AI Medicine
https://www.sciltp.com/journals/aim

Article

Revisiting Dual-Encoder Fusion and Optimal Transport Regularization in PET-CT Tumor Segmentation: A Matched-Control Study under Extended Training

Man Li, Yongwei Tang, Xiaoqing Zhao, Yangyang Zhao and Meng Wang *

School of Electronic and Information Engineering, Liaoning Technical University, Xingcheng 125105, China
* Correspondence: wangmengdx@lntu.edu.cn

How To Cite: Li, M.; Tang, Y.; Zhao, X.; et al. Revisiting Dual-Encoder Fusion and Optimal Transport Regularization in PET-CT Tumor Segmentation: A Matched-Control Study under Extended Training. *AI Medicine* **2026**, *3*(2), 12. https://doi.org/10.53941/aim.2026.100012

Received: 30 June 2026
Revised: 6 October 2026
Accepted: 15 October 2026
Published: 30 October 2026

Abstract: Dual-encoder fusion and explicit cross-modal alignment are two common strategies for multimodal PET-CT segmentation, but their relative contributions remain unclear when training duration and model capacity are controlled. We therefore reassess dual-encoder fusion and Optimal Transport (OT) regularization in DpDNet using an extended 200-epoch protocol with convergence monitoring and five-fold cross-validation on 75 AutoPET-derived PET-CT cases. The dual encoder shows no significant Dice advantage over a parameter-matched, approximately compute-matched weight-shared two-stream encoder (ΔDice = 0.0076, $p = 0.61$), but achieves higher Dice than a parameter-matched early-fusion encoder (ΔDice = 0.0465, $p = 0.016$). Because the latter requires $1.24\times$ the FLOPs, this difference cannot be attributed uniquely to fusion position. OT yields no significant late-training Dice improvement on either the dual encoder (ΔDice = 0.0091, $p = 0.58$) or the weight-shared two-stream encoder (ΔDice = 0.0080, $p = 0.45$), and a 20-epoch pilot advantage does not persist under the 200-epoch protocol. The previously observed pre-training–OT interaction is likewise not reproduced and remains exploratory. Voxel Dice does not distinguish the matched-capacity configurations, whereas lesion-level detection shows only nominal differences that do not survive multiple-comparison correction. These results indicate that conclusions about multimodal PET-CT segmentation depend strongly on control design, training duration, and evaluation metric.

Keywords: PET-CT segmentation; multimodal fusion; dual-encoder architecture; optimal transport; convergence analysis; lesion-level detection

1. Introduction

Positron Emission Tomography and Computed Tomography (PET-CT) are central to oncological imaging. PET offers metabolic sensitivity but limited spatial resolution; CT provides anatomical detail but limited functional specificity. Their complementarity makes PET-CT indispensable for radiotherapy target delineation, surgical planning, and treatment response assessment [1, 2]. Accurate tumor segmentation is therefore critical: under-segmentation risks recurrence, over-segmentation increases dose to organs at risk, and manual delineation remains slow and variable (Dice 0.60–0.85). Automated segmentation is thus an urgent clinical need.

The core difficulty lies in fusing two modalities with fundamentally different physical mechanisms. CT Hounsfield units produce sharp anatomical boundaries, while PET SUV values produce diffuse metabolic signals, differing in intensity distribution, spatial resolution, and noise [3, 4]. The literature has pursued two largely independent directions: modality-specific dual encoders [3, 5, 6] and explicit cross-modal alignment, including contrastive learning, mutual information maximization, and Optimal Transport (OT) [7–9]. Which contributes more to final accuracy, and whether the two interact, remains unresolved.

A methodological confound obscures this question. A reported training budget is not evidence of convergence:

Copyright: © 2026 by the authors. This is an open access article under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).
Publisher's Note: Scilight stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

none of the compared works states a convergence criterion, and the epoch is not comparable across studies because patch and batch definitions differ. A model evaluated at a fixed early budget may therefore reflect faster convergence rather than better late-training performance. We quantify this directly rather than attributing it to prior work. Whether OT follows this pattern is tested, not assumed: on our own curves, the original OT arm is behind at every logged checkpoint, including the earliest, so the convergence-acceleration account does not describe our measurements. If convergence acceleration is read as accuracy improvement, the evidence base for methodological decisions weakens. A short budget is inexpensive and its gain is real at that budget; we therefore report the outcome of applying the test in our setting, with competing readings stated side by side.

