

Vision Foundation Models for Multi-Modal Medical Image Classification

Daniel Scholz

`scholz@cs.tum.edu`

Technical University of Munich

Ayhan Can Erdur

Technical University of Munich

Camillo Saueressig

TUM Klinikum

Esther Utasa

TUM Klinikum

Jan C. Peeken

TUM Klinikum

Daniel Rueckert

Technical University of Munich

Benedikt Wiestler

TUM Klinikum

Article

Keywords:

Posted Date: October 23rd, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11204482/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

VISION FOUNDATION MODELS FOR MULTI-MODAL MEDICAL IMAGE CLASSIFICATION

Daniel Scholz^{1,2,3,*}, Ayhan Can Erdur^{3,4}, Camillo Saueressig¹, Esther Utasa¹, Jan C. Peeken^{4,5,6}, Daniel Rueckert^{2,3,†}, Benedikt Wiestler^{1,2,*,†}

¹Chair for AI for Image-Guided Diagnosis and Therapy, Technical University of Munich (TUM) and TUM University Hospital, Munich, Germany

²Munich Center for Machine Learning (MCML), Munich, Germany

³Chair for AI in Healthcare and Medicine, Technical University of Munich (TUM) and TUM University Hospital, Munich, Germany

⁴Department of Radiation Oncology, TUM University Hospital, Munich, Germany

⁵Deutsches Konsortium für Translationale Krebsforschung (DKTK), Partner Site Munich, Munich, Germany

⁶Institute of Radiation Medicine (IRM), Department of Radiation Sciences (DRS), Helmholtz Center Munich, Munich, Germany

[†]D.R. and B.W. contributed equally as senior authors.

*Corresponding authors: Daniel Scholz (scholz@cs.tum.edu) and Benedikt Wiestler (b.wiestler@tum.de)

ABSTRACT

Clinicians combine several imaging modalities to reach a diagnosis. Yet assembling large annotated datasets for every modality combination is rarely feasible, and modalities are frequently missing in practice. Vision foundation models (FMs) could ease both problems by transferring from few labels, but this has mainly been shown for single images. We therefore compare FM initialisation (DINOv2 and MedSigLIP) against random initialisation, across fusion and training strategies, on three tasks: glioma subtyping from multi-sequence MRI, HER2 scoring from paired histopathology stains, and liver-lesion classification from multi-phase MRI. At full data, FM initialisation matched or outperformed random initialisation, raising liver-lesion sensitivity from 54 % to 76 %. The advantage held as labels grew scarce and was largest under missing modalities, where FM initialisation kept sensitivity near 73 % as random initialisation fell to 29 %. For multi-modal medical imaging, FM initialisation is a strong default, most valuable when labels are scarce or acquisition protocols incomplete.

1 Introduction

Many diagnostic tasks in medicine rely on several complementary types of imaging, which we refer to as multi-modal imaging [1, 2]. Examples span multi-sequence MRI for glioma subtyping [3], paired histopathology stains for HER2 scoring [4], and multi-phase MRI for liver-lesion classification [5]. Typically, a separate task-specific deep learning model is trained from random initialisation for each modality combination [6, 7, 8, 9]. However, annotated data is scarce in medicine [10], which makes it impractical to train a dedicated model for every new combination of sequences, phases, or stains. Acquisitions are also frequently incomplete, with modalities missing at inference because patients undergo different examinations and acquisition protocols vary across sites [11, 12, 13]. Missing modalities are the norm rather than the exception in clinical practice. When labels are scarce or a modality is missing, a less accurate model can lead to missed diagnoses.

Vision foundation models (FMs), such as DINOv2 [14], offer an alternative. Trained on large unlabelled image corpora, they learn general-purpose visual features that transfer to downstream tasks with few labels. In medical imaging, FMs often perform as well as task-specific networks [15] and remain accurate under shifts between sites, scanners, and cohorts [16, 17].

However, no foundation model is readily available for a new combination of imaging modalities. Medical-domain FMs operate on only a single image [16, 18, 19, 20, 21]. The few multi-modal medical FMs either pair one imaging modality with text [22, 23, 24] or are tied to a single anatomical region [25, 26]. Given these restrictions, building a

multi-modal imaging classifier from an FM amounts to adapting single-image models with a fusion step. Because this adaptation is difficult, most multi-modal medical-imaging models still start from random initialisations, with special-purpose methods for missing modalities [27, 28, 29, 30, 11, 31]. Surveys of the field organise this literature around the choice of fusion scheme for task-specific networks trained from random initialisation [32]. Prior work shows this adaptation is possible [33, 34], but no study has yet systematically tested whether it outperforms random initialisation, and whether the answer holds across fusion strategies and training strategies. Nor has any study tested whether the gains hold under label or modality scarcity. Our question is therefore separate from the choice of fusion scheme: we hold fusion as one axis and vary initialisation, rather than proposing a new fusion scheme.

Here, we adapt general and medical vision FMs into multi-modal medical-imaging classifiers. We compare three initialisations: random, a general-purpose vision FM (DINOv2 [14]), and a medical vision FM (MedSigLIP [21]). We test each under two fusion strategies: early fusion with modality-aware tokens and attention across modalities, or late fusion with one backbone per modality. Each is trained with supervised fine-tuning, semi-supervised training, or linear probing. We evaluate on three clinically distinct tasks: glioma subtyping from multi-sequence brain MRI, HER2 scoring from paired multi-stain histopathology (BCI) [4], and liver-lesion classification from multi-phase abdominal MRI (LLD-MMRI) [5]. On every task, we test performance with all modalities, with only a subset of the training labels, and with a modality missing at inference. On the glioma subtyping task, where several cohorts are available, we additionally test generalisation to an unseen external site.

Our results show that FM initialisation, combined with appropriate fusion and semi-supervised adaptation, should be the default choice for multi-modal medical-imaging classifiers. This holds even without a large paired pretraining corpus, and especially when labels are scarce or modalities are missing.

2 Results

We evaluate foundation-model-initialised against randomly initialised models on three multi-modal imaging tasks (§4.1) (Figure 1): glioma subtyping on multi-sequence brain MRI, HER2 scoring on paired haematoxylin and eosin (H&E) and immunohistochemistry (IHC) histopathology (BCI), and seven-class lesion classification on multi-phase abdominal MRI (LLD-MMRI). We compare the models across four experiments, each representing a common challenge in clinical practice. The full-data experiment (§2.1) sets the main comparison between foundation-model and random initialisation. The label-efficiency experiment (§2.2) tests how model performance degrades in limited-label scenarios, which are common in clinical reality. Individual modalities are often missing in clinical reality, so we test the missing-modality robustness (§2.3) of each setting by dropping one modality at inference. When deployed, models must generalise across sites, and MRI in particular varies between manufacturers and hospitals [35]. We therefore evaluate cross-site generalisation (§2.4) on an out-of-distribution hospital in the glioma dataset.

We compare three vision transformer (ViT) initialisations (random, DINOv2 [14], MedSigLIP [21]) across training strategy (supervised fine-tuning (FT), semi-supervised (SSL), linear probing) and fusion (early channel or modality-token concatenation, late per-modality fusion). Within each configuration of training strategy and fusion we compare initialisations, which isolates the effect of foundation-model (FM) versus random weights. We name each model by its initialisation, fusion, and training strategy, written for example as (MedSigLIP, late fusion, SSL). A randomly initialised ResNet [36], an established architecture for medical image classification [37], serves as an additional baseline. We also compare the best randomly initialised model against the best FM-initialised model, selected on validation MCC (§4.3, Supplementary Table 1).

We report performance as the Matthews correlation coefficient (MCC) [38, 39], a balanced classifier metric on $[-1, +1]$ with 0 at chance-level prediction and +1 at perfect classification (§4.8). MCC is especially suitable for imbalanced scenarios such as glioma subtyping, where some subtypes occur more commonly than others [40]. We also report the macro-averaged area under the receiver operating characteristic curve (AUROC) as a secondary metric. AUROC and sensitivity views of every experiment are provided in the Supplementary Information (Supplementary Figures 1–8). Error bars in all figures are 95 % bootstrap confidence intervals.

2.1 FM initialisation outperforms random initialisation at full data

We train every (training strategy, fusion) configuration on the full training set, using all images and the complete label set where training is supervised, and report MCC on the held-out test set. With full labels, FM initialisation beats random initialisation within all four matched configurations on all three datasets (Figure 2).

On LLD-MMRI, every FM-initialised ViT beats its random-initialised counterpart, with within-group gains from +0.17 to +0.36 MCC across the four (training strategy, fusion) groups. Best-versus-best, the best FM configuration (MedSigLIP, late fusion, SSL) beats the strongest random-initialisation baseline, ResNet-random (FT-early), by

