar, P., Topol, E.J.: Self-supervised learning in medicine and healthcare. Nature Biomedical Engineering **6**(12), 1346–1352 (2022)

[25] Lee, H.H., Bao, S., Huo, Y., Landman, B.A.: 3d ux-net: A large kernel volumetric convnet modernizing hierarchical transformer for medical image segmentation. In: The Eleventh International Conference on Learning Representations (2022)

[26] Lee, H.H., Liu, Q., Bao, S., Huo, Y., Landman, B.A.: Rep3d: Re-parameterize large 3d kernels with low-rank receptive modeling for medical imaging. arXiv preprint arXiv:2505.19603 (2025)

[27] Lee, H.H., Liu, Q., Bao, S., Yang, Q., Yu, X., Cai, L.Y., Li, T., Huo, Y., Koutsoukos, X., Landman, B.A.: Scaling up 3d kernels with bayesian frequency re-parameterization for medical image segmentation. arXiv preprint arXiv:2303.05785 (2023)

[28] Liu, Q., Zheng, X., Muennighoff, N., Zeng, G., Dou, L., Pang, T., Jiang, J., Lin, M.: Regmix: Data mixture as regression for language model pre-training. arXiv preprint arXiv:2407.01492 (2024)

[29] Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: International Conference on Learning Representations (2019), https://arxiv.org/abs/1711.05101

[30] Marcus, D.S., Wang, T.H., Parker, J., Csernansky, J.G., Morris, J.C., Buckner, R.L.: Open access series of imaging studies (oasis): cross-sectional mri data in young, middle aged, nondemented, and demented older adults. Journal of cognitive neuroscience **19**(9), 1498–1507 (2007)

[31] Quetin, S., Enger, S.A.: Automatic lesion and lymph node segmentation from pet and ct scans of the head and neck region: a hecktor 2025 challenge report. In: Fourth Head and Neck Cancer Tumor Lesion Segmentation, Diagnosis and Prognosis (2025)

[32] Raghu, M., Zhang, C., Kleinberg, J., Bengio, S.: Transfusion: Understanding transfer learning for medical imaging. In: Advances in Neural Information Processing Systems. vol. 32 (2019), https://proceedings.neurips.cc/paper_files/paper/2019/hash/eb1e78328c46506b46a4ac4a1e378b91-Abstract.html

[33] Rosenfeld, J.S., Rosenfeld, A., Belinkov, Y., Shavit, N.: A constructive prediction of the generalization error across scales. In: International Conference on Learning Representations (2020), https://openreview.net/forum?id=ryenvpEKDr

[34] Rui, S., Chen, L., Tang, Z., Wang, L., Liu, M., Zhang, S., Wang, X.: Multi-modal vision pre-training for medical image analysis. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 5164–5174 (2025)

[35] Shukor, M., Bethune, L., Busbridge, D., Grangier, D., Fini, E., El-Nouby, A., Ablin, P.: Scaling laws for optimal data mixtures. arXiv preprint arXiv:2507.09404 (2025)

[36] Taleb, A., Loetzsch, W., Danz, N., Severin, J., Gaertner, T., Bergner, B., Lippert, C.: 3d self-supervised methods for medical imaging. arXiv preprint arXiv:2006.03829 (2020)

[37] Tang, Y., Yang, D., Li, W., Roth, H.R., Landman, B., Xu, D., Nath, V., Hatamizadeh, A.: Self-supervised pre-training of swin transformers for 3d medical image analysis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20730–20740 (2022)

[38] Wald, T., Ulrich, C., Suprijadi, J., Ziegler, S., Nohel, M., Peretzke, R., Kohler, G., Maier-Hein, K.: An openmind for 3d medical vision self-supervised learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 23839–23879 (2025)

[39] Wang, D., Wang, X., Wang, L., Li, M., Da, Q., Liu, X., Gao, X., Shen, J., He, J., Shen, T., et al.: A real-world dataset and benchmark for foundation model adaptation in medical image classification. Scientific Data **10**(1), 574 (2023)

[40] Wang, G., Wu, J., Luo, X., Liu, X., Li, K., Zhang, S.: Mis-fm: 3d medical image segmentation using foundation models pretrained on a large-scale unannotated dataset. arXiv preprint arXiv:2306.16925 (2023)

[41] Wang, J., Wang, X., Lyu, L., Chen, J., Ma, F.: Fedmeki: A benchmark for scaling medical foundation models via federated knowledge injection. Advances in Neural Information Processing Systems **37**, 1082–1116 (2024)

[42] Wu, B., Xie, Y., Zhang, Z., Ge, J., Yaxley, K., Bahadir, S., Wu, Q., Liu, Y., To, M.S.: Bhsd: A 3d multi-class brain hemorrhage segmentation dataset. In: International Workshop on Machine Learning in Medical Imaging. pp. 147–156. Springer (2023)

[43] Xie, S.M., Pham, H., Dong, X., Du, N., Liu, H., Lu, Y., Liang, P.S., Le, Q.V., Ma, T., Yu, A.W.: Doremi: Optimizing data mixtures speeds up language model pretraining. Advances in Neural Information Processing Systems **36**, 69798–69818 (2023)

[44] Xie, Y., Zhang, J., Xia, Y., Wu, Q.: Unimiss: Universal medical self-supervised learning via breaking dimensionality barrier. In: European Conference on Computer Vision. pp. 558–575. Springer (2022)

[45] Xu, T., Hosseini, S., Anderson, C., Rinaldi, A., Krishnan, R.G., Martel, A.L., Goubran, M.: A generalizable 3d framework and model for self-supervised learning in medical imaging. npj Digital Medicine **8**(1), 639 (2025)

[46] Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., Barua, A., Raffel, C.: mt5: A massively multilingual pre-trained text-to-text transformer. In: Proceedings of the 2021 conference of the North American chapter of the association for computational linguistics: Human language technologies. pp. 483–498 (2021)

[47] Yang, J., Shi, R., Wei, D., Liu, Z., Zhao, L., Ke, B., Pfister, H., Ni, B.: Medmnist v2-a large-scale lightweight benchmark for 2d and 3d biomedical image classification. Scientific data **10**(1), 41 (2023)

[48] Ye, J., Liu, P., Sun, T., Zhan, J., Zhou, Y., Qiu, X.: Data mixing laws: Optimizing data mixtures by predicting language modeling performance. arXiv preprint arXiv:2403.16952 (2024)

[49] Ye, Y., Xie, Y., Zhang, J., Chen, Z., Wu, Q., Xia, Y.: Continual self-supervised learning: Towards universal multi-modal medical data representation learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11114–11124 (2024)