Four questions follow. Past the plateau, does the dual-encoder architecture still dominate regularization-based approaches, or do OT constraints add independent value? Does OT genuinely improve late-training quality, or merely accelerate convergence? Do MAE pre-training and OT complement each other, or does pre-training already achieve implicit alignment, making OT redundant or harmful? And if Dice barely moves, does any lesion-level metric separate configurations, or are they equivalent by every computable metric?

This paper addresses these questions through a 200-epoch controlled study on the AutoPET dataset [10] (75 cases, 5-fold cross-validation) using the DpDNet framework [11]. Our findings are threefold.

First, the dual encoder shows no statistically significant Dice advantage over a parameter-matched, approximately compute-matched weight-shared two-stream encoder (ΔDice = 0.0076, $p = 0.61$). Although it outperforms a parameter-matched early-fusion encoder (ΔDice = 0.0465, $p = 0.016$, uncorrected), the latter requires $1.24\times$ FLOPs, so the difference cannot be attributed uniquely to fusion position or computational budget.

Second, hierarchical OT regularization provides no statistically significant late-training Dice improvement under the 200-epoch protocol (ΔDice = 0.0091 and 0.0080 for the dual and weight-shared encoders, respectively), and we find no clear evidence that it accelerates convergence. The confidence intervals do not exclude potentially meaningful positive effects. An apparent 0.062 Dice advantage observed in a separate 20-epoch pilot disappears after extended training, highlighting the importance of evaluating training dynamics alongside final performance.

Third, voxel-level Dice does not distinguish the matched-capacity configurations, whereas lesion-level metrics reveal nominal differences that do not survive multiple-comparison correction. An additional single-fold analysis of the MAE–OT interaction remains exploratory and does not establish a systematic effect.

Together, these findings emphasize the importance of matched architectural controls, extended training, and multi-metric evaluation when assessing multimodal PET-CT segmentation methods.

2. Related Work

2.1. PET-CT Segmentation: From Single to Dual Encoders

Early PET-CT segmentation methods relied on handcrafted radiomic features and traditional machine learning [2,12]. The advent of deep learning, particularly U-Net [13] and its many reported variants [14–16], established encoder-decoder architectures as the de facto standard. For multimodal inputs, three fusion paradigms exist: input-level channel concatenation [13], feature-level fusion with independent modality encoders [3,5], and decision-level combination. Feature-level fusion via dual encoders has gained prominence for PET-CT [6], as it allows CT and PET to be processed by modality-specific feature extractors before bottleneck concatenation. Transformer-based architectures [17–19] have further expanded the architectural design space, though their comparative advantage over well-tuned CNN dual encoders at matched training budgets remains underexplored.

2.2. Cross-Modal Alignment: Convergence Accelerator or Accuracy Driver?

To mitigate the distribution gap between CT and PET features [12], researchers have introduced explicit alignment constraints: contrastive learning [7], mutual information maximization, adversarial domain adaptation [20,21], and Optimal Transport (OT)-based distribution matching [8,22]. OT is particularly appealing for multimodal alignment because the Wasserstein distance provides a geometrically meaningful metric between probability distributions with guaranteed gradients [23,24]. The closest recent work, C^{2}MAOT [9], applies hierarchical OT to cross-modal masked autoencoding for PET-CT and reports an improvement over its own baseline; we do not contest that result at the budget at which it was measured. Our concern is narrower: the training budget in such studies is fixed before the study and is usually reported without an accompanying convergence criterion, so it cannot be determined from the report whether the gain reflects a genuine accuracy improvement or merely faster optimization. Alignment losses can provide strong gradient signals during early training and may alter convergence dynamics [25]. Whether the reported gains reflect genuine accuracy improvement or merely faster optimization cannot be determined from a training budget whose relation to convergence is not stated. Deciding the question requires re-running the comparison to convergence under a stated criterion, which is what the present study does for the OT component.