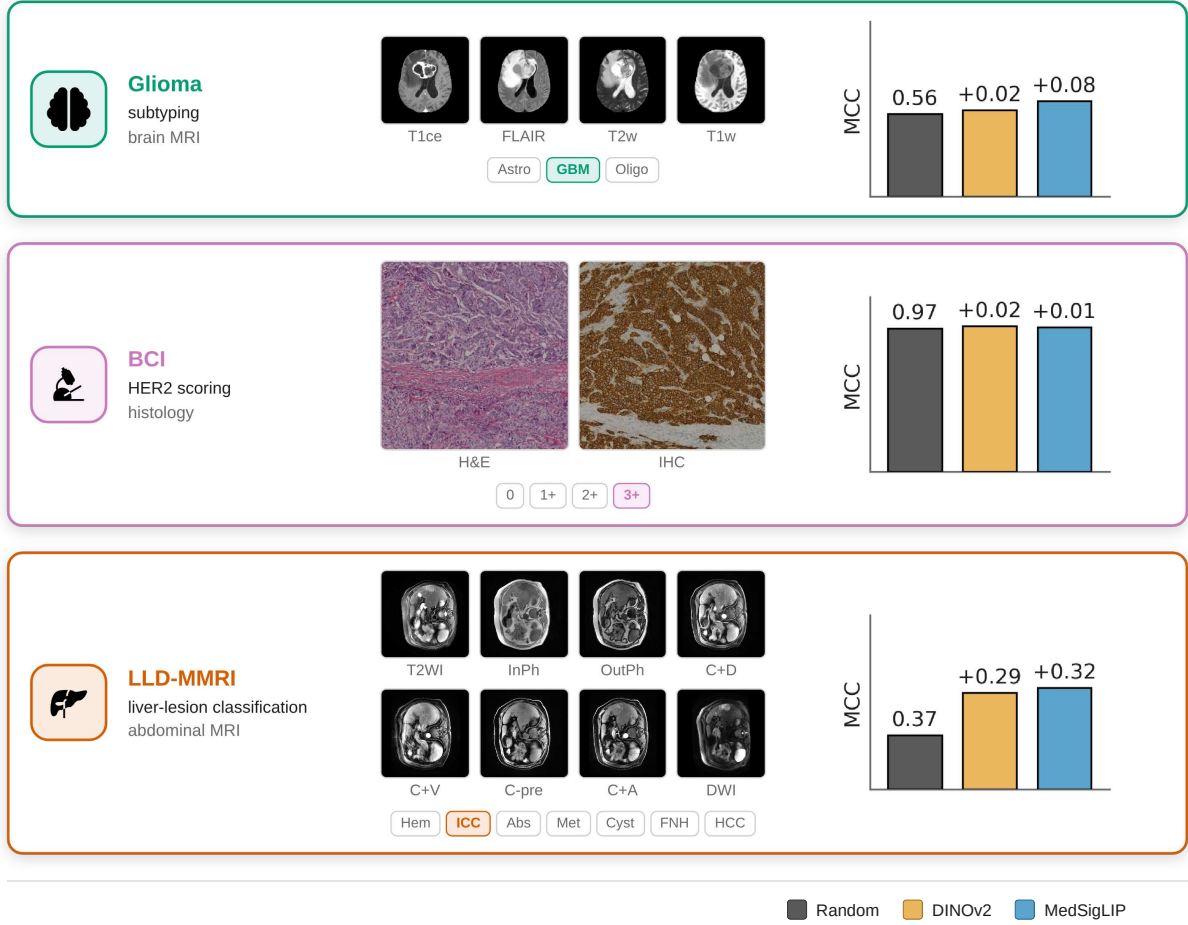


Figure 1: The three multi-modal imaging benchmarks, one per row, each with its full set of imaging modalities: glioma subtyping (brain MRI), HER2 scoring (breast-cancer histology, BCI), and liver-lesion classification (abdominal MRI, LLD-MMRI). The class boxes below each example list the task categories, with the shown case’s true label highlighted. Bars give the best Matthews correlation coefficient (MCC) for random, DINOv2, and MedSigLIP initialisation. The random bar shows its absolute MCC, and each foundation-model bar its improvement over random. Foundation-model initialisation matches or exceeds random initialisation on all three benchmarks, with the largest margin on LLD-MMRI. Icons: Healthicons (CC0) and Font Awesome (CC BY 4.0).

82 Δ MCC +0.21 (FDR-corrected $p < 0.01$). The same comparison raises mean per-class sensitivity from 0.54 to 0.76
 83 (Supplementary Figure 1b).

84 On BCI, every FM-initialised ViT beats its random-initialised counterpart. Best-versus-best, the strongest random-
 85 initialisation baseline is semi-supervised ViT-random with late fusion, and the best FM configuration leads it by only
 86 +0.02 MCC, short of significance (FDR-corrected $p = 0.10$).

87 On glioma, every FM-initialised ViT beats its random-initialised counterpart, with within-group gains from +0.03 in
 88 the supervised fine-tuning configurations to +0.29 in SSL-early (semi-supervised, early fusion). In SSL-early the
 89 DINOv2 bar, the early-fusion DINOv2 semi-supervised model (MM-DINOv2 [33]), pulls clear of the random one.
 90 Every FM-initialised SSL configuration matches or exceeds the strong ResNet-random baseline (MCC 0.64), and the
 91 best FM and best random-initialisation model are statistically tied at full labels. On glioma and BCI, where MCC is
 92 matched, sensitivity is comparable too, at 0.75 versus 0.68 on glioma and above 0.97 for both models on BCI.

93 Across the figure, FM initialisation wins all 12 within-configuration comparisons, the gap widest under semi-
 94 supervised training and narrowest under fine-tuning. Best-versus-best, the FM lead reaches significance only on
 95 LLD-MMRI, and on glioma and BCI the two are matched.

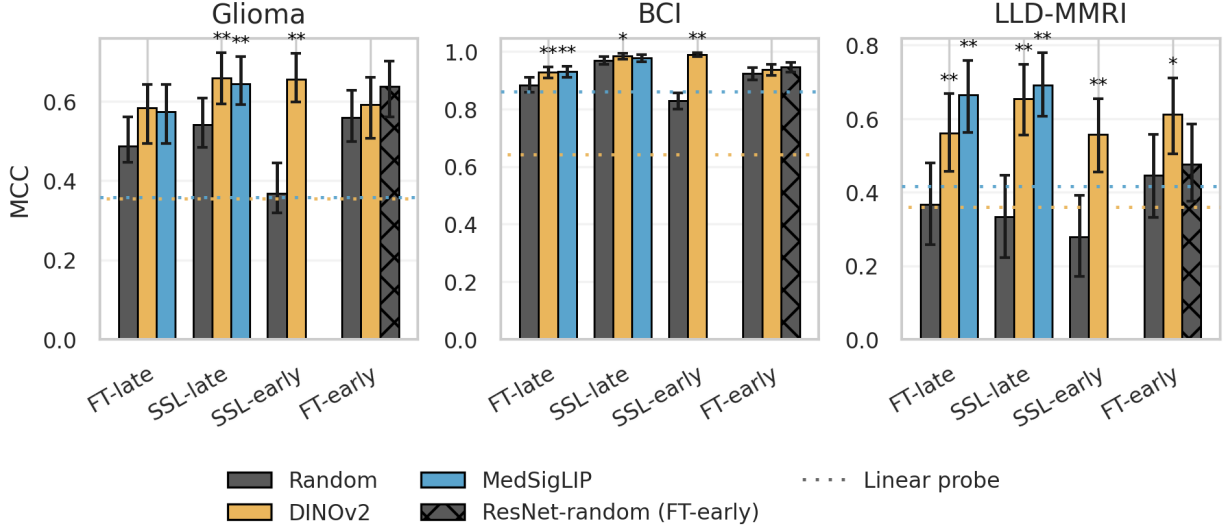


Figure 2: Matthews correlation coefficient (MCC) on the three benchmarks. Each panel is one dataset: glioma subtyping, breast-cancer histology (BCI), and liver-lesion MRI (LLD-MMRI). The x-axis groups runs by (training strategy, fusion) configuration, and the coloured bars in each group contrast the random, DINOv2 [14], and MedSigLIP [21] initialisations, so the foundation-model-vs-random effect reads off one group at a time. MedSigLIP has no early-fusion configuration, so only the random and DINOv2 bars are drawn in those cells. The cross-hatched bar, shown only in the FT-early (supervised fine-tuning, early fusion) column, is the ResNet-random baseline, and horizontal dotted lines mark the DINOv2 and MedSigLIP linear-probe baselines. Error bars are 95 % bootstrap confidence intervals. Stars above a foundation-model bar mark Benjamini–Hochberg-corrected significance of that bar against its matched random-initialisation bar in the same group (* $p < 0.05$, ** $p < 0.01$).

Beyond full labels, the FM advantage persists down to low label budgets (§2.2) and widens sharply under missing modalities (§2.3).

2.2 FM initialisation keeps its advantage at low label budgets

We subsample the training labels to 10 %, 25 %, 50 %, and 100 % and retrain every (training strategy, fusion) configuration at each fraction (§4.3). The 100 % fraction corresponds to the full-data setting of §2.1.

Within every matched (training strategy, fusion) configuration, foundation-model initialisation improves on random initialisation on all three datasets across the label fractions (Figure 3). The gain is significant in most settings and on par in the others, and no configuration falls significantly below its random-initialisation counterpart at any fraction. The advantage holds down to the smallest label budgets, and it is largest and most consistent on LLD-MMRI.

On LLD-MMRI, the FM-initialisation advantage holds at every labelled fraction. MedSigLIP late-fusion SSL gains Δ MCC +0.27 at 10 % ($p = 0.001$), +0.25 at 25 % ($p = 0.002$), and +0.35 at 50 % ($p < 0.001$) over its random-initialisation counterpart. The same pattern holds under fine-tuning, where DINOv2 late-fusion FT significantly gains between +0.20 and +0.34 MCC across the reduced fractions. On the best-versus-best axis, selected at each fraction (Supplementary Table 5), the lead persists at every reduced fraction, where the best FM beats the strongest random-initialisation model by +0.23 at 10 %, +0.21 at 25 %, and +0.37 at 50 % (all significant).

On BCI, the FM gain is large under fine-tuning and early-fusion semi-supervised training. MedSigLIP late-fusion FT gains Δ MCC +0.40 at 10 %, +0.40 at 25 %, and +0.38 at 50 % over its random-initialisation counterpart, all significant. DINOv2 early-fusion SSL (MM-DINOv2) gains +0.42 at 10 % down to +0.21 at 50 %, also significant at every fraction. The gain is largest at small budgets and narrows as labels grow. Under late-fusion SSL the FM matches its random-initialisation counterpart. On the best-versus-best axis the FM is nominally ahead at every fraction (+0.01 to +0.04 MCC) but significant only at 25 %, in line with the otherwise near-tie at full data.