[50] Yu, X., Yang, Q., Zhou, Y., Cai, L.Y., Gao, R., Lee, H.H., Li, T., Bao, S., Xu, Z., Lasko, T.A., et al.: Unest: local spatial representation learning with hierarchical transformer for efficient medical segmentation. Medical Image Analysis p. 102939 (2023)

[51] Zhou, Z., Sodha, V., Pang, J., Gotway, M.B., Liang, J.: Models genesis. Medical image analysis **67**, 101840 (2021)

A Experimental Details

A.1 Dataset Description

Pretraining datasets. Our six pretraining domains are drawn from three publicly available 3D medical imaging collections spanning CT, MRI, and PET (Table 4). To simulate realistic data imbalance across domains, we subsample each source dataset to a controlled size: 500 volumes for high-availability domains (ABD-CT, BRAIN-T1, BRAIN-T2, HN-CT) and 150 volumes for low-availability domains (ABD-MRI, HN-PET). This $3.3\times$ imbalance ratio reflects practical clinical settings where some modalities (e.g., CT) are far more abundant than others (e.g., PET or MRI), and ensures that data-proportional allocation produces meaningfully different mixtures from uniform allocation. The specific volume counts are chosen to be large enough for stable MAE training at our proxy scale with time and cost efficiency (i.e. Finished pretraining within a day for empirical observation) while remaining computationally tractable for $K^2 = 36$ source-target configurations across calibration and evaluation budgets. All pretraining volumes are unlabeled and used exclusively for self-supervised MAE reconstruction.

Table 4: Pretraining domain statistics. "Available" refers to the total unlabeled volumes in the source dataset; "Selected" is the number subsampled for our experiments. All volumes are resampled to $1.5\,\mathrm{mm}$ isotropic resolution.

Domain	Modality	Available	Selected	Source
ABD-CT	CT	1100	500	AMOS 2022 [20]
ABD-MRI	MRI	700	150	AMOS 2022 [20]
BRAIN-T1	MRI (T1 + ce)	2052	500	BraTS 2023 [1]
BRAIN-T2	MRI (T2 + FLAIR)	2052	500	BraTS 2023 [1]
HN-CT	CT	700	500	HECKTOR 2025 [31]
HN-PET	PET	700	150	HECKTOR 2025 [31]

Downstream finetuning datasets. To validate that the MAE-derived mixture structure transfers beyond the proxy objective, we finetune pretrained encoders on seven datasets covering six segmentation tasks and three classification tasks across all three modalities (Table 5). Downstream subjects are held out from pretraining to avoid data leakage. For segmentation, we use the labeled subsets of the same source collections (AMOS, BraTS, HECKTOR) as well as the BHSD brain hemorrhage CT dataset. For classification, we use three benchmarks that probe complementary aspects of 3D representation transfer: structural brain MRI (OASIS-1), functional brain PET (SynBrain-PET), and abdominal CT organ recognition (OrganMNIST3D).

Table 5: Downstream finetuning datasets. Segmentation tasks report Dice (%); classification tasks report accuracy and AUC. All finetuning protocols are held fixed across pretraining strategies.

Dataset	Task	Classes	Number of Subjects (Train/Val/Test)	Metric	Domain
Segmentation					
AMOS 2022 [20]	Organ seg.	16	200 (160/20/20)	Dice	ABD-CT / MRI
BraTS 2023 [1]	Tumor seg.	3	480 (400/40/40)	Dice	BRAIN-T1 / T2
HECKTOR 2025 [31]	Tumor seg.	3	524 (418/53/53)	Dice	HN-CT
BHSD [42]	Hemorrhage seg.	3	92 (72/10/10)	Dice	Brain CT
Classification					
OASIS-1 [30]	Brain class.	2	235 (165/35/35)	Acc / AUC	BRAIN-T1
SynBrain-PET [4]	PET class.	5	500 (348/76/76)	Acc / AUC	Brain PET
OrganMNIST3D [47]	Organ class.	11	1,743 (972/161/610)	Acc / AUC	ABD-CT

A.2 Preprocessing

Pretraining preprocessing. All pretraining volumes are reoriented to RAS and resampled to $1.5\,\mathrm{mm}$ isotropic resolution using trilinear interpolation. Domain-specific intensity normalization maps all volumes to $[-1, 1]$ based on modality and anatomy, as summarized in Table 6. After normalization, foreground regions are cropped, volumes are padded to at least 96^3 voxels, and a random 96^3 patch is extracted for training. Light augmentation is applied during pretraining: random

intensity scaling ($\pm 10\%$, $p = 0.3$), random contrast adjustment ($p = 0.2$), and random Gaussian noise ($\sigma = 0.02$, $p = 0.15$). A 75% random patch masking ratio is applied for masked autoencoding.

Table 6: Domain-specific intensity preprocessing. All domains share the same spatial preprocessing (RAS orientation, $1.5\,\mathrm{mm}$ isotropic resampling, foreground cropping, 96^3 patching).

Domain	Family	Intensity normalization	Output range
ABD-CT	CT	HU window $[-160, 240]$, linear rescale	$[-1, 1]$
HN-CT	CT	HU window $[-200, 300]$, linear rescale	$[-1, 1]$
ABD-MRI	MRI	z-norm (nonzero), clip $\pm 3\sigma$, rescale	$[-1, 1]$
BRAIN-T1	MRI	z-norm (nonzero), clip $\pm 5\sigma$, rescale	$[-1, 1]$
BRAIN-T2	MRI	z-norm (nonzero), clip $\pm 5\sigma$, rescale	$[-1, 1]$
HN-PET	PET	z-norm (nonzero), clip $\pm 5\sigma$, rescale	$[-1, 1]$

Finetuning preprocessing. Finetuning transforms are designed to match the pretraining intensity and resolution pipeline exactly, ensuring that pretrained encoder weights see the same input distribution. Labels are resampled using nearest-neighbor interpolation to preserve discrete class values. Table 7 summarizes the domain-specific finetuning configuration. The key design choices are:

- **BraTS (BRAIN-T1/T2):** Labels $\{0, 1, 2, 4\}$ are converted to three overlapping binary channels (WT, TC, ET) *after* spatial transforms and random cropping, enabling multi-label sigmoid training. The conversion is placed after `RandCropByPosNegLabeld` in MONAI because the cropper requires single-channel integer labels to identify foreground voxels.
- **CT lesion tasks (HECKTOR, BHSD):** Small tumor/lesion targets use `pos=4, neg=1` with `num_samples=4` to ensure sufficient foreground exposure during training.
- **AMOS (ABD-CT/MRI):** Standard organ segmentation with `pos=1, neg=1` and `num_samples=2`.