2.3. Self-Supervised Pre-Training and Its Interaction with Regularization

Masked Autoencoding (MAE) [26] and related self-supervised paradigms [27–29] learn transferable visual representations from unlabeled data. In medical imaging, pre-training on domain-specific images has demonstrated consistent benefits for downstream segmentation [28]. However, the interaction between pre-training and downstream task-specific regularization is poorly understood. Conceptually, both pre-training and explicit alignment constraints aim to improve feature representations; whether they complement each other or produce redundant—or even conflicting—gradient signals has not been empirically tested. This question is practically significant either way: if pre-training already achieves implicit cross-modal alignment, downstream OT constraints add computation for no benefit, and if it does not, the two should compound. Our own measurements do not resolve which holds—the pre-training arms are single-fold and the effect is smaller than we can distinguish from noise—so we state the question rather than an answer.

2.4. The Convergence Gap in Comparative Evaluation

A recurring but underappreciated methodological issue in the medical image segmentation literature is the evaluation of methods at training budgets that differ across studies and are rarely tied to a convergence criterion [16,30]. Different architectures and regularization strategies have different convergence speeds. Epoch counts are not comparable across studies: some frameworks define one epoch as a fixed number of patch-level iterations and others as one pass over the training patches, and the batch and patch sizes differ, so a fixed budget of 30, 50 or 100 epochs denotes very different amounts of optimization, and any comparison that rests on the epoch count alone inherits that ambiguity. When two methods are compared at a budget at which neither has been shown to be converged, the faster-converging one can appear superior regardless of its late-training capability. The present study does not propose a new architecture or loss function; instead, it revisits one published comparison under a 200-epoch protocol with matched-capacity controls, and reports where the converged measurement agrees with the original and where it does not. We make no claim about how often the same pattern holds elsewhere.

3. Methods

3.1. Base Framework: DpDNet

This study employs DpDNet (Dual-prompt Driven Segmentation Network), proposed in [11], as the baseline framework for all experiments. DpDNet is designed for multi-cancer PET-CT segmentation with a mature prompt-driven decoder module that enables segmentation of multiple tumor types through task-specific prompts without training separate models. The original DpDNet uses a single encoder that receives CT and PET concatenated along the input channel dimension, that is, fusion at the input; we refer to this configuration throughout as the *early-fusion single encoder*, and it is the configuration the original paper reports. Whether this encoder is in fact inferior to modality-specific encoders is the empirical question this study addresses, so we do not assume it here.

3.2. Dual-Encoder Architecture Design

The core architectural modification of this paper is replacing DpDNet's single shared encoder with a dual independent-encoder architecture, as illustrated in Figure 1.

The design comprises four key components:

(1) **Modality-specific encoders:** Independent encoder pathways are allocated for CT and PET respectively. Both encoders share an identical hierarchical structure (6 encoding stages with channel dimensions of 16, 32, 64, 128, 256, and 256, respectively), but maintain separate learnable parameters.

(2) **Independent feature extraction:** The CT encoder exclusively processes CT input (1 channel), learning anatomical structural features; the PET encoder exclusively processes PET input (1 channel), learning metabolic activity features. This design enables each encoder to develop modality-specialized feature extractors without cross-modal interference.

(3) **Bottleneck fusion:** The outputs of the two encoders are concatenated along the channel dimension at the bottleneck layer (deepest features) and subsequently fed into the shared prompt-driven decoder.

(4) **Parameter analysis:** The dual encoder is not compared here against the original 22M single encoder, because that comparison confounds the fusion architecture with the parameter count. Its capacity is instead bounded by two matched-capacity controls, described next and listed with their parameter and FLOP counts in Table 1.