On glioma, the FM-initialisation advantage holds from 25 % labels upward but not at the smallest budget. MedSigLIP late-fusion SSL gains Δ MCC +0.15 at 25 % ($p < 0.001$) and +0.14 at 50 % ($p = 0.001$) over its random-initialisation counterpart, both significant. At 10 % labels the two are statistically tied (Δ MCC -0.06). Under fine-tuning the

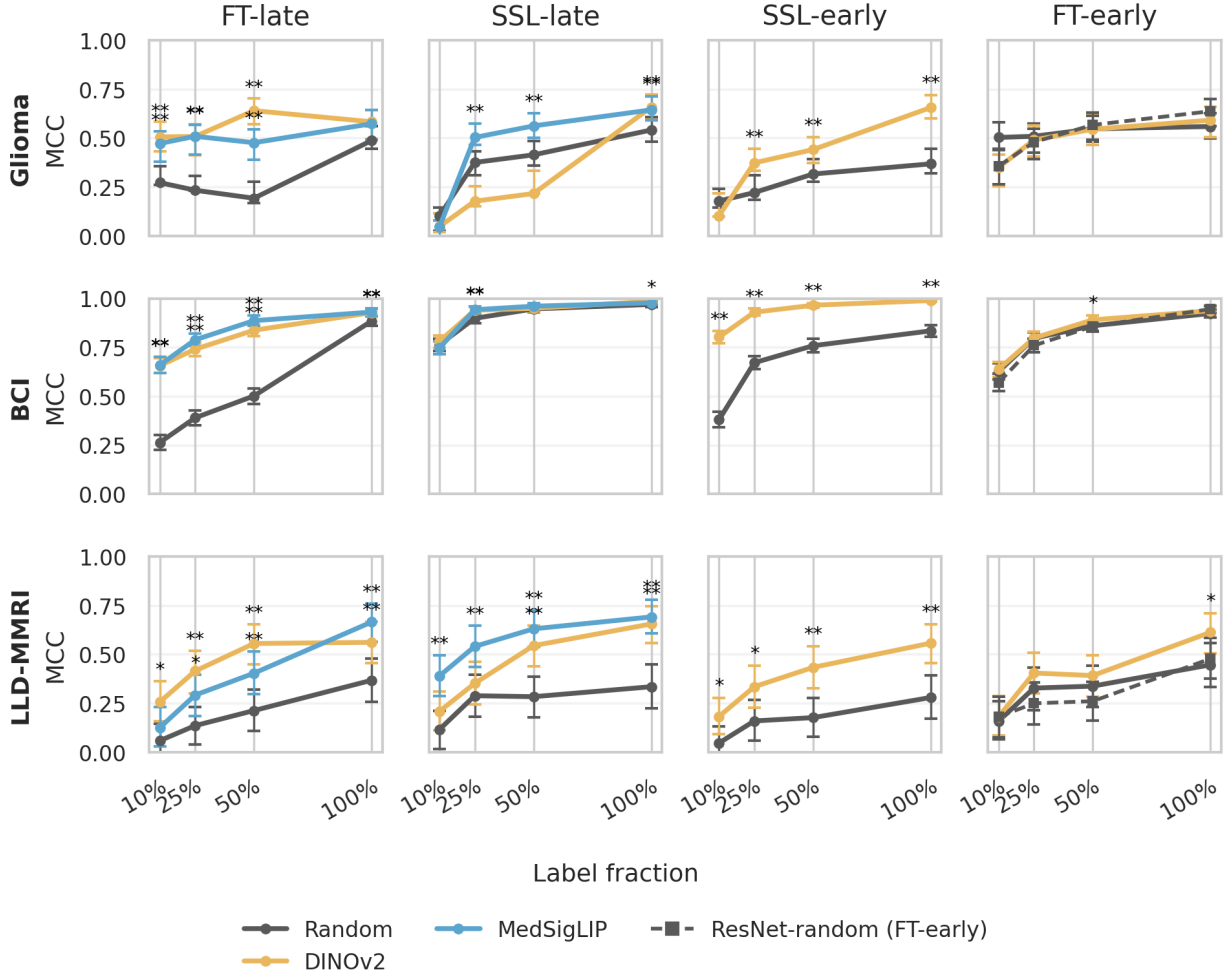


Figure 3: Label efficiency of foundation-model versus random initialisation. MCC across four labelled fractions on the three benchmarks, with datasets along the rows and (training strategy, fusion) configurations along the columns. The x-axis of each panel is the labelled fraction, and coloured lines contrast the three initialisations. The dashed line with square markers in FT-early is the ResNet-random baseline. MedSigLIP has no early-fusion configuration, so only the random and DINOv2 curves are drawn in those cells. Error bars are 95 % bootstrap confidence intervals. Stars mark Benjamini-Hochberg-corrected significance of each foundation-model curve against the randomly initialised curve at the same labelled fraction (* $p < 0.05$, ** $p < 0.01$).

advantage reaches even the smallest budget, where DINOv2 late-fusion FT gains +0.20 MCC ($p=0.001$). On the best-versus-best axis the two are tied at every fraction (-0.002 to $+0.022$ MCC, none significant). The FM advantage on glioma grows further under modality dropout (§2.3), and on the out-of-distribution cohort it concentrates in the semi-supervised configurations (§2.4).

Across the three datasets the direction is consistent: foundation-model initialisation is significantly better in most settings and never significantly worse. How large the gain is, and the smallest budget at which it reaches significance, depend on the dataset and the training strategy.

2.3 FM initialisation improves robustness to missing modalities

We evaluate every trained (training strategy, fusion) configuration with one modality withheld at inference and no re-training, to test feature robustness (Figure 4).

Foundation-model initialisation stays close to its full-input performance under modality dropout, while random initialisation often collapses. The gap is significant on all three datasets, both within matched configurations and best-versus-best.

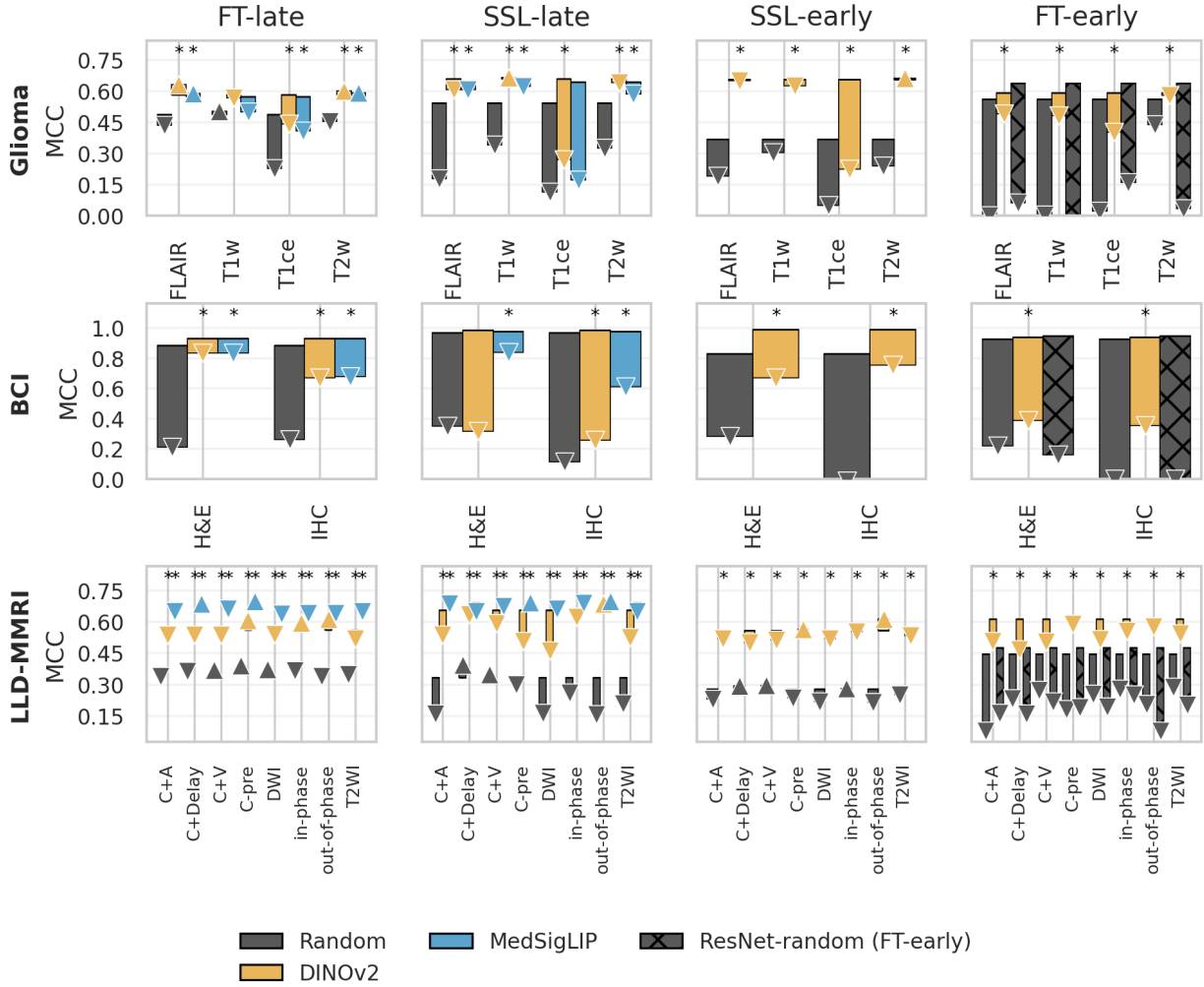


Figure 4: MCC under single-modality dropout on all three datasets. Panels form a 3×4 grid: datasets along the rows, (training strategy, fusion) configurations along the columns. The x-axis of each panel groups bars by masked modality. Each panel uses floating bars. For every method a bar spans the y-axis between its full-input MCC and its MCC with one modality masked at inference (no re-training), so the bar height is the size of the change. A short horizontal tick marks the full-input baseline, and a triangle marks the masked value, pointing down for a drop and up for a gain. Most bars point down, the loss incurred by masking. MedSigLIP has no early-fusion configuration, so only the random and DINOv2 bars are drawn in those cells. The cross-hatched bar in FT-early is the ResNet-random baseline. Stars in the top row of each panel mark Benjamin-Hochberg-corrected significance of each foundation-model initialisation against the randomly initialised reference for that masked modality (* $p < 0.05$, ** $p < 0.01$). The matching AUROC view is in Supplementary Figure 4.

On LLD-MMRI (one of eight MRI phases masked), the randomly initialised FT-early ViT loses half of its full-input MCC under single-phase dropout, its mean across the eight masks falling from 0.45 to 0.22, while MedSigLIP late-fusion SSL is essentially invariant. The per-mask FM-vs-random gain is significant on every mask for both DINOv2 and MedSigLIP and ranges from +0.21 to +0.54 MCC. For sensitivity, the gap widens under masking: MedSigLIP late fusion holds a mean of 0.73 across the eight masks, down only from 0.76 at full input, while the randomly initialised baseline falls to 0.29 from 0.54 (Supplementary Figure 6c). The same drops expressed as relative percentages, for every method and both metrics, are in Supplementary Figure 7.