Table 7: Finetuning configuration per downstream domain.

Domain	Intensity	pos	neg	samples	Loss	Dice metric
ABD-CT	HU $[-160, 240] \to [-1, 1]$	1	1	2	DiceCE (softmax)	incl. background
ABD-MRI	z-norm, $\pm 3\sigma \to [-1, 1]$	1	1	2	DiceCE (softmax)	incl. background
BRAIN-T1	z-norm, $\pm 5\sigma \to [-1, 1]$	4	1	2	DiceCE (sigmoid)	excl. background
BRAIN-T2	z-norm, $\pm 5\sigma \to [-1, 1]$	4	1	2	DiceCE (sigmoid)	excl. background
HN-CT	HU $[-200, 300] \to [-1, 1]$	4	1	4	DiceCE (softmax)	excl. background

Finetuning augmentation. All segmentation tasks share the same augmentation pipeline: random intensity shift (± 0.1, $p = 0.5$) and random affine transform (rotation $\pm \pi/30$, scale $\pm 10\%$, $p = 1.0$). No augmentation is applied during validation.

Inference. Training-time validation uses sliding-window inference with `sw_batch_size`$= 1$ and overlap $= 0.25$ for computational efficiency. Final evaluation uses `sw_batch_size`$= 4$ and overlap $= 0.7$ for higher accuracy. All predictions and metrics are computed at the resampled $1.5\,\mathrm{mm}$ resolution.

A.3 Model Architecture

We distinguish the 12M-parameter proxy used to estimate the scaling law and transfer matrix from the approximately 121M-parameter target model pretrained with the selected mixture for downstream evaluation. The following architecture and parameter breakdown describe the target model. Its Vision Transformer (ViT) [9] has 768-dimensional hidden size, 12 transformer layers, 12 attention heads, MLP dimension 3072, and 16^3 patch size. The ViT encoder is embedded within the UNETR [13] framework, which provides a CNN decoder for masked patch reconstruction during pretraining. Each domain uses a dedicated convolutional input adapter that projects single-channel 96^3 volumes to the shared patch embedding space. During pretraining, 75% of patches are randomly masked and the model reconstructs masked patches using per-patch MSE loss [14].

Table 8: Target UNETR parameter breakdown. The approximately 121M count includes the ViT encoder and CNN decoder. Only the ViT encoder is initialized from the pretrained MAE checkpoint; the CNN decoder and task-specific heads are randomly initialized for each downstream task.

Component	Parameters	Initialization
ViT encoder (`vit.*`)	116,679,936	Pretrained (MAE)
CNN decoder (`encoder1-4, decoder2-5`)	4,443,140	Random
Output head (`out`)	varies by task	Random
Total	$\sim$121M	–

A.4 Pretraining Optimization

All pretraining runs use AdamW [29] with learning rate 1.5×10^{-4}, batch size 4. Training exposure is indexed by the volume-draw budget T, with domain sampling governed by the prescribed mixture weights $\boldsymbol{h}$. We use a cache rate of 1.0 for MONAI CacheDataset to maximize throughput. Each configuration is run with 2–3 random seeds and results are averaged to reduce variance in both scaling-law fitting and final evaluation.

A.5 Downstream Model Initialization

Segmentation: UNETR with pretrained ViT encoder. For segmentation tasks, we use UNETR [13], which combines a ViT encoder with a CNN decoder connected through skip connections at multiple resolutions. The pretrained MAE checkpoint stores the full UNETR state dict (ViT encoder + CNN reconstruction decoder). During finetuning initialization, we load **only the ViT encoder weights** (`vit.*` keys) and **randomly initialize** the CNN decoder branches (`encoder1`–`encoder4`, `decoder2`–`decoder5`, `out`). This is deliberate: the ViT encoder learns spatial representations from MAE that transfer well to segmentation, but the CNN decoder was trained for pixel-level reconstruction (regression), which differs fundamentally from segmentation (per-voxel classification). Loading reconstruction decoder weights can cause negative transfer. The segmentation output head is initialized with the correct number of output channels for the target task. Weight loading proceeds as follows:

1. Load the MAE checkpoint containing `unetr_state_dict` and metadata (`scale_factor`, `num_layers`).
2. Build a fresh UNETR with `out_channels = num_classes` for the target task.
3. Extract all keys matching `vit.*` from the checkpoint and load them into the new UNETR's ViT encoder, verifying shape compatibility.
4. All remaining parameters (CNN decoder, skip connections, output head) retain random initialization.

Classification: ViT encoder with MLP head. For classification tasks, we extract the same pretrained ViT encoder and attach a lightweight MLP classification head. The ViT processes the input 96^3 volume into a sequence of patch embeddings, which are globally average-pooled to produce a single feature vector of dimension $d = 768$. The pooled feature is passed through LayerNorm and an MLP head to map it to the target classes:

$$\hat{y} = f_{\text{MLP}}\left(\text{LN}\left(\text{GAP}(\text{ViT}(x))\right)\right),$$

where GAP denotes global average pooling over the patch token sequence. The MLP head is randomly initialized, while the ViT encoder is loaded from the same pretrained checkpoint and finetuned according to the downstream training schedule.

From-scratch baseline. Both segmentation and classification baselines use the identical architecture with all parameters randomly initialized (no pretrained weights), isolating the effect of pretraining from architectural differences.

A.6 Downstream Finetuning Protocol

Finetuning uses AdamW with learning rate $1e{-}4$, batch size 2, and 200 epochs for all downstream tasks. For segmentation, we use DiceCE loss with softmax for multi-class tasks (AMOS, HECKTOR, BHSD) and DiceCE loss with sigmoid for multi-label segmentation (BraTS). For classification, we use cross-entropy loss. All finetuning protocols (i.e. optimizer, learning rate, epochs, augmentation, and evaluation) are held fixed across pretraining strategies to ensure fair comparison.

Segmentation augmentation. Training augmentation follows [25]: foreground cropping, spatial padding to 96^3, random cropping via `RandCropByPosNegLabeld` (pos$=1$, neg$=1$ for organ segmentation; pos$=4$, neg$=1$ for small lesion/tumor tasks) in MONAI package, random intensity shifts (± 0.1, $p = 0.5$), and random affine transforms (rotation $\pm\pi/30$, scale $\pm 10\%$). Sliding-window evaluation follows the stage-specific inference settings in Appendix A.2.