Single-encoder controls. Two different single-encoder models are used, because each isolates a different variable, and they are not interchangeable. (i) The *weight-shared two-stream encoder* pushes CT and PET through one shared set of encoder weights and concatenates the two resulting feature branches along the channel dimension

at the bottleneck, exactly as the dual encoder does, so the only difference from the dual encoder is whether the two modalities share weights. The two modalities are still processed in two separate passes through the encoder and share only their parameters, so this control isolates shared from independent parameterization rather than the number of processing streams; the input-level single-stream design is the second control below. Its first four stages keep the dual encoder's channel widths (16/32/64/128); only the two deepest stages are widened (256 to 298) so that the total parameter count matches (33.07 M versus 33.11 M, a difference of 0.11%). Because the first four stages are identical and the two widened deep stages together account for under 2% of the FLOPs, its FLOP count is 1.01 times the dual encoder's rather than equal to it: only those two stages differ, and apart from them both networks run the same single-pass computation once per modality. (ii) The *parameter-matched early-fusion single encoder* keeps the original DpDNet design and fuses the modalities at the input, so it consumes a channel-concatenated CT+PET tensor; it is widened to 33.10 M. It matches the parameter count but costs 1.24 times the FLOPs of the dual encoder, because a concatenated input carries both modalities through every layer at once. We report both controls rather than one, and we state the effective parameter counts directly in Table 1 so that the source of any capacity difference is unambiguous.

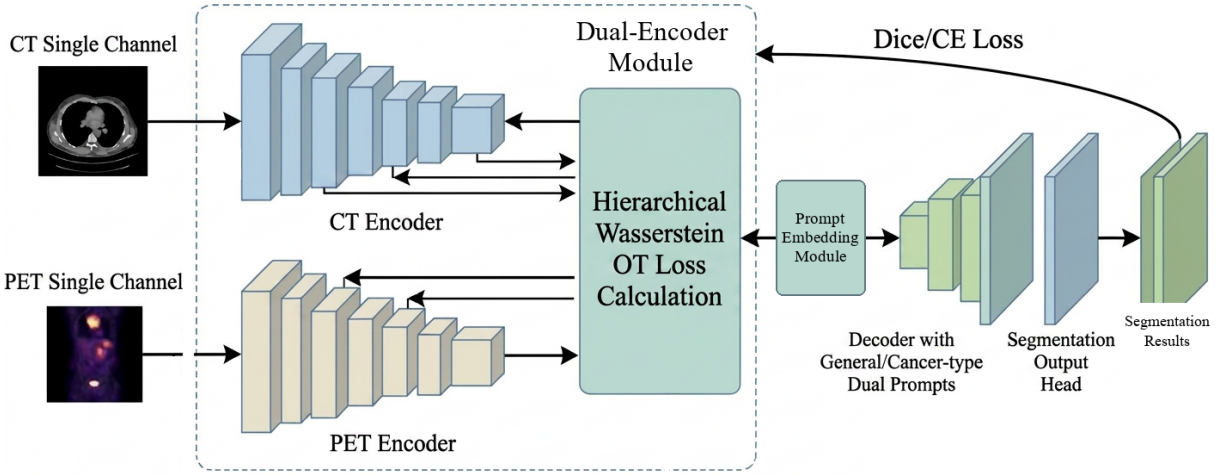

Figure 1. The dual-encoder PET-CT segmentation network architecture. CT and PET images are processed by independent modality-specific encoders (6 stages each, channel widths 16, 32, 64, 128, 256, 256), fused at the bottleneck via channel concatenation, and decoded by a shared prompt-driven decoder that carries the general prompt and the cancer-type prompt shown in the figure. Hierarchical Wasserstein OT losses ($\mathcal{L}_{OT}^{l}$) enforce cross-modal feature alignment at all six encoder depths. Only this configuration is drawn. The two matched-capacity controls it is compared against are two different models, neither of which is the original 22 M baseline encoder, and both are specified here. The *weight-shared two-stream encoder* sends CT and PET through one shared set of encoder weights and keeps stages 1–4 at the widths shown above (16, 32, 64, 128), widening only the two deepest stages (256 to 298) so that it matches the dual encoder, 33.07 M against 33.11 M. The *early-fusion* single encoder instead concatenates CT and PET at the input, so a single 2-channel stream traverses every layer; it is widened to 33.10 M, and it is the configuration the original DpDNet reports. Both controls are described in Section 3.2 and listed with their parameter and FLOP counts in Table 1.