On BCI (mask H&E or IHC), the randomly initialised ViT leans on IHC, collapsing to MCC=0.00 when IHC is removed, and still drops to 0.22 when H&E is removed. MedSigLIP late-fusion FT gains +0.42 MCC under IHC masking and MedSigLIP late-fusion SSL gains +0.49 MCC under H&E masking over its random-initialisation counterpart, both significant. In sensitivity the collapse is starker. Removing IHC drops the randomly initialised ViT to 0.25, the chance rate for four balanced classes, while MedSigLIP late-fusion FT still detects 0.72 of cases.

On glioma (mask one of FLAIR, T1w, T1ce, T2w), the randomly initialised ViT drops by 0.12 to 0.56 MCC depending on the masked sequence. On three of the four sequences it falls to near zero. The best-per-mask FM advantage is +0.37 to +0.49 MCC and is significant on every mask for at least one FM initialisation.

The gap holds best-versus-best as well, with representatives selected per masked modality (Supplementary Tables 2–4). The best FM leads significantly under every mask, by +0.11 to +0.17 MCC on glioma, +0.49 on BCI, and +0.17 to +0.36 on LLD-MMRI.

In absolute numbers, FM initialisation achieves the best classification performance under missing modalities. This missing-modality robustness has two independent sources: initialisation and fusion. Under the same fusion, the foundation-model late-fusion variants are more robust than the randomly initialised late-fusion baseline (ViT-random, late fusion), which isolates the contribution of initialisation. Under the same initialisation, even the randomly initialised late-fusion model degrades far less than channel concatenation, which isolates the contribution of fusion.

Neither factor is a purpose-built missing-modality mechanism. MM-DINOv2 is the only configuration trained with an explicit modality-masking objective (Methods, §4.4.2). Yet the strongest foundation-model late-fusion variants match its robustness without ever seeing a masked input during training (see Discussion, §3). Beyond missing modalities, deployed models must also generalise to unseen sites, which we examine next (§2.4).

2.4 FM initialisation improves cross-site generalisation only under semi-supervised training

On glioma, where site labels are available, we test how random and FM initialisations generalise to an unseen site (Figure 5). To this end, we compare downstream performance on sites available during training (EGD and UCSF-PDGM, both in-distribution) and an unseen one (TCGA, out-of-distribution). The FM-initialisation benefits are most pronounced in the semi-supervised configurations.

The best in-distribution configuration is ResNet-random on EGD [41] (MCC 0.75) and MedSigLIP late-fusion SSL on UCSF-PDGM [42] (0.80). Selection of the best model per initialisation is based on the validation set, comprising in-

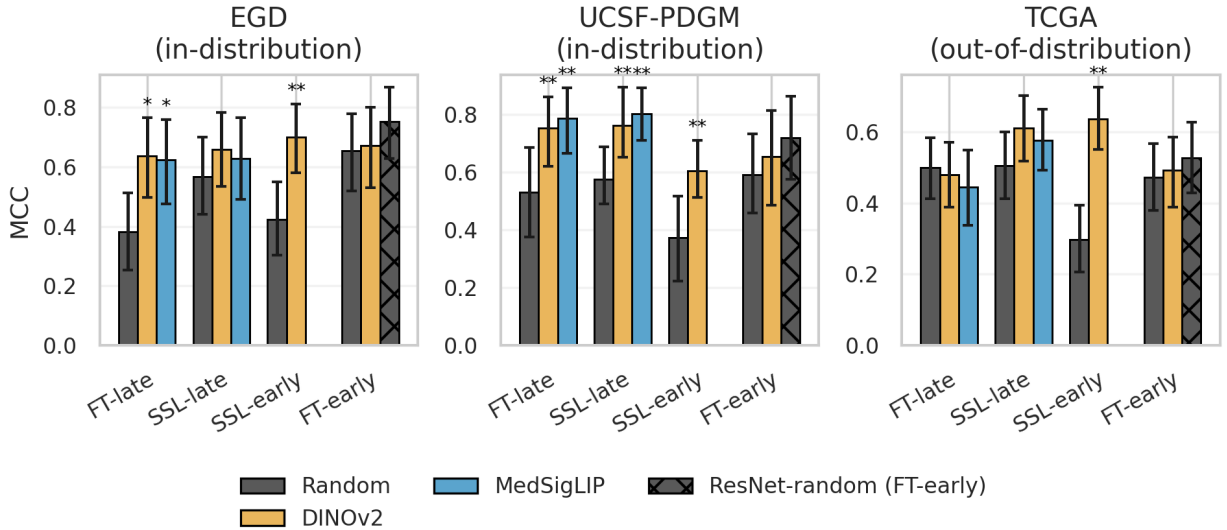


Figure 5: Per-site MCC on glioma. Each panel is one site: EGD [41] and UCSF-PDGM [42] are in-distribution (seen at training time), and TCGA is held out as an out-of-distribution test cohort. UPenn-GBM is excluded (single-class site, MCC undefined). Layout, bar grouping, and figure conventions within each panel match Figure 2. Error bars are 95 % bootstrap confidence intervals. Stars above a foundation-model bar mark Benjamini–Hochberg-corrected significance against its matched random-initialisation bar in the same group (* $p < 0.05$, ** $p < 0.01$).

distribution data and an in-house out-of-distribution cohort (validation MCC for every glioma model in Supplementary Table 1). On UCSF-PDGM the best FM configuration, MedSigLIP late-fusion SSL (0.80), leads the ResNet-random baseline (0.72). On EGD the best FM configuration, DINOv2 early-fusion SSL (0.70), sits just below ResNet-random (0.75). On the in-distribution sites, FM initialisation significantly improves over its matched random-initialisation baseline for several configurations.

On the out-of-distribution TCGA [43] cohort, the benefit of FM initialisation concentrates in the semi-supervised configurations. Every semi-supervised FM variant improves on its matched random-initialisation counterpart, by +0.07 to +0.34 MCC. The gain is significant for the early-fusion variant MM-DINOv2 (+0.34 MCC, FDR-corrected $p < 0.001$). The fine-tuned FM configurations show no such advantage, scoring between -0.05 and $+0.02$ MCC relative to their random-initialisation counterparts. Comparing the single best FM configuration against the strongest random-initialisation model, the two are tied on TCGA (FDR-corrected $p = 0.49$).

3 Discussion

We investigated whether foundation-model initialisation should be the default choice for multi-modal medical-imaging classifiers, which in practice must be built from few labels and incomplete imaging. We compared DINOv2 [14] and MedSigLIP [21] against random-initialisation baselines on glioma subtyping (multi-sequence MRI), HER2 scoring (paired H&E and IHC histopathology), and liver-lesion classification (multi-phase MRI). Across the four experiments (full-data accuracy, label efficiency, missing-modality robustness, cross-site generalisation) all matched comparisons result in on-par or improved performance under foundation-model initialisation. At full labels, foundation-model initialisation gains +0.21 MCC on liver lesions over the best random-initialisation baseline ($p < 0.01$). Mean per-class sensitivity rises from 0.54 to 0.76 on this task. Under label scarcity it never falls significantly below random initialisation at any labelled fraction on any dataset, with the largest best-versus-best gap of +0.37 MCC on liver lesions at 50 % labels. Of the four experiments, the widest gap is under missing-modality dropout, where randomly initialised models collapse while the foundation-model variants maintain their performance. On the out-of-distribution TCGA cohort for glioma, the semi-supervised foundation-model variants improve over their matched random-initialisation counterparts. The gain is significant for the early-fusion semi-supervised variant MM-DINOv2 (+0.34 MCC, $p < 0.001$).

We attribute these gains to foundation-model initialisation acting as a prior, which helps most where labelled data is scarce. When labelled data is plentiful, training a model from random initialisation becomes more feasible, so foundation-model and random initialisation tie on glioma and breast cancer, while FM initialisation leads only on liver lesions. However, the FM backbones do require fine-tuning, since a linear probe on frozen features stays at or below random initialisation on every primary metric. Among training strategies, the gain is largest under semi-supervised training, where foundation-model initialisation and learning from unlabelled data reinforce one another.

The gap is widest under missing modalities, where it has two independent sources. This robustness is a property of the pretrained features themselves: under the same settings, the foundation-model variants degrade far less under modality dropout than their random-initialisation counterparts. Yet, the fusion architecture adds a second, independent source, since even from random initialisation a late-fusion model degrades far less than a channel- or modality-token-concatenation model, which can collapse to near-chance performance. Neither source is a mechanism built for missing inputs: the strongest late-fusion foundation-model variants never saw a masked input during training, yet they match MM-DINOv2 and its explicit modality-masking objective. No single fusion strategy dominates: late fusion gives the strongest foundation-model configuration at full data on both glioma and liver lesions (MedSigLIP), whereas early fusion (MM-DINOv2) stands out on glioma under cross-site shift.

The two settings where foundation-model initialisation helps most, scarce labels and missing modalities, are also the two that matter most in the clinic. Expert annotations are expensive and slow to collect [10], and complete multi-modal imaging acquisitions are the exception once patient motion, time pressure, and differing protocols across hospitals are accounted for. Missing modalities in particular have motivated a dedicated line of work, with architectures and training objectives built specifically to stay accurate when an input is absent [27, 28, 29, 30, 11, 31]. A pretrained backbone minimises both failure modes at once, and provides the robustness clinical deployment needs from the choice of initialisation rather than from a special-purpose solution engineered for each failure mode. In practice, a new site can then deploy the model with the imaging and labels it already has, rather than assembling the large paired datasets random initialisation would need.

Reza et al. [31] propose a parameter-efficient adaptation procedure that improves missing-modality robustness on top of pretrained multi-modal networks. Our isolation result shows that generic foundation-model initialisation already supports much of this robustness through initialisation alone, without a dedicated adaptation procedure. Prior work shows that 2D foundation models, once properly adapted, transfer to medical imaging [33] and can match medical-specific models on 3D classification benchmarks [34]. However, the field has lacked a systematic analysis

of whether the initialisation advantage holds across multi-modal medical imaging. These settings combine distinct imaging modalities, may lack a modality at inference, and span acquisition shifts across sites. We provide that test across initialisations, fusion strategies, and training strategies, isolating the effect of initialisation rather than relying on a single adaptation recipe.