A.7 Fixed Model Size

We fit the scaling law at a fixed proxy capacity and optimize over data mixture and budget, rather than jointly optimizing model size as in Chinchilla-style scaling laws [17]. This is deliberate: 3D medical imaging at 96^3 resolution is GPU-memory-bound, and model architecture is typically standardized across benchmarks [13, 25]. The practical question facing practitioners is how to allocate a fixed data budget across heterogeneous domains, not how to trade off model size against data. Our scaling law therefore models $L(\boldsymbol{h}, T)$ at fixed model size N, making the surrogate directly actionable for the regime where most 3D medical pretraining operates. The downstream experiments test whether the selected allocation remains useful for a larger target model; they do not refit the transfer matrix at multiple capacities. Joint estimation of $L(N, \boldsymbol{h}, T)$ remains future work.

B Additional Results

B.1 Domain Characterization

Table 9: Domain characterization from the transfer matrix and prescribed allocations.

Domain	$\bar{\tau}_{\text{out}}$	$\bar{\tau}_{\text{in}}$	**Role**	h^*_{TA}	h_{DP}
ABD-CT	0.29	0.32	Hub	23.7%	21.7%
ABD-MRI	0.24	0.22	–	5.0%	6.5%
BRAIN-T1	0.21	0.25	–	5.0%	21.7%
BRAIN-T2	0.26	0.21	–	5.0%	21.7%
HN-CT	0.25	0.25	–	5.0%	21.7%
HN-PET	0.09	0.09	Island	56.3%	6.5%

Table 9 summarizes the role of each domain as determined by the estimated transfer matrix. ABD-CT acts as a hub with the highest mean outgoing transfer ($\bar{\tau}_{\text{out}} = 0.29$), contributing meaningfully to four other domains. HN-PET is an island with negligible incoming and outgoing transfer ($\bar{\tau} = 0.09$), within the fitted reconstruction proxy. BRAIN-T1 and BRAIN-T2 have comparatively small fitted exponents, indicating slower decay of reducible reconstruction loss over the evaluated range; this does not establish anatomical regularity or intrinsic task difficulty.

B.2 Transfer Amplification

Table 10 shows the transfer amplification effect at $B = 100K$. Floor domains ($h = 5\%$) receive only $5{,}000$ direct volume draws but achieve 3.7–$6.1\times$ effective budget through cross-domain transfer. Despite this minimal direct allocation, all floor domains achieve lower MAE loss under Transfer-Aware than under Data-Proportional, demonstrating a Pareto improvement across all domains simultaneously. The largest improvement is on HN-PET (-66%), where Transfer-Aware allocates 56.3% of the budget versus only 6.5% under Data-Proportional.

Table 10: Transfer amplification at $B = 100K$. Floor domains ($h = 5\%$) achieve 3.7–6.1× effective budget through cross-domain transfer, and still outperform Data-Proportional.

Domain	$h_i \cdot T$	T_{eff}	Boost	TA (Ours)	DP	Δ
ABD-CT	23,700	33,652	1.4×	**0.0399**	0.0525	−24%
ABD-MRI	5,000	20,928	4.2×	**0.0297**	0.0449	−34%
BRAIN-T1	5,000	20,212	4.0×	**0.0180**	0.0213	−15%
BRAIN-T2	5,000	18,334	3.7×	**0.0174**	0.0201	−13%
HN-CT	5,000	30,289	6.1×	**0.0479**	0.0631	−24%
HN-PET	56,300	59,248	1.1×	**0.0057**	0.0169	−66%

Table 11: Descriptive power-law fits to the observed mixture curves ($L_i = C/T^{\beta}$). These mixture-specific exponents are distinct from the self-domain β_i used to calibrate Eq. (6).

	TA (Ours)		**DP**		
Domain	β	C	β	C	Ratio
ABD-CT	0.407	5.02	0.302	2.05	1.35×
ABD-MRI	0.487	9.51	0.352	3.01	1.38×
BRAIN-T1	0.209	0.21	0.173	0.15	1.21×
BRAIN-T2	0.243	0.30	0.225	0.27	1.08×
HN-CT	0.436	8.12	0.330	3.24	1.32×
HN-PET	0.724	39.80	0.466	4.96	1.55×
Mean	**0.418**	–	0.308	–	1.36×

B.3 Scaling Exponent Comparison

Table 11 compares the fitted scaling exponents under Transfer-Aware and Data-Proportional allocation. The arithmetic mean of the displayed per-domain exponents is 0.418 for Transfer-Aware and 0.308 for Data-Proportional; these differ from the fits to aggregate mean-loss curves in Figure 2. The fitted slope alone does not determine the absolute marginal loss reduction. The largest exponent ratio is on HN-PET ($1.55\times$), reflecting that direct investment in the island domain produces the steepest returns. Brain domains show the smallest ratios (1.08–$1.21\times$), indicating smaller differences in empirical log-log slopes.

B.4 Interpreting Effective Budgets

Effective-budget decompositions describe how the fitted additive model combines direct and transferred exposure. They are model-based quantities, rather than independent measurements of representation quality. Validation of the surrogate therefore requires comparing predicted and observed losses on configurations excluded from calibration; Section 5 reports the held-out $200K$ prediction. A correlation between effective-budget ratios and loss ratios alone does not establish calibration of each target-domain prediction.

B.5 Per-Domain Scaling Curves

Figure 4 shows per-domain MAE loss across all four mixture strategies on log-log axes. Transfer-Aware achieves a steeper scaling exponent on all six domains, with the largest gains on HN-PET (β : 0.724 vs. 0.466) and ABD-MRI (β : 0.487 vs. 0.352). Brain domains show the smallest inter-strategy gap, indicating smaller differences in the empirical log-log slopes. The $3.5\times$ heterogeneity in β across domains (0.209–0.724) motivates allowing domain-specific exponents in heterogeneous 3D medical data mixtures.

B.6 Floor Constraint Ablation

Figure 5 shows the effect of varying the floor constraint ϵ from 2.5% to 12.5%. The optimal value is $\epsilon = 5\%$: lower floors (2.5%) over-concentrate budget on just two domains (ABD-CT and HN-PET), while higher floors ($\geq 10\%$) force budget into domains that receive sufficient signal through transfer,