Table 1. Five-fold cross-validation Dice under the extended 200-epoch protocol. Folds 0–4 report per-fold results; mean ± SD is calculated across folds. Relative FLOPs are normalized to the dual encoder. In the dual-encoder row marked *(original)* the transport term was detached from the encoder and contributes no gradient; *(corrected)* is the differentiable loss.

Configuration	Params (M)	FLOPs	Fold 0	Fold 1	Fold 2	Fold 3	Fold 4	Mean ± SD
Early-fusion single, no OT	33.10	1.24×	0.584	0.557	0.663	0.708	0.803	0.663 ± 0.099
Weight-shared two-stream, no OT	33.07	1.01×	0.638	0.650	0.684	0.705	0.831	0.702 ± 0.077
Weight-shared two-stream, + OT	33.07	1.01×	0.649	0.652	0.727	0.708	0.814	0.710 ± 0.068
Dual encoder, no OT	33.11	1.00×	0.641	0.624	0.733	0.735	0.815	0.709 ± 0.078
Dual encoder, + OT (original)	33.11	1.00×	0.641	0.617	0.758	0.718	0.813	0.709 ± 0.081
Dual encoder, + OT (corrected)	33.11	1.00×	0.622	0.689	0.727	0.724	0.830	0.718 ± 0.075

A design consequence of this layout is that the parameter-matched early-fusion control has no OT counterpart. The OT term is defined between the layer-wise features of two separate modality encoders; the early-fusion model

produces a single feature stream and therefore has no such pair of feature sequences to align. That cell of the design is undefined rather than omitted for convenience, which is why no such row appears in Table 1; the single-encoder OT control is the weight-shared two-stream row instead.

The core intuition underlying the dual-encoder design is that CT and PET possess fundamentally different physical meanings and statistical properties. CT Hounsfield units reflect tissue density with sharp anatomical boundaries, while PET SUV values represent metabolic activity with diffuse spatial patterns. Employing independent encoders allows each pathway to learn dedicated feature extraction operators optimized for their respective imaging physics.

3.3. Hierarchical Optimal Transport Feature Alignment

We introduce a hierarchical Optimal Transport (OT) regularization term to reduce the distribution shift between CT anatomical features and PET metabolic features. The OT loss is applied at multiple encoder hierarchy levels.

(1) **Hierarchical first-order Wasserstein distance**

Let $\mathbf{F}_l^{\mathrm{CT}}$ and $\mathbf{F}_l^{\mathrm{PET}}$ denote the feature outputs of the CT and PET encoders at layer l, respectively. The 1-Wasserstein distance between the two feature distributions is:

$$W_1(\mu_l^{\mathrm{CT}}, \mu_l^{\mathrm{PET}}) = \inf_{\gamma \in \Pi(\mu_l^{\mathrm{CT}}, \mu_l^{\mathrm{PET}})} \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim\gamma} \left[\|\mathbf{x} - \mathbf{y}\|_1\right] \tag{1}$$

where μ_l^{CT} and μ_l^{PET} represent the marginal distributions of the two modality features at layer l, and $\Pi(\cdot, \cdot)$ denotes the set of all joint distributions with these marginals.

(2) **Efficient sampling-and-sorting approximation**

Directly solving the optimal transport plan via the Sinkhorn algorithm [31] incurs $O(K^2)$ complexity for K spatial feature locations. We adopt an efficient sampling-and-sorting approximation [24]:

$$\widehat{W}_1(\mathbf{F}_l^{\mathrm{CT}}, \mathbf{F}_l^{\mathrm{PET}}) = \frac{1}{KC_l} \sum_{c=1}^{C_l} \sum_{k=1}^{K} \left| F_{l,(k,c)}^{\mathrm{CT}} - F_{l,(k,c)}^{\mathrm{PET}} \right| \tag{2}$$

where C_l is the number of feature channels at layer l and $F_{l,(k,c)}$ denotes the k-th order statistic of channel c after sorting over the sampled spatial locations. The two quantities being compared are therefore the per-channel empirical marginals of the two modality feature maps: the K spatial locations are drawn once and shared by both modalities, each channel is then sorted independently on either side, and the absolute differences are averaged over both the sample index k and the channel index c. Equation (2) is therefore not a sample estimator of the joint quantity in Equation (1), but a channel-wise marginal one-dimensional surrogate for it: Equation (1) requires a transport plan that couples the two feature maps across all channels at once, and the two agree only when the channels are statistically independent, which the feature maps of a convolutional encoder are not. We adopt the surrogate for its cost, $O(K \log K)$ per layer at $K = 4096$ shared locations.