Four constraints bound the scope of these results. Experiments are restricted to 2D slices because the evaluated backbones are pretrained on 2D natural images, yet an extension to 3D has been studied in the literature [44, 45, 34]. The SDR-Former benchmark [5] reports a 3D ResNet-50 AUROC of 0.93 on liver lesions versus our 2D ResNet-50 AUROC of 0.86 (Supplementary Figure 1a), which bounds the cost of the 2D simplification. Missing modalities are handled by a simple minimum-value imputation at inference, so the robustness gap we measure is a lower bound on what foundation-model initialisation buys. We only evaluate on classification tasks, so transfer to dense prediction or time-to-event modelling remains open. Our benchmarks do not include other clinically common imaging pairs such as OCT with fundus photography or radiology with histopathology [32]. Because foundation-model initialisation acts as a modality-agnostic prior, we expect its benefit to extend to these pairs, though the best fusion strategy would still be task-specific.

Two questions remain open for future work. Matching fusion to a task’s modality structure is one, since no single strategy dominated across our benchmarks. Whether to use a medical vision FM (MedSigLIP) or a general-purpose vision FM (DINOv2) for initialisation is the other. MedSigLIP outperforms DINOv2 only on liver lesions (MCC 0.69 versus 0.66), and the two tie on glioma and breast cancer. MedSigLIP is both pretrained on medical images and built on the larger SigLIP SO400M backbone, so disentangling pretraining domain from capacity remains future work.

In summary, we advocate for foundation-model initialisation as the default choice for new multi-modal medical-imaging classifiers. It matched or exceeded training from random initialisation in nearly every setting we tested, across multi-sequence MRI, multi-stain histopathology, and multi-phase MRI. This strategy gained the most where random-initialisation pipelines are weakest: under label scarcity and missing modalities. These are also where the two approaches diverge most, and the scenarios most relevant for clinical translation. Based on our results, training a new multi-modal medical-imaging model from random initialisation should become the exception rather than the default.

4 Methods

4.1 Datasets

4.1.1 Breast Cancer IHC (BCI)

BCI contains paired haematoxylin and eosin (H&E) and HER2 immunohistochemistry (IHC) patches from registered whole-slide images (WSIs) of breast cancer tissue [4]. We follow the predefined train/test split released with the BCI Grand Challenge, which is constructed at the slide/patient level so that patches from a given WSI appear in only one split. We derive the train/validation split from the training portion using the same patient/slide-level disjointness. Images are resized to 224×224 . Counts, splits, labels, and class distribution appear in Table 1.

4.1.2 Liver-Lesion Classification on Multi-phase MRI (LLD-MMRI)

LLD-MMRI comprises multi-phase MRI with eight phases per patient for liver-lesion classification [5]. We use patient-level train/validation/test splits and crop lesions from bounding boxes, with intensities clipped to the 1st–99th percentile and rescaled to $[0, 1]$. During training we sample a 2D slice independently for each phase, so the slices do not need to align spatially across phases. Evaluation uses 98×98 centre slices without inter-phase registration. Cohort sizes, classes, and distributions are provided in Table 1.

4.1.3 Glioma Brain MRI

We unify seven public cohorts (BraTS2021 [46, 47], LUMIERE [48], UPenn-GBM [49], REMBRANDT [50], UCSF-PDGM [42], EGD [41], and TCGA [43]) for glioma subtype classification aligned with the WHO 2021 scheme [3]. The label set has three subtypes: astrocytoma, glioblastoma, and oligodendroglioma. Modalities are T1w, T2w, T1ce, and FLAIR. Preprocessing uses skull stripping (HD-BET [51]), intensity clipping to the 1st–99th percentile with rescaling to $[0, 1]$, and rigid registration to the SRI atlas [52] at 1 mm isotropic. For training we sample 2D slices that contain at least 500 tumour voxels.

The validation and test splits each combine an in-distribution portion, held out from the multi-site training pool, with a held-out external cohort. We hold the in-house cohort out of training entirely and use it as the out-of-distribution validation set. TCGA, also kept out of training, is the out-of-distribution test set. A subset of the validation cohort (87

Table 1: Dataset characteristics. Splits are patient-disjoint. Prevalence is reported over the full labelled cohort, which for glioma spans all labelled subjects across the train, validation, and test splits (1,898 in total), not only the 1,020 labelled training subjects. Patient-level age demographics were not reported by the source datasets.

BCI. Breast cancer HER2 scoring [4].

H&E and IHC stained histopathology patches (1024×1024), 4,873 registered pairs from 51 WSI pairs drawn from a cohort of 319 breast cancer patients at Capital Medical University (Beijing). Predefined train/test split with validation held out from train.

Split (image pairs)		HER2 score (prevalence)	
Train	3,116	0	2,143 (44 %)
Val	780	1+	1,336 (27 %)
Test	977	2+	1,154 (24 %)
		3+	240 (5 %)

LLD-MMRI. Liver-lesion classification on multi-phase MRI [5].

498 patients with one annotated lesion each, from Ningbo Medical Center Lihuili Hospital (Zhejiang, China). Each lesion is acquired across 8 MR phases (T2WI, in-phase, out-of-phase, C-pre, C+A, C+V, C+Delay, DWI). Predefined patient-level train/val/test split.

Split (patients)		Lesion class (prevalence)	
Train	316	Hepatocellular carcinoma	157 (31 %)
Val	78	Hepatic hemangioma	79 (16 %)
Test	104	Intrahepatic cholangiocarcinoma	58 (12 %)
		Hepatic cyst	53 (11 %)
		Hepatic abscess	54 (11 %)
		Hepatic metastasis	51 (10 %)
		Focal nodular hyperplasia	46 (9 %)

Glioma MRI. WHO 2021 glioma subtype classification.

Multi-sequence MRI (T1w, T2w, T1ce, FLAIR) unifying 7 public cohorts, with scans collected between 2004 and 2021. Six cohorts form the in-distribution pool used for training, validation, and test, while the seventh and an additional in-house cohort are held out from training. Semi-supervised training pool of 2,918 subjects, of which 1,020 carry WHO 2021 labels used for the supervised loss. The remaining unlabelled subjects contribute only to the self-supervised objective. The validation and test splits each combine an in-distribution portion held out from the multi-site training pool with a held-out external cohort: the in-house cohort is the out-of-distribution validation source and TCGA the out-of-distribution test source. In-distribution labelled subjects are split 70/10/20 stratified by class.

Split (patients)		Subtype (prevalence, labelled)	
Train (SSL pool)	2,918	Glioblastoma	1,524 (80 %)
of which labelled	1,020	Astrocytoma	240 (13 %)
Val (ID + OOD)	146 + 226	Oligodendroglioma	134 (7 %)
Test (ID + OOD)	292 + 214		

of 372 subjects) lacks the T2w sequence. We retain these subjects and treat the absent sequence as a missing modality at evaluation (§4.8). The training and test splits are complete across all four sequences. Evaluation uses the axial slice at the tumour centre of mass cropped to 98×98 . Slicing the test predictions by acquisition site supports the cross-site generalisation analysis (§2.4). We report per-site MCC and aggregated in-distribution versus out-of-distribution metrics alongside the overall numbers. Full cohort sizes, splits, and class distributions are summarised in Table 1.

4.2 Ethics oversight

This study used the BCI and LLD-MMRI datasets without additional ethics approval, as both are publicly released retrospective collections. The in-house glioma cohort was a retrospective analysis of an ongoing cohort of brain tumour patients, approved by the Ethics Committee of the Technical University of Munich (approval number 340/16S). The committee waived the requirement for informed consent because of the retrospective design. All methods were carried out in accordance with relevant guidelines and regulations.

4.3 Comparison framework

The headline question is whether foundation-model (FM) initialisation helps over training from random initialisation. We test random initialisation, general-purpose vision FM DINOv2 [14] pretraining, and medical vision FM MedSigLIP [21] pretraining. We compare over two further axes: fusion (early versus late) and training strategy (supervised fine-tuning, semi-supervised training, or linear probing on frozen features).

Within each (initialisation, dataset) pair the representative method is selected by best validation MCC, with macro-F1 as a tie-break. The best-versus-best comparisons reported in the Results (best FM-initialised configuration against the strongest random-initialisation model) use these validation-selected representatives. The test set is never consulted for model selection, and we report the corresponding test metric of the selected configuration. Selecting on validation rather than test means the reported strongest random-initialisation model is occasionally not the random-initialisation model with the highest test score at that setting. We keep the validation-selected choice, because picking the baseline by its test score would tune it on the same data the comparison reports.

Linear probes on frozen FM features are reference numbers for representation quality, not a fair head-to-head against random initialisation, and are flagged separately throughout. Throughout the Results we identify each model by its (initialisation, fusion, training strategy). The method names defined in this section are the matching labels used in Table 2 and the figures. Table 2 lists every method with its backbone, fusion architecture, and schedule.

For the label-efficiency experiment we subsample the training labels to 10 %, 25 %, 50 %, and 100 % with per-class stratification and retrain every configuration at each fraction, with the validation and test sets held fixed. The supervised and fine-tuning configurations train only on the retained labelled samples. The semi-supervised configurations keep every image and treat the dropped ones as unlabelled data, so a semi-supervised run at 10 % labels still sees the full set of images.

4.4 Foundation-model-initialised methods

4.4.1 Initialisations

We compare two pretrained backbones. DINOv2 [14] is a ViT-B/14 self-supervised on natural images. MedSigLIP [21] is a SigLIP SO400M ViT pretrained on biomedical image-text pairs. We use the public google/medsiglip-448 checkpoint, trained on a mixture of dermatology, radiology, ophthalmology, and pathology images paired with text. MedSigLIP is a strong medical vision foundation model and uses a larger backbone than DINOv2, so this initialisation reflects strong medical pretraining rather than an architecture-matched control, whereas DINOv2 against ViT-random isolates the effect of initialisation at a fixed architecture. Backbone parameter counts and feature dimensions are listed in Table 2.