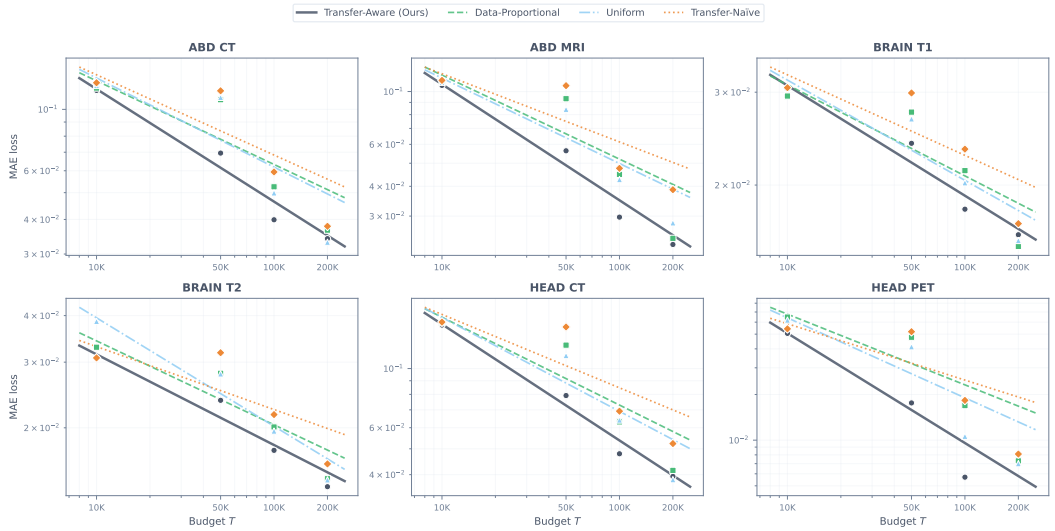

Figure 4: Per-domain scaling laws (log-log) comparing transfer-aware and data-proportional allocation. Transfer-aware achieves a steeper scaling exponent β on all 6 domains, with the largest gains on HEAD PET (β: 0.724 vs. 0.466) and ABD MRI (β : 0.487 vs. 0.352). BRAIN T1 and BRAIN T2 show the smallest differences, indicating smaller differences in the empirical log-log slopes. The 3.5× heterogeneity in β across domains (0.209–0.724) motivates allowing domain-specific exponents in heterogeneous medical data mixtures.

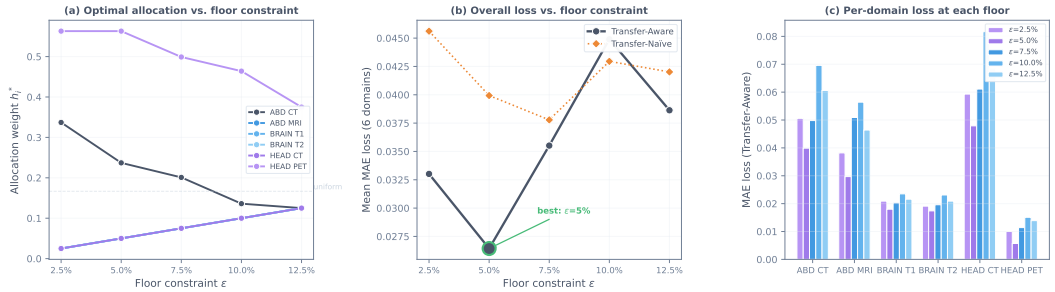

Figure 5: Floor constraint ablation. (a) As the floor ϵ increases, the optimizer loses flexibility: HEAD PET allocation drops from 56.3% to 37.5% and ABD CT from 33.7% to 12.5%. **(b)** Mean MAE loss is minimized at ϵ=5%; higher floors force budget into domains that receive sufficient signal through transfer, degrading overall performance. **(c)** Per-domain losses confirm that ϵ=5% achieves the best or near-best loss on every domain simultaneously.

eliminating the advantage of the transfer matrix. At $\epsilon = 10\%$, Transfer-Aware actually underperforms Transfer-Naïve, confirming that an over-constrained floor removes the optimizer's ability to exploit transfer structure.

C Scaling-Law Clarifications and Objective Sensitivity

C.1 An empirical effective-data model

Our scaling law is an empirical surrogate for a fixed training setup. Its starting point is the shifted power-law form used to describe learning curves across data budgets [16, 33, 22]. We first fit each target domain's self-domain curve, then express training on another source as an equivalent target-domain data budget. Finally, we approximate the contributions of multiple sources by their sum. These steps motivate the surrogate; they do not establish a universal law of medical-image transfer.

Budget units and coefficient normalization. The budget T counts volume draws, rather than distinct subjects or distinct acquired volumes. To make the power-law argument dimensionless, let

$T_0 = 10{,}000$ volume draws and $u = T/T_0$. The self-domain law can then be written as

$$L_i(T) = E_i + C_i T^{-\beta_i} = E_i + \widetilde{C}_i u^{-\beta_i}, \qquad \widetilde{C}_i = \frac{C_i}{T_0^{\beta_i}}. \tag{14}$$

The coefficient changes with the budget convention; C_i and $\widetilde{C}_i$ must not be assigned the same numerical value. In the normalized form, E_i and $\widetilde{C}_i$ have the units of the evaluated loss, while u and β_i are dimensionless. The floor E_i and the reducible term need not have similar magnitudes. The three self-domain parameters are fitted jointly from the observed learning curves.

Directed transfer and the first-order mixture approximation. For a positive fitted transfer coefficient, a source-only run is modeled by replacing u with $\tau_{ij}u$ while retaining the target parameters $(E_i, \widetilde{C}_i, \beta_i)$. Thus τ_{ij} measures source-j training in target-i effective-data units, with $\tau_{ii} = 1$ and no symmetry requirement. For a mixture, the corresponding approximation is

$$\widehat{L}_i(\boldsymbol{h}, T) = E_i + \widetilde{C}_i \left(\frac{T}{T_0} \sum_{j=1}^{K} \tau_{ij} h_j \right)^{-\beta_i}. \tag{15}$$

A zero coefficient contributes no effective budget in this sum. A positive coverage floor and the self-transfer term ensure that the mixture's effective budget is positive. The additive sum is a first-order modeling assumption: it does not represent nonlinear synergy, interference, or changes in transfer induced by the mixture itself. These omitted effects are one reason to assess the surrogate against learned mixture baselines as well as allocation heuristics.

Self-domain curves and directed transfer coefficients are estimated from pure-source calibration measurements. For the extrapolation experiment, fitting uses $10K$, $50K$, and $100K$ budgets, with $200K$ held out. The predicted and observed mean losses at $200K$ are 0.0227 and 0.0224, respectively, with reported per-domain correlation $r = 0.989$. This checks extrapolation over the tested budget interval; it does not establish invariance to model capacity, training objective, or additional imaging domains.

C.2 Comparability of target losses

Here MAE abbreviates masked autoencoding; the reconstruction objective is per-patch mean squared error (MSE), as specified in Appendix A.3. We refer to the reported quantity as reconstruction loss throughout the following tables. The evaluated losses are computed after the modality-specific preprocessing in Appendix A.2, which maps all inputs to $[-1, 1]$. They are not computed directly in CT Hounsfield units, MRI scanner units, or PET standardized uptake values. A common numerical range still does not give every target the same semantic difficulty or clinical importance.