We do not claim an accuracy guarantee for this estimator. Under the identical fold-0 protocol we trained the dual encoder with this estimator, with sliced Wasserstein over 32 random one-dimensional projections, and with log-domain entropic OT solved by Sinkhorn iterations, holding the architecture, the optimizer, the schedule and the budget fixed and varying only the transport term. All three transports land inside the fold-to-fold standard deviation (SD) of the 200-epoch protocol, so this comparison can screen for gross estimator failure but cannot rank them. It is reported as a single-fold screen in the supplementary material, and we make no claim about the relative accuracy of the transport estimators.

(3) **Hierarchically weighted OT loss**

Following the hierarchical representation principles of deep networks, we apply OT matching at multiple encoder stages and aggregate with depth-increasing weights:

$$\mathcal{L}_{\mathrm{OT}} = \sum_{l=1}^{L} w_l \cdot \widehat{W}_1(\mathbf{F}_l^{\mathrm{CT}}, \mathbf{F}_l^{\mathrm{PET}}), \quad w_l = \frac{2^l}{\sum_{j=1}^{L} 2^j} \tag{3}$$

where $L = 6$ is the total number of encoder stages. The weights w_l increase exponentially with depth, reflecting the greater semantic importance of deeper features.

(4) **Training dynamic weight scheduling**

To balance early-stage modality alignment with late-stage segmentation task convergence, an adaptive OT weight schedule $\lambda(t)$ is employed:

$$\lambda(t) = \begin{cases} \lambda_{\max} \dfrac{t}{W}, & 0 \le t \le W \\ \lambda_{\max} - (\lambda_{\max} - \lambda_{\min}) \dfrac{t - W}{T - W}, & W < t \le T \end{cases} \tag{4}$$

where t is the current epoch, $T = 200$ is the total number of epochs and $W = 5$ is the length of the warm-up, with $\lambda_{\max} = 0.1$ and $\lambda_{\min} = 0.01$. The schedule is piecewise linear: λ rises linearly from 0 to 0.1 over the first five epochs, so that the alignment term does not perturb the randomly initialized representation, and then decays linearly to 0.01 at epoch 200, so that the segmentation objective dominates once cross-modal features have been aligned. Both segments are linear in t; the exact values used by the implementation are $\lambda(5) = 0.1$ and $\lambda(200) = 0.01$.

Total loss. The complete training objective combines Dice loss, cross-entropy loss, and the adaptively scheduled OT regularization:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{Dice}} + \mathcal{L}_{\text{CE}} + \lambda(t) \cdot \mathcal{L}_{\text{OT}} \tag{5}$$

3.4. Self-Supervised Pre-Training Strategy

To investigate the interaction effects between self-supervised pre-training and OT regularization, we design an independent Masked Autoencoding (MAE) pre-training scheme [26] for the single encoder. During pre-training, random patches of CT and PET images are masked, and the encoder learns to reconstruct the missing patches from visible context. The masking ratio is 25%, applied to 128^3 patches, and the reconstruction term carries a weight of 0.05 relative to the loss on visible patches. Pre-training ran for 20 epochs of 1200 iterations each at batch size 2, that is, 24000 optimization steps, with a learning rate of 10^{-4}. This is the quantity we report: the number of epochs alone would be ambiguous for the reason given in Section 2.

The pre-training corpus is the same 75 PET-CT cases used downstream, and it is not split by fold: the pre-trained weights are shared by all five folds and are therefore exposed to the images of every case, including those that serve as a fold's validation set. This is a self-supervised objective with no labels, but it is still information from the validation images, so the pre-training arms should be read as an upper bound rather than as a clean cross-validated estimate. After pre-training, the encoder weights are loaded as initialization for the downstream segmentation fine-tuning.