4.4.2 Fusion variants

Early fusion (MM-DINOv2). MM-DINOv2 [33] extends DINOv2 with three modifications. First, **multi-modal tokenisation**: each modality is patchified independently, pre-trained DINOv2 positional embeddings are applied per modality, and a learnable modality embedding is added to each patch token. Second, **modality masking**: in addition to the iBOT patch-level masking objective [53], whole-modality masking is applied to half of the training samples, each dropping at most one modality, encouraging the model to impute missing-modality information. Third, **semi-supervised joint training**: the supervised and self-supervised objectives are optimised together in a single stage, without a separate pretraining phase (§4.4.3). Two minor adaptations are applied as defaults rather than from dedicated ablations: modality embeddings are zero-initialised instead of randomly, and the frozen-backbone warmup phase is omitted, with no degradation observed from either in our runs. Unless stated otherwise, backbone and optimisation settings follow DINOv2 defaults. Deviations are reported in §4.6.

Late fusion. The late-fusion variants apply a single encoder with shared weights independently to each modality, yielding per-modality class (CLS) and patch tokens. Per-modality features are concatenated in a fixed modality order and passed through two linear layers with a LeakyReLU activation, applied with separate weights to the CLS and patch token streams. Missing modalities are replaced with the minimum input value at inference. This architecture is shared across both initialisations, DINOv2 and MedSigLIP, under both semi-supervised training and fine-tuning.

4.4.3 Training strategies

Semi-supervised joint training. Cross-entropy on labelled examples and the DINO/iBOT self-supervised loss on all examples are optimised in a single stage [54], with no separate pretraining phase. The labelled cross-entropy term

is evaluated on the global crops with temperature 0.1 and label smoothing 0.1. It is added to the DINO and iBOT losses (each weighted 1.0) with a weight of 2.0, scaled by the fraction of labelled examples in the batch. Under this strategy we train MM-DINOv2 (DINOv2, early fusion) and the late-fusion models initialised with DINOv2 and with MedSigLIP.

Supervised fine-tuning. Only labelled examples are used, with cross-entropy loss. Backbone weights are unfrozen and updated with a lower learning rate than the classification head (Table 2). Each fine-tuned configuration uses the same fusion architecture as its semi-supervised counterpart: early fusion for DINOv2 (the fine-tuned MM-DINOv2) and late fusion for DINOv2 and for MedSigLIP.

Linear probing on frozen features. The linear-probe configurations, with DINOv2 and with MedSigLIP, freeze the FM encoder and extract per-modality CLS features. The per-modality feature vectors are concatenated in a fixed modality order, with no learnable projection between the frozen backbone and the classifier, and classified with L2-regularised logistic regression (cuML). The regularisation strength C is selected from 45 log-spaced candidates in $[10^{-6}, 10^5]$ on a 10 % held-out portion of the training set. The model is then retrained on all training data ($\text{max_iter} = 10,000$, $\text{tol} = 10^{-12}$), matching the protocol from [15].

4.5 Random-initialisation baselines

The random-initialisation baselines establish the reference that foundation-model initialisation is measured against. ResNet-random and ViT-random train a ResNet-50 [36] and a ViT-B/14 [55] from random initialisation. ViT-random shares the ViT-B/14 architecture of the DINOv2-initialised methods, so the random-initialisation comparison is architecture-matched rather than confounding initialisation with a ResNet-versus-ViT difference. ResNet-random is retained as a conventional convolutional reference. Modalities are stacked as channels and the first layer (Conv1 for ResNet-50 or patch embedding for ViT-B/14) is adapted to the resulting channel count.

We also train a randomly initialised late-fusion baseline, a ViT-B/14 with the same per-modality late-fusion architecture used by the FM late-fusion variants (§4.4.2). It isolates the contribution of foundation-model initialisation from the late-fusion architecture itself and is reported alongside the channel-concatenation random-initialisation baselines wherever missing-modality robustness is compared. These baselines use only labelled examples, with cross-entropy loss and a single learning rate shared between backbone and classification head.

To match the foundation-model semi-supervised configurations within each fusion, we also train two random-initialisation semi-supervised counterparts with the same joint-training recipe (§4.4.3): a late-fusion model and an early-fusion model. The early-fusion counterpart uses modality-token concatenation, the MM-DINOv2 architecture with random weights, rather than the channel concatenation of the supervised early-fusion baseline. The early-fusion initialisation comparison is therefore architecture-matched under semi-supervised training, whereas under fine-tuning it pairs a channel-concatenation random baseline against a modality-token foundation model. These counterparts supply the random-initialisation bars in the semi-supervised columns of every figure and are listed in Table 2. Augmentations match the global view augmentations of the DINOv2-based methods (the same random flips and crop scales, except for the BCI global-crop scale noted in §4.6), with no additional intensity augmentations beyond the dataset normalisation described in §4.6.

4.6 Implementation details

DINOv2-based methods use ViT-B/14. MedSigLIP-based methods use the SigLIP SO400M encoder. ResNet-random uses a ResNet-50 backbone. Each run uses a single NVIDIA A100 (80 GB) or A40 (46 GB) GPU. MedSigLIP-based methods require an A100 due to memory requirements. Because the MedSigLIP checkpoint is trained at 448×448 , we keep each task’s native input resolution and interpolate its positional embeddings to the corresponding token grid at every forward pass, rather than upsampling inputs to 448×448 .

For glioma, inputs are 2D axial slices cropped to 98×98 at the tumour centre of mass, with global crops at scale $[0.5, 1.0]$ and 56×56 local crops at scale $[0.1, 0.5]$. For BCI, H&E and IHC patches are resized to 224×224 with 98×98 local crops. The BCI global crop scale is $[0.5, 1.0]$ for supervised and MedSigLIP methods and the DINOv2 default $[0.32, 1.0]$ for DINOv2-based semi-supervised methods. For LLD-MMRI, lesion centre slices are cropped to 98×98 from bounding boxes. Intensity augmentation, namely DINOv2 colour jitter and greyscale, is applied on BCI only, since the dataset consists of RGB images. Glioma and LLD-MMRI intensities are clipped to the 1st–99th percentile and rescaled to $[0, 1]$, with no intensity augmentation.

The semi-supervised methods follow the DINOv2 multi-crop scheme with two global and eight local crops per image, and the iBOT objective masks 10–50 % of patches at random. For the semi-supervised joint training the DINO and

iBOT projection heads output as many prototypes as there are classes (3, 4, and 7 for glioma, BCI, and LLD-MMRI), rather than the DINOv2 default of 65,536. The joint objective shares these prototypes with the supervised classifier, so it requires one prototype per class. Optimisation follows DINOv2 (AdamW [56]) for the semi-supervised methods, whereas the random-initialisation baselines and supervised fine-tunes use Adam [57] without weight decay and a cosine learning-rate schedule without warmup. These settings are inherited from each strategy’s native recipe rather than harmonised across strategies. The same applies to the labelled cross-entropy, which keeps the DINOv2 temperature and label smoothing in the semi-supervised runs and is plain cross-entropy under supervised fine-tuning and the random-initialisation baselines.

All training schedules, batch sizes, and epoch counts are reported in Table 2.

4.7 Fairness and controls

All methods use a shared fixed random seed and identical train/validation/test splits. Each configuration is trained once, so the reported confidence intervals and significance tests reflect variability from test-set sampling. Every classification cross-entropy loss is weighted by negative log class frequency, with class k given weight $-\log(\pi_k + 10^{-6})$ where π_k is the empirical frequency of class k in the labelled training data. In the semi-supervised configurations this weighting applies to the labelled cross-entropy term. The weights are recomputed from the retained labels in each label-efficiency run, but because the subsampling is per-class stratified they stay essentially unchanged across labelled fractions. Input resolutions and global view augmentations are matched across methods, except that DINOv2-based semi-supervised runs retain the DINOv2 default global-crop scale on BCI (§4.6). Parameter counts are reported in Table 2. The missing-modality evaluation protocol (§4.8) is applied uniformly across methods.

4.8 Metrics and statistical analysis

Metrics. We choose the primary metric to stay reliable under class imbalance, following the Metrics Reloaded recommendations [58]. The primary metric is the Matthews correlation coefficient (MCC), the multiclass extension by Gorodkin [39] of the original coefficient [38], computed on the held-out test set at the checkpoint with best validation MCC. For the multiclass setting, MCC is a correlation coefficient with theoretical range $[-1, +1]$. A value of 0 corresponds to chance-level prediction and +1 to perfect classification, while the minimal attainable value depends on the empirical class distribution. Unlike accuracy or F1, MCC accounts for all cells of the confusion matrix and is not inflated by class imbalance [59, 60]. We additionally report macro-averaged area under the receiver operating characteristic curve (AUROC) as a secondary metric for context. For a clinically interpretable view we also report sensitivity, the macro-averaged per-class recall, equivalently balanced accuracy: the fraction of cases in each class that the model correctly identifies. MCC remains the primary metric on which model selection and the headline comparisons rest, and sensitivity is reported alongside it with 95 % bootstrap confidence intervals. Macro-F1, the unweighted mean of per-class F1 scores, is used only as a tie-break in model selection and for the per-class comparisons.

Statistical analysis. Uncertainty is quantified through 95 % confidence intervals via the non-parametric stratified percentile bootstrap ($n = 2,000$ resamples) on test-set predictions [61]. All significance tests are one-sided paired permutation tests on the held-out test predictions with H_1 : the foundation-model method outperforms the random-initialisation baseline, one-sided because the hypothesis is directional by construction, and Benjamini–Hochberg false discovery rate (FDR) correction [62] is applied at $\alpha=0.05$. Each test swaps the two methods’ paired predictions sample by sample under the null and estimates the p -value from $n = 10,000$ Monte Carlo permutations [63, 64]. MCC, macro-F1, and sensitivity differences are computed from the argmax predictions, and the AUROC difference from the predicted probabilities as macro one-vs-rest (one-vs-rest because all three tasks are multi-class).