To examine sensitivity to target loss scale, we also optimize dimensionless relative-loss objectives at $T = 100K$. For each target i, define a positive reference loss s_i and minimize

$$J_{\mathrm{rel}}(\boldsymbol{h}, T) = \frac{1}{K} \sum_{i=1}^{K} \frac{\widehat{L}_i(\boldsymbol{h}, T)}{s_i}. \tag{16}$$

We use either the lowest target-i loss among pure-source runs at the same budget, $s_i = \min_j L_{i \leftarrow j}(T)$, or target i's loss under Uniform pretraining. The first reference need not be the target's own-domain run. Reference losses are fixed while optimizing the mixture.

Reference allocation for the supplementary analyses. The objective-sensitivity and bootstrap summaries use the reported reference allocation $\boldsymbol{h}_{\mathrm{ref}} = (26.7, 5, 5, 5, 5, 53.3)\%$. This is distinct from the primary allocation $(23.7, 5, 5, 5, 5, 56.3)\%$ used in Table 1. We retain these reported configurations separately: the sensitivity results and bootstrap intervals below refer to $\boldsymbol{h}_{\mathrm{ref}}$, and do not replace the primary effective-budget decomposition.

The sensitivity analysis shows that the two-domain concentration pattern persists under both tested normalizations, but it also shows that exact allocations depend on the objective. Equal target weighting is a default evaluation choice. The weighted objective in Eq. (13) allows application-specific priorities; neither equal weighting nor normalization establishes equal clinical value across targets.

Table 12: Objective sensitivity at $T = 100K$ with $\epsilon = 0.05$. All allocations are percentages. Cosine similarity is measured against the reference allocation in the first row. ABD-CT and HN-PET remain the two domains above the floor, although their exact weights change.

Objective	ABD CT	ABD MRI	Brain T1	Brain T2	HN CT	HN PET	Cosine Similarity
Mean loss (reference)	26.7	5.0	5.0	5.0	5.0	53.3	1.000
Relative to best pure source	6.0	5.0	5.0	5.0	5.0	74.0	0.928
Relative to Uniform	14.2	5.0	5.0	5.0	5.0	65.8	0.969

D Stability Across Seeds and Calibration Resamples

D.1 Replicated mixture-training results

Table 13 reports an additional three-seed summary for the four original mixture strategies. These means and standard deviations are a separate result summary from the primary reconstruction values in Table 2; several entries differ, and the two summaries are not pooled. Transfer-Aware has the lowest reported mean at every budget in this replicated summary. The variability nevertheless matters, especially for the small Transfer-Aware–Uniform difference at $200K$. A standard deviation describes run-to-run dispersion and is not a confidence interval or a paired significance test.

Table 13: Additional three-seed summary: mean reconstruction loss $\pm$ standard deviation. These entries are kept separate from the primary result summaries. Seed-level uncertainty is not reported for the added learned baselines.

Mixture	$10K$	$50K$	$100K$	$200K$
Uniform	0.0837 ± 0.0016	0.0533 ± 0.0138	0.0345 ± 0.0013	0.0227 ± 0.0075
Data-Proportional	0.0852 ± 0.0009	0.0639 ± 0.0073	0.0365 ± 0.0075	0.0283 ± 0.0043
Transfer-Naïve	0.0845 ± 0.0010	0.0696 ± 0.0102	0.0399 ± 0.0064	0.0283 ± 0.0050
Transfer-Aware	0.0812 ± 0.0011	0.0496 ± 0.0045	0.0264 ± 0.0031	0.0224 ± 0.0022

D.2 Bootstrap uncertainty in the optimized allocation

To propagate calibration uncertainty into the allocation, we use $1{,}000$ run-level bootstrap resamples, refit (E_i, C_i, β_i) and τ_{ij}, and solve the constrained mixture optimization again for each resample. The calibration uses three budget levels and two to three seeds per budget, giving six to nine measurements for each directed source–target configuration. Table 14 summarizes the resulting allocations for the supplementary reference configuration.

Table 14: Bootstrap allocation summary for $\boldsymbol{h}_{\mathrm{ref}}$, with $\epsilon = 0.05$. All values are percentages. The wide intervals indicate uncertainty in the precise optimum.

Domain	Reference weight	Bootstrap mean	Reported 95% interval
ABD-CT	26.7	25.4	$[5.0, 75.0]$
ABD-MRI	5.0	5.8	$[5.0, 13.7]$
BRAIN-T1	5.0	8.9	$[5.0, 27.5]$
BRAIN-T2	5.0	5.8	$[5.0, 15.7]$
HN-CT	5.0	6.3	$[5.0, 17.3]$
HN-PET	53.3	47.8	$[5.0, 73.9]$

HN-PET receives more than the floor in 95.6% of resamples and has the largest weight in 82.5%. ABD-CT is above the floor in 83.7%, and ABD-CT together with HN-PET forms the two highest-weight domains in 67.8%. The reported median cosine similarity of the allocation vectors is 0.96, and the reported median Spearman correlation of the transfer coefficients is 0.78. These summaries support a recurring allocation pattern more strongly than they support any exact weight. In particular, the broad intervals for ABD-CT and HN-PET caution against interpreting either point allocation as uniquely determined.

Role of the coverage floor. The floor is a tunable coverage constraint. The sweep in Appendix B.6 considers ϵ from 2.5% to 12.5%, with the lowest observed mean reconstruction loss at 5% in this setting. Lower floors permit greater concentration, while higher floors reduce the optimizer's freedom to exploit incoming transfer. This empirical choice is not a universal recommendation: suitable coverage depends on the domain count and the intended downstream uses.

E Expanded Learned-Baseline and Downstream Comparisons

E.1 Learned mixture baselines

We extend the comparison with RegMix [28] and Data Mixing Laws [48]. In the RegMix adaptation, anchored and Dirichlet mixtures are sampled subject to the same 5% allocation floor. Each mixture is trained with the same 3D MAE proxy and evaluated on all six targets. One LightGBM regressor is fitted per target, and the mixture is selected by minimizing the mean predicted target loss. Data Mixing Laws models the joint effect of mixture proportions and training-data scale. Transfer-Aware instead separates target-specific scaling parameters from directed source-to-target effective-data coefficients.

Table 15 retains the primary reconstruction summaries for the original strategies and adds the learned baselines. Transfer-Aware is approximately 14.3% and 19.5% lower than RegMix at $10K$ and $100K$, respectively, using RegMix's loss as the denominator. RegMix is approximately 3% lower at $50K$, and the two methods are tied at the reported precision at $200K$. These results support competitive performance, rather than uniform dominance over a flexible learned mixture predictor. Equal final pretraining budgets do not, by themselves, establish equal total calibration cost.