3.5. Experimental Configuration and Evaluation Protocol

To ensure fairness and reproducibility, all experiments uniformly adopt consistent data processing, training strategies, and evaluation protocols. The cohort is a 75-case subset of the AutoPET public dataset [1,10], which comprises 1014 whole-body FDG-PET/CT studies with expert voxel-level lesion annotations and covers more cancer types than the three studied here. Two criteria define the subset. The first is lesion positivity: because the task is lesion segmentation, only studies containing at least one annotated tumor lesion were eligible, and lesion-negative studies were excluded. The second restricts the cohort to the three cancer types covered by the DpDNet prompt bank—melanoma, lymphoma, and lung cancer, which the dataset records in the case-identifier prefix. Within each type, cases were drawn at random in numbers chosen to balance the three types, giving 23 melanoma, 25 lymphoma, and 27 lung-cancer cases. The subset is therefore balanced by design rather than being a random sample of the source collection: the draw used no information beyond the two criteria above, the case list was fixed before any model in this study was trained and has not been modified since, and no case was added or removed on the basis of model performance. All images are resampled to a common resolution and normalized to zero mean and unit variance. The training pipeline is constructed based on the nnU-Net v2 self-configuring framework [30], with automatic preprocessing, resampling, and data augmentation configuration. All experiments employ 5-fold cross-validation, training for a full 200 epochs with a batch size of 2. The optimizer is stochastic gradient descent with Nesterov momentum, momentum 0.99, weight decay 3×10^{-5}, and a polynomial learning-rate schedule with exponent 0.9 over the 200 epochs; the initial learning rate is 10^{-4}, below the nnU-Net default of 10^{-2}. All arms, including the OT and pre-training arms, use this identical recipe, so that a difference in convergence behavior is attributable to the architecture or regularizer under test rather than to the optimizer.

We state the convergence criterion explicitly, because the claim that a comparison is made at convergence is otherwise unfalsifiable. For every run we form the exponential moving average of the per-epoch validation Dice—the patch-level Dice computed over the validation patches during training and used here for model selection, which is a different quantity from the per-case Dice of the result tables—with $\alpha = 0.9$, which suppresses the epoch-to-epoch fluctuation of the validation pass, and we call the first epoch e after which no later epoch improves

that average by more than $\delta = 0.005$ Dice the *plateau point* of the run. The test is necessarily retrospective—it is evaluated once the run has finished—but it is stated in advance of the runs it judges and it is applied identically to all arms.

The definition has a property that limits what it can prove, and we state it rather than leave it to be found. The last epoch of any finite run satisfies the condition vacuously, because no epoch follows it, so a run cannot fail the test and the bare observation that all 30 runs pass is uninformative on its own. What the numbers carry is where the plateau point falls and what the criterion permits after it. Across the 30 runs reported here the plateau point ranges from epoch 116 to epoch 197, so no configuration in the result tables is read at a partial-progress budget; and because the same δ defines the plateau point, no epoch after it raises any run's metric by more than 0.005 Dice. That ceiling is what we rely on, not the size of the remaining budget: tightening δ to 0.002 moves the latest plateau point only to epoch 198, but the last runs to plateau would then have almost no margin, so we state the criterion and the ceiling rather than claim that 200 epochs is a generous budget.

Results are reported at two levels, and the two are not interchangeable. At the *fold level* we first average the per-case metric within each fold and then take the mean and sample SD over the five folds ($n = 5$); every significance statement in this paper is made at this level. The aggregate Dice quoted in the tables is the mean of the five fold means, not the mean over the 75 pooled per-case values. Pooling would give 75 predictions produced by five different fold models, which are not independent samples and would understate the variance; for the reference dual-encoder configuration the pooled per-case average is 0.7070 against a fold-level mean of 0.7093.

Segmentation quality is measured by Dice, by intersection over union, and by the 95th-percentile Hausdorff distance in millimetres. Detection quality is measured separately and at the lesion level, after connected-component analysis of both the reference and the prediction: recall at any overlap, recall at 50 percent overlap, lesion precision, and the number of false-positive lesions per case. Two conventions are fixed in advance and applied identically to every arm: an empty prediction scores zero precision rather than being discarded, since $tp/(tp + fp)$ is defined as 0 when the numerator and denominator vanish together; an empty reference produces no defined sensitivity, and those cases are excluded from the sensitivity average. The per-fold number of contributing cases is reported alongside the multi-metric table.