We report two comparison families. The first compares each foundation-model configuration against the randomly initialised model in the same training-strategy and fusion cell (the matched pairs whose stars appear in every figure). The second is the best-versus-best comparison of §4.3, one comparison per dataset rather than one per cell. At full labels we test both comparison families on every metric, and we test per-class F1 against the best random-initialisation baseline for every class. For label efficiency we test at each labelled fraction. For missing-modality robustness we remove each modality in turn, replace it with the minimum input value at test time, and retain per-sample predictions, so each single-modality-removed case is tested with a paired permutation test rather than a descriptive confidence-interval-overlap heuristic. For cross-site generalisation we slice the full-data glioma test predictions by acquisition site and run a paired permutation test within each site, keeping model selection global and varying only the test evaluation.

FDR families are scoped per condition rather than pooled across datasets: one family per (dataset, metric, comparison family) for the full-data tests, refined further per mask, per labelled fraction, or per site. The per-class F1 family is

Table 2: Training hyperparameters, with columns giving each configuration’s training strategy, initialisation, and fusion. Backbone parameters exclude classification heads. Semi-supervised (SSL) configurations take `base_lr` from the dataset YAML and apply linear warmup for 10% of total epochs followed by cosine decay. Supervised configurations (random-initialisation and FM fine-tunes) override `base_lr` via launcher CLI flags and use cosine decay without warmup. FM fine-tunes report two learning rates as *backbone-LR / head-LR*. All runs use a single NVIDIA A100 (80 GB) or A40 (46 GB) GPU. MedSigLIP-based configurations require an A100.

	Init	Fusion	Backbone	Params	<i>d</i>
	Random (ResNet)	Early (channel-concat)	ResNet-50 (MONAI)	23.5 M	2,048
	Random	Early (channel-concat)	ViT-B/14 (MONAI)	86.6 M	768
	Random	Early (modality emb.)	ViT-B/14 (MM-DINOv2)	86.6 M	768
	Random	Late (cat + MLP)	ViT-B/14 (random init)	86.6 M	768
	DINOv2	Early (modality emb.)	DINOv2 ViT-B/14	86.6 M	768
	DINOv2	Late (cat + MLP)	DINOv2 ViT-B/14	86.6 M	768
	MedSigLIP	Late (cat + MLP)	MedSigLIP SO400M	~428 M	1,152

Dataset	Strategy	Init	Fusion	Epochs	Warmup	Batch/GPU	LR
Glioma	FT	Random (ResNet)	Early	500	0	96	10^{-4}
		Random	Early	500	0	96	10^{-4}
		Random	Late	500	0	96	10^{-4}
		DINOv2	Early	500	0	96	$10^{-5} / 10^{-4}$
		DINOv2	Late	500	0	96	$10^{-5} / 10^{-4}$
		MedSigLIP	Late	500	0	16	$10^{-5} / 10^{-4}$
	SSL	Random	Early	500	50	64	10^{-4}
		Random	Late	500	50	64	4×10^{-4}
		DINOv2	Early	500	50	64	10^{-4}
		DINOv2	Late	500	50	64	4×10^{-4}
		MedSigLIP	Late	500	50	16	10^{-4}
BCI	FT	Random (ResNet)	Early	500	0	96	10^{-4}
		Random	Early	500	0	96	10^{-4}
		Random	Late	500	0	96	10^{-4}
		DINOv2	Early	500	0	96	$10^{-5} / 10^{-4}$
		DINOv2	Late	500	0	96	$10^{-5} / 10^{-4}$
		MedSigLIP	Late	500	0	32	$10^{-5} / 10^{-4}$
	SSL	Random	Early	500	50	80	10^{-4}
		Random	Late	500	50	80	3.6×10^{-4}
		DINOv2	Early	500	50	80	10^{-4}
		DINOv2	Late	500	50	80	3.6×10^{-4}
		MedSigLIP	Late	500	50	16	8×10^{-4}
LLD-MMRI	FT	Random (ResNet)	Early	2,500	0	48	10^{-4}
		Random	Early	2,500	0	48	10^{-4}
		Random	Late	2,500	0	48	10^{-4}
		DINOv2	Early	2,500	0	48	$10^{-5} / 10^{-4}$
		DINOv2	Late	2,500	0	48	$10^{-5} / 10^{-4}$
		MedSigLIP	Late	2,500	0	32	$10^{-5} / 10^{-4}$
	SSL	Random	Early	2,500	250	48	10^{-4}
		Random	Late	2,500	250	48	10^{-4}
		DINOv2	Early	2,500	250	48	10^{-4}
		DINOv2	Late	2,500	250	48	10^{-4}
		MedSigLIP	Late	2,500	250	16	2.4×10^{-3}

scoped *per dataset*, because class sets differ across datasets (3 / 4 / 7 classes on glioma / BCI / LLD-MMRI) and pooling would let LLD-MMRI’s seven-class bloc distort the glioma and BCI families.

Data availability

The BCI dataset [4] is publicly available at <https://bupt-ai-cz.github.io/BCI/>. The LLD-MMRI dataset [5] is publicly available at <https://github.com/LMMEng/LLD-MMRI-Dataset>. The glioma cohort is a retrospective collection assembled from seven public sources (BraTS2021, LUMIERE, UPenn-GBM, REMBRANDT, UCSF-PDGM, EGD, TCGA). All constituent datasets are individually available at their respective repositories. See Methods for accession details. All seven sources are publicly available from their original repositories under the data-use terms specified there. No additional access restrictions apply to the assembled cohort. The in-house cohort used for out-of-distribution validation contains patient data and is not publicly available. It may be obtained from the corresponding author on reasonable request, subject to institutional approval.

Code availability

Our implementation builds on MM-DINOv2 [33]. The full training and evaluation code, including all configurations and instructions for reproducing the reported experiments, is available at <https://github.com/daniel-scholz/mm-dinov2> and will be made publicly available upon acceptance.

References

- [1] Bischof, J. *et al.* Multimodal bioimaging across disciplines and scales: Challenges, opportunities and breaking down barriers. *npj Imaging* **2**, 5 (2024).
- [2] Baltrušaitis, T., Ahuja, C. & Morency, L.-P. Multimodal Machine Learning: A Survey and Taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **41**, 423–443 (2019). URL <https://ieeexplore.ieee.org/document/8269806/>.
- [3] Wen, P. Y. & Packer, R. J. The 2021 WHO Classification of Tumors of the Central Nervous System: clinical implications. *Neuro-Oncology* **23**, 1215 (2021). URL <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8328017/>. Publisher: Oxford University Press.
- [4] Liu, S. *et al.* BCI: Breast Cancer Immunohistochemical Image Generation through Pyramid Pix2pix. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 1814–1823 (IEEE, New Orleans, LA, USA, 2022).
- [5] Lou, M. *et al.* SDR-Former: A siamese dual-resolution transformer for liver lesion classification using 3D multi-phase imaging. *Neural Networks* 107228 (2025).
- [6] Yang, C. *et al.* Deep learning model for the detection of prostate cancer and classification of clinically significant disease using multiparametric MRI in comparison to PI-RADs score. *Urologic Oncology: Seminars and Original Investigations* **42**, 158.e17–158.e27 (2024).
- [7] Cong, C. *et al.* MRI -Based Breast Cancer Classification and Localization by Multiparametric Feature Extraction and Combination Using Deep Learning. *Journal of Magnetic Resonance Imaging* **59**, 148–161 (2024).
- [8] Cai, J. C. *et al.* Fully Automated Deep Learning Model to Detect Clinically Significant Prostate Cancer at MRI. *Radiology* **312**, e232635 (2024).
- [9] Varma, A. *et al.* VariViT: A vision transformer for variable image sizes. In *Medical Imaging with Deep Learning*, 1571–1583 (PMLR, 2024).
- [10] Willemink, M. J. *et al.* Preparing Medical Imaging Data for Machine Learning. *Radiology* **295**, 4–15 (2020).
- [11] Zhang, C. *et al.* M3Care: Learning with Missing Modalities in Multimodal Healthcare Data. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2418–2428 (ACM, Washington DC USA, 2022). URL <https://dl.acm.org/doi/10.1145/3534678.3539388>.
- [12] Wu, R., Wang, H., Chen, H.-T. & Carneiro, G. Deep Multimodal Learning with Missing Modality: A Survey (2024). URL <http://arxiv.org/abs/2409.07825>. ArXiv:2409.07825 [cs].
- [13] Koch, V. *et al.* UNICORN: A deep learning model for integrating multi-stain data in histopathology. *npj Digital Medicine* **9** (2026). URL <https://www.nature.com/articles/s41746-026-02829-6>.