Table 15: Primary mean reconstruction loss across six target domains, with learned baselines added. Lower is better. This table is separate from the three-seed summary in Table 13.

Mixture	$10K$	$50K$	$100K$	$200K$
Uniform	0.0853	0.0671	0.0345	0.0228
Data-Proportional	0.0843	0.0712	0.0365	0.0233
Transfer-Naïve	0.0835	0.0799	0.0399	0.0283
Data Mixing Laws	0.0855	0.0739	0.0395	0.0269
RegMix	0.09346	**0.04367**	0.03278	**0.0224**
Transfer-Aware	**0.0801**	0.0451	**0.0264**	**0.0224**

Interpretation of Transfer-Naïve. Transfer-Naïve keeps the same fitted self-domain parameters but replaces the transfer matrix with the identity, $\tau_{ij} = \delta_{ij}$. It is a deliberately misspecified ablation of directed transfer, rather than a recommended allocation rule. In heterogeneous settings it may allocate poorly because it omits contributions from shared sources. Its gap relative to Transfer-Aware shows the consequence of that omission within this surrogate family. The comparisons with RegMix and Data Mixing Laws provide a complementary assessment against learned alternatives. RegMix can represent mixture-to-loss relationships without the explicit additive effective-budget decomposition; conversely, the Transfer-Aware parameters expose the source–target structure used to choose an allocation. Combining these approaches remains a possible extension, not an evaluated result.

E.2 Complete downstream comparisons

Tables 16 and 17 report all six mixture strategies at the $100K$ pretraining budget. Accuracy is shown outside parentheses and AUC inside parentheses. Task protocols and downstream architectures are held fixed across mixture strategies. The tables report point estimates without additional significance annotations.

Transfer-Aware has the highest Dice on all six segmentation evaluations, exceeding RegMix by 0.8, 0.5, 2.1, 2.3, 1.8, and 1.9 points, respectively, for a mean difference of 1.57 points. It also has the highest mean classification accuracy and the highest accuracy on OASIS-1 and SynBrain-PET. RegMix performs better on OrganMNIST3D accuracy and OASIS-1 AUC, while RegMix and Transfer-Naïve have the highest reported OrganMNIST3D AUC. Thus reconstruction quality does not determine every downstream ranking.

Table 16: Classification after $100K$ mixture pretraining: accuracy (AUC), in percent. The final column is the unweighted mean accuracy across the three tasks. Higher is better. Rankings differ between accuracy and AUC.

Mixture	OASIS-1	SynBrain-PET	OrganMNIST3D	Mean acc.
Uniform	66.9 (73.6)	69.7 (94.5)	88.9 (99.4)	75.17
Data-Proportional	67.5 (76.0)	71.9 (93.9)	89.7 (99.4)	76.37
Transfer-Naïve	66.3 (76.3)	73.6 (95.3)	90.1 (99.5)	76.67
Data Mixing Laws	67.5 (79.3)	73.2 (94.2)	89.8 (99.4)	76.83
RegMix	68.8 (**79.7**)	74.6 (95.3)	**91.5** (**99.5**)	78.30
Transfer-Aware	**69.7** (78.4)	**78.1** (**95.9**)	90.3 (99.4)	**79.37**

Table 17: Segmentation Dice (%) after $100K$ mixture pretraining. The final column is the unweighted mean over the six evaluations. Higher is better. RegMix entries use the completed six-task evaluation.

Mixture	AMOS CT	AMOS MRI	BraTS T1	BraTS T2	HECKTOR	BHSD	Mean
Uniform	81.6	68.8	50.4	61.4	67.8	57.2	64.53
Data-Proportional	82.0	68.8	52.7	63.0	68.7	59.0	65.70
Transfer-Naïve	81.3	66.9	50.5	61.1	68.0	57.5	64.22
Data Mixing Laws	80.9	67.9	50.1	61.0	68.5	58.5	64.48
RegMix	82.3	68.9	53.0	63.1	69.3	59.3	65.98
Transfer-Aware	**83.1**	**69.4**	**55.1**	**65.4**	**71.1**	**61.2**	**67.55**

For a descriptive assessment, we pair each of the six methods' $100K$ reconstruction losses in Table 15 with its unweighted mean downstream score. The Spearman rank correlations are $\rho = -0.943$ for mean Dice and $\rho = -0.486$ for mean classification accuracy, computed from the displayed aggregate results. These are six method-level observations, not subject-level or seed-level estimates. They summarize association within this comparison and do not establish a causal relationship or a clinically calibrated proxy. Averaging across tasks likewise provides a compact descriptive summary, rather than a common scale of clinical utility.

E.3 Budget dependence and relative-gap conventions

The relative excess loss of a baseline over Transfer-Aware is

$$g_b = \frac{L_b - L_{\mathrm{TA}}}{L_{\mathrm{TA}}}, \qquad r_b = \frac{L_b - L_{\mathrm{TA}}}{L_b} = \frac{g_b}{1 + g_b}, \tag{17}$$

where r_b is the conventional loss reduction relative to the baseline. For example, at $50K$, the primary Data-Proportional and Transfer-Aware losses are 0.0712 and 0.0451. Data-Proportional is 57.9% higher than Transfer-Aware (g_b), equivalently a 36.7% reduction for Transfer-Aware relative to Data-Proportional (r_b). At $100K$, the corresponding quantities are 38.3% and 27.7%. These denominators should not be interchanged.

The primary excess-loss gaps against Data-Proportional are 5.2%, 57.9%, 38.3%, and 4.0% at $10K$, $50K$, $100K$, and $200K$. The largest advantage is therefore in the intermediate-budget regime. The narrowing gap at $200K$ is consistent with diminishing returns for the fixed proxy, potentially from its loss floor or capacity, but does not identify which mechanism is responsible or establish a universal saturation point.

F Transfer Interpretation, Related Work, and Scope

F.1 Imaging heterogeneity and the meaning of the fitted parameters

Table 18 gives the reported native imaging summaries, before the shared spatial preprocessing and modality-specific intensity normalization. They illustrate differences in acquisition resolution, volume geometry, and intensity conventions. The source summaries do not specify a population-level aggregation statistic; they should not be interpreted as universal acquisition settings or population-wide intensity bounds.

Table 18: Reported native imaging summaries. These descriptive values precede resampling and intensity normalization, and are not the ranges used to evaluate reconstruction loss. MRI values are in arbitrary units (a.u.); SUV denotes standardized uptake value.

Domain	Spacing (mm)	Matrix size	Native intensity summary
ABD-CT	$0.67 \times 0.67 \times 2.0$	$512 \times 512 \times 201$	$[-1018, 439]$ HU
ABD-MRI	$1.95 \times 1.95 \times 7.85$	$192 \times 192 \times 24$	$[0, 1.6 \times 10^5]$ a.u.
BRAIN-T1	$1 \times 1 \times 1$	$240 \times 240 \times 155$	$[0, 1943]$ a.u.
BRAIN-T2	$1 \times 1 \times 1$	$240 \times 240 \times 155$	$[0, 1299]$ a.u.
HN-CT	$0.98 \times 0.98 \times 1.5$	$512 \times 512 \times 238$	$[-1000, 244]$ HU
HN-PET	$4 \times 4 \times 4$	$144 \times 144 \times 87$	$[0, 2.3]$ SUV

Scaling exponents describe a proxy learning curve. For the fitted self-domain law, β_i is the rate of relative decay of the reducible reconstruction loss: $\partial \log(L_i - E_i)/\partial \log T = -\beta_i$. A smaller exponent means slower relative decay, rather than proving that a domain is anatomically simple or rapidly saturated. Loss levels depend on E_i, C_i, β_i, and the budget. For fixed fitted parameters, the additive floor E_i cancels from marginal loss reductions and mixture comparisons; the allocation depends on C_i, β_i, the transfer coefficients, target weights, and coverage constraints. It is not determined by an anatomical interpretation of β_i alone. Likewise, "hub" and "island" describe estimated outgoing and incoming transfer in the evaluated setup; they are not intrinsic clinical properties of a modality.

F.2 Relationship to Transfusion

Transfusion [32] studies when transfer learning from natural images benefits medical-image tasks, and analyzes the usefulness of the transferred representations. Its findings motivate evaluating transfer empirically instead of assuming that pretraining automatically improves medical performance. Our study shares that concern but addresses a different allocation question: under a fixed self-supervised budget, how should training be distributed across medical domains whose directed transfer effects differ?

Table 19: Positioning relative to Transfusion [32]. The questions are complementary; our contribution is a quantitative mixture-allocation model in the studied 3D setting.

Aspect	Transfusion	This study
Transfer setting	Supervised ImageNet initialization for 2D retinal and chest X-ray classification.	Self-supervised pretraining across 3D CT, MRI, and PET domains.
Training question	When and how transfer learning benefits a target medical task.	How to allocate a fixed pretraining budget across multiple medical sources and targets.
Analysis	Transfer utility and reuse of pretrained representations.	Target-specific loss scaling and directed effective-data coefficients used for mixture optimization.
Evaluation role	Assess the value of transferred representations on medical tasks.	Assess proxy-loss prediction and the downstream usefulness of the selected mixture.

The distinction is therefore not that medical transfer itself has been unstudied. We estimate a directional source–target surrogate and use it to make a budget-dependent allocation decision. Downstream segmentation and classification provide evidence about the usefulness of that decision, while the limitations of reconstruction as a proxy remain relevant to both the interpretation and deployment of the resulting representations.

F.3 Clinical and scaling limits

The six-domain experiments use controlled subsets of 150–500 volumes per domain. They establish evidence in this heterogeneous 3D setting, rather than across arbitrary clinical corpora, institutions, patient populations, or training objectives. Dice evaluates the overlap of predicted anatomical or lesion regions and therefore adds structured task evidence beyond reconstruction loss. It does not directly validate topology, surface consistency, preservation of small lesions, or clinical decision quality. No direct anatomical-consistency analysis is established by these aggregate results. Such

evaluation, together with qualitative prediction review, remains necessary before stronger anatomical or clinical claims are made.

The small proxy and larger downstream pretraining model serve different roles, as discussed in Appendix A.3. Transferring a selected allocation to the larger model does not demonstrate that every coefficient τ_{ij} is invariant to capacity. A full transfer-matrix refit across multiple model sizes has not been evaluated. Jointly modeling $L(N, \boldsymbol{h}, T)$, or allowing transfer coefficients to depend on scale, remains future work.

Increasing the domain count. There are K self-domain and $K(K-1)$ directed source–target evaluation configurations, or K^2 in total: 36 at $K = 6$, 400 at $K = 20$, and 2,500 at $K = 50$, before accounting for budgets and seeds. These are evaluation configurations, not necessarily distinct training jobs, since evaluating one source-trained model on multiple targets could permit reuse. The quadratic number of directed relationships remains a limitation. Sparse, low-rank, hierarchical, or uncertainty-guided calibration could reduce the required measurements, but these extensions are not evaluated here. The coverage constraint must also satisfy $K\epsilon \leq 1$: a fixed 5% floor leaves only the uniform mixture at $K = 20$ and is infeasible at $K = 50$. Larger-domain extensions must adapt the floor or use other coverage constraints.

F.4 Code and implementation resources

The accompanying repository is

`https://anonymous.4open.science/r/transfer-aware-scaling-3d-med-4564`.

The implementation resources cover scaling-law fitting, mixture optimization, bootstrap analysis, preprocessing, and the pretraining and downstream finetuning workflow, with usage instructions. Medical datasets remain subject to their original access conditions and licenses and are obtained from their respective sources. The repository and experimental details should be used together to distinguish calibration budgets, the primary mixture results, and the supplementary stability summaries.

G Compute Cost

The exhaustive calibration design covers $K^2 = 36$ source-target evaluation configurations for six domains, with multiple budgets and seeds. The 12M-parameter proxy is used for calibration; the approximately 121M-parameter target model is pretrained with the selected mixture for downstream evaluation. A proxy training run is reported to require approximately 10–20 GPU-hours on a single NVIDIA A100, depending on domain and run configuration. Once the law is fitted, optimizing a new mixture takes less than one second on a single CPU core using 100 SLSQP restarts. A target-model pretraining run at $T = 200K$ is reported to require approximately three days. These are per-run timings, rather than an aggregate accounting of all calibration, baseline, and downstream experiments. The quadratic number of source-target configurations and the additional budgets and seeds limit direct expansion to many domains; see Appendix F.

H Broader Impact

By replacing exhaustive mixture search with principled scaling-law optimization, our framework reduces the compute cost of training 3D medical foundation models, lowering the barrier for resource-constrained institutions. However, the framework optimizes a self-supervised proxy (MAE loss) rather than clinical outcomes directly. Practitioners should validate downstream performance on their specific clinical tasks before deployment. Additionally, the optimized mixture inherits any biases present in the source datasets (i.e If certain patient populations are underrepresented in specific imaging domains), the allocation will not address this imbalance. Incorporating fairness-aware objectives into the mixture optimization is an important direction for future work.