- [14] Oquab, M. *et al.* DINOv2: Learning Robust Visual Features without Supervision. *Transactions on Machine Learning Research* (2023). URL <https://openreview.net/forum?id=a68SUt6zFt>.
- [15] Baharoon, M. *et al.* Evaluating General Purpose Vision Foundation Models for Medical Image Analysis: An Experimental Study of DINOv2 on Radiology Benchmarks (2024). URL <http://arxiv.org/abs/2312.02366>. ArXiv:2312.02366 [cs].
- [16] Pérez-García, F. *et al.* Exploring scalable medical image encoders beyond text supervision. *Nature Machine Intelligence* **7**, 119–130 (2025). URL <https://www.nature.com/articles/s42256-024-00965-w>. Publisher: Nature Publishing Group.
- [17] Azizi, S. *et al.* Robust and data-efficient generalization of self-supervised machine learning for diagnostic imaging. *Nature Biomedical Engineering* **7**, 756–779 (2023). URL <https://www.nature.com/articles/s41551-023-01049-7>. Publisher: Nature Publishing Group.
- [18] Bannur, S. *et al.* MAIRA-2: Grounded Radiology Report Generation (2024). URL <http://arxiv.org/abs/2406.04449>. ArXiv:2406.04449 [cs].
- [19] Ma, J. *et al.* Segment anything in medical images. *Nature Communications* **15**, 654 (2024).
- [20] Ma, J. *et al.* MedSAM2: Segment anything in 3D medical images and videos. *arXiv preprint arXiv:2504.03600* (2025).
- [21] Sellergren, A. *et al.* MedGemma technical report. *arXiv preprint arXiv:2507.05201* (2025).
- [22] Zhang, G. *et al.* A multi-modal foundation model for brain disease diagnosis and medical imaging. *Patterns* **7**, 101538 (2026).
- [23] Wu, C. *et al.* Towards generalist foundation model for radiology by leveraging web-scale 2D & 3D medical data. *Nature Communications* **16** (2025).
- [24] Sun, K. *et al.* Medical multimodal foundation models in clinical diagnosis and treatment: Applications, challenges, and future directions. *Artificial Intelligence in Medicine* **170**, 103265 (2025). URL <https://doi.org/10.1016/j.artmed.2025.103265>.
- [25] Luo, L. *et al.* A large model for non-invasive and personalized management of breast cancer from multiparametric MRI. *Nature Communications* **16** (2025).
- [26] Morano, J. *et al.* Multimodal foundation model and benchmark for comprehensive retinal OCT image analysis. *npj Digital Medicine* **8** (2025).
- [27] Havaei, M., Guizard, N., Chapados, N. & Bengio, Y. HeMIS: Hetero-Modal Image Segmentation. In Ourselin, S., Joskowicz, L., Sabuncu, M. R., Unal, G. & Wells, W. (eds.) *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016*, vol. 9901, 469–477 (Springer International Publishing, Cham, 2016). URL https://link.springer.com/10.1007/978-3-319-46723-8_54. Series Title: Lecture Notes in Computer Science.
- [28] Liu, S., Wang, H., Li, S. & Zhang, C. Mixture-of-experts and semantic-guided network for brain tumor segmentation with missing MRI modalities. *Medical & Biological Engineering & Computing* **62**, 3179–3191 (2024). URL <https://link.springer.com/10.1007/s11517-024-03130-y>.
- [29] Zhou, T., Canu, S., Vera, P. & Ruan, S. Latent Correlation Representation Learning for Brain Tumor Segmentation with Missing MRI Modalities. *IEEE Transactions on Image Processing* **30**, 4263–4274 (2021). URL <http://arxiv.org/abs/2104.06231>. ArXiv:2104.06231 [eess].
- [30] Liu, H. *et al.* M3AE: Multimodal Representation Learning for Brain Tumor Segmentation with Missing Modalities (2023). URL <http://arxiv.org/abs/2303.05302>. ArXiv:2303.05302 [eess].
- [31] Reza, M. K., Prater-Bennette, A. & Asif, M. S. Robust Multimodal Learning With Missing Modalities via Parameter-Efficient Adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **47**, 742–754 (2025). URL <https://ieeexplore.ieee.org/document/10713849/>.
- [32] Li, Y. *et al.* A review of deep learning-based information fusion techniques for multimodal medical image classification. *Computers in Biology and Medicine* **177**, 108635 (2024). URL <https://doi.org/10.1016/j.compbiomed.2024.108635>.
- [33] Scholz, D. *et al.* MM-DINOv2: Adapting Foundation Models for Multi-modal Medical Image Analysis. In Gee, J. C. *et al.* (eds.) *Medical Image Computing and Computer Assisted Intervention – MICCAI 2025*, 320–330 (Springer Nature Switzerland, Cham, 2025).
- [34] Liu, H. *et al.* Revisiting 2D foundation models for scalable 3D medical image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 30021–30031 (2026).

- [35] Kushol, R., Parnianpour, P., Wilman, A. H., Kalra, S. & Yang, Y.-H. Effects of MRI scanner manufacturers in classification tasks with deep learning models. *Scientific Reports* **13**, 16791 (2023).
- [36] He, K., Zhang, X., Ren, S. & Sun, J. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770–778 (2016).
- [37] Chen, C., Mat Isa, N. A. & Liu, X. A review of convolutional neural network based methods for medical image classification. *Computers in Biology and Medicine* **185**, 109507 (2025).
- [38] Matthews, B. W. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochimica et Biophysica Acta (BBA) - Protein Structure* **405**, 442–451 (1975).
- [39] Gorodkin, J. Comparing two K-category assignments by a K-category correlation coefficient. *Computational Biology and Chemistry* **28**, 367–374 (2004).
- [40] Tamura, J., Itaya, Y., Hayashi, K. & Yamamoto, K. Statistical inference of the Matthews correlation coefficient for multiclass classification. *Journal of Applied Statistics* 1–26 (2026).
- [41] van der Voort, S. R. *et al.* The Erasmus Glioma Database (EGD): Structural MRI scans, WHO 2016 subtypes, and segmentations of 774 patients with glioma. *Data in Brief* **37**, 107191 (2021). URL <https://www.sciencedirect.com/science/article/pii/S2352340921004753>.
- [42] Calabrese, E. *et al.* The University of California San Francisco Preoperative Diffuse Glioma MRI Dataset. *Radiology: Artificial Intelligence* **4**, e220058 (2022). URL <https://pubs.rsna.org/doi/10.1148/ryai.220058>. Publisher: Radiological Society of North America.
- [43] Bakas, S. *et al.* Advancing The Cancer Genome Atlas glioma MRI collections with expert segmentation labels and radiomic features. *Scientific Data* **4**, 170117 (2017). URL <https://www.nature.com/articles/sdata2017117>. Number: 1 Publisher: Nature Publishing Group.
- [44] Müller-Franzes, G. *et al.* Medical Slice Transformer: Improved Diagnosis and Explainability on 3D Medical Images with DINOv2 (2024). URL <http://arxiv.org/abs/2411.15802>. ArXiv:2411.15802 [eess].
- [45] Yu, Y., Li, G. & Yang, J. PlaneCycle: Training-Free 2D-to-3D Lifting of Foundation Models Without Adapters. arXiv:2603.04165 (2026).
- [46] Menze, B. H. *et al.* The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS). *IEEE transactions on medical imaging* **34**, 1993–2024 (2015).
- [47] Baid, U. *et al.* The RSNA-ASNR-MICCAI BraTS 2021 Benchmark on Brain Tumor Segmentation and Radiogenomic Classification (2021). URL <http://arxiv.org/abs/2107.02314>. ArXiv:2107.02314 [cs].
- [48] Suter, Y. *et al.* The LUMIERE dataset: Longitudinal Glioblastoma MRI with expert RANO evaluation. *Scientific Data* **9**, 768 (2022). URL <https://www.nature.com/articles/s41597-022-01881-7>. Publisher: Nature Publishing Group.
- [49] Bakas, S. *et al.* The University of Pennsylvania glioblastoma (UPenn-GBM) cohort: advanced MRI, clinical, genomics, & radiomics. *Scientific Data* **9**, 453 (2022). URL <https://www.nature.com/articles/s41597-022-01560-7>. Number: 1 Publisher: Nature Publishing Group.
- [50] Gusev, Y. *et al.* The REMBRANDT study, a large collection of genomic data from brain cancer patients. *Scientific Data* **5**, 180158 (2018). URL <https://www.nature.com/articles/sdata2018158>. Publisher: Nature Publishing Group.
- [51] Isensee, F. *et al.* Automated brain extraction of multisequence MRI using artificial neural networks. *Human Brain Mapping* **40**, 4952–4964 (2019).
- [52] Rohlfing, T., Zahr, N. M., Sullivan, E. V. & Pfefferbaum, A. The SRI24 multichannel atlas of normal adult human brain structure. *Human Brain Mapping* **31**, 798–819 (2009). URL <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2915788/>.
- [53] Zhou, J. *et al.* Image BERT Pre-training with Online Tokenizer. In *International Conference on Learning Representations (ICLR)* (2022). URL <https://openreview.net/forum?id=ydopy-e6Dg>.
- [54] Fini, E. *et al.* Semi-Supervised Learning Made Simple With Self-Supervised Clustering. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 3187–3197 (2023). URL https://openaccess.thecvf.com/content/CVPR2023/html/Fini_Semi-Supervised_Learning_Made_Simple_With_Self-Supervised_Clustering_CVPR_2023_paper.html.
- [55] Dosovitskiy, A. *et al.* An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations (ICLR)* (2021). URL <https://openreview.net/forum?id=YicbFdNTTy>.

- [56] Loshchilov, I. & Hutter, F. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations (ICLR)* (2019).
- [57] Kingma, D. P. & Ba, J. Adam: A Method for Stochastic Optimization. In Bengio, Y. & LeCun, Y. (eds.) *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings* (2015).
- [58] Maier-Hein, L. *et al.* Metrics reloaded: recommendations for image analysis validation. *Nature Methods* **21**, 195–212 (2024). URL <https://www.nature.com/articles/s41592-023-02151-z>. Publisher: Nature Publishing Group.
- [59] Chicco, D. & Jurman, G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics* **21**, 6 (2020). URL <https://doi.org/10.1186/s12864-019-6413-7>.
- [60] Chicco, D., Tötsch, N. & Jurman, G. The Matthews correlation coefficient (MCC) is more reliable than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix evaluation. *BioData Mining* **14**, 13 (2021). URL <https://doi.org/10.1186/s13040-021-00244-z>.
- [61] Efron, B. & Tibshirani, R. J. *An Introduction to the Bootstrap* (Chapman & Hall, New York, 1993).
- [62] Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)* **57**, 289–300 (1995).
- [63] Ernst, M. D. Permutation methods: A basis for exact inference. *Statistical Science* **19** (2004).
- [64] Phipson, B. & Smyth, G. K. Permutation P-values should never be zero: Calculating exact P-values when permutations are randomly drawn. *Statistical Applications in Genetics and Molecular Biology* **9** (2010).

Acknowledgements

This study was supported by the DFG, grant #504320104. Computational experiments were performed on our laboratory’s local GPU cluster equipped with NVIDIA A100 (80 GB) and A40 (46 GB) GPUs. All training runs used a single GPU.

Author contributions

D.S.: Conceptualization, Methodology, Software, Formal analysis, Investigation, Data curation, Visualization, Writing—original draft. A.C.E.: Conceptualization, Writing—review & editing. C.S.: Conceptualization, Writing—review & editing. E.U.: Data curation, Writing—review & editing. J.C.P.: Supervision, Writing—review & editing. D.R.: Supervision, Writing—review & editing. B.W.: Supervision, Writing—review & editing, Project administration, Funding acquisition. D.R. and B.W. contributed equally as senior authors.

Additional information

Competing interests The authors declare no competing interests.

Supplementary Files

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

- [supplementary.pdf](#)