The 95th-percentile Hausdorff distance characterizes the upper tail of the surface-distance distribution rather than average overlap. Because it is an upper-tail statistic it can be inflated by sufficiently extensive distant false-positive components even when the bulk of the boundary is accurate, and in whole-body PET-CT the largest values here are in fact set by spurious predicted components far from any reference lesion rather than by imprecise boundaries, so we do not read it as a measure of boundary precision.

The five folds are not stratified by cancer type, and we state their composition rather than asserting otherwise. Fold 0 contains a single lymphoma case, and it is the fold on which every pilot ablation and every pre-training arm was run; this is why the per-cancer table reports fold-level dispersion beside each cancer type, and why we attach no significance claim to any single cancer type. The full per-fold, per-cancer breakdown is given in the supplementary material.

Statistical significance is assessed with a paired t-test over the five folds, comparing the fold-level means of two arms with four degrees of freedom, and every effect size is reported with its 95% confidence interval as well as its p value. At $n = 5$ the test has low power, so no result is treated as established on the strength of a p value alone. The intervals carry more information here than the p values do: with four degrees of freedom the interval half-width is 2.776 times the standard error of the paired differences, which is 0.019 to 0.042 Dice across the comparisons we report, so a non-significant difference states that the data are consistent with a range of effects that includes zero rather than that the effect is zero. Accordingly we do not read a non-significant comparison as establishing equivalence between two arms; the intervals in the results state how large an effect the data exclude. The between-fold SD printed in the tables and figures describe how much each configuration varies across the five splits; they are not the reference variability for any comparison between configurations, which is the SD of the paired fold-level differences quoted with each interval. Where a family of comparisons is made, we report the number of tests and the corresponding Bonferroni threshold.

4. Experimental Results

4.1. Main Results: Architecture versus Regularization at 200 Epochs

Table 1 presents the 5-fold cross-validation results for all six 200-epoch configurations we ran. Four findings follow: the first two concern the architectural comparison, the third the regularization term, and the fourth, on the metric that does separate these configurations, is given in Section 4.5.

(1) **Independent modality-specific encoder parameters do not explain the architectural advantage.** Our starting comparison set a 22 M single encoder against a 33 M dual encoder, so it could not separate the

fusion architecture from the parameter count. Against single-encoder controls of matched capacity the picture changes. Relative to the *weight-shared two-stream* encoder, which shares one set of encoder weights across the two modalities and is matched to 33.07 M and to within 1% of the FLOP budget, the dual encoder's fold-level advantage is 0.0076 Dice (paired over the five folds, $t = 0.55$, df $= 4$, $p = 0.61$, 95% CI -0.031 to $+0.046$; both arms without OT)—not significant, and consistent with an effect of up to 0.05 Dice rather than showing the effect to be zero. Relative to the *parameter-matched early-fusion* single encoder, which consumes a channel-concatenated CT+PET tensor at 33.10 M but costs 1.24 times the FLOPs, the dual encoder is higher by 0.0465 Dice ($t = -3.98$, $p = 0.016$, 95% CI $+0.014$ to $+0.079$, again both arms without OT)—the only comparison in this table whose interval excludes zero. The two controls therefore lead to different conclusions rather than to one. The second control differs from the dual encoder in fusion position and in computational cost at the same time—and the FLOP excess is itself a consequence of fusing at the input—so that comparison cannot separate the two, and its 0.0465 Dice cannot be assigned to either. Their intervals overlap only between $+0.014$ and $+0.046$ Dice, so five folds do not pin the fusion-position effect to a single value; what they establish is that one interval contains zero and the other does not. The effect of decoupling the two encoders is not separable from the capacity difference in this setting, and these controls do not establish it. Viewed table-wide rather than pair by pair, the five matched-capacity rows are not separated by this metric: all ten of their pairwise fold-level differences are at most 0.017 Dice and none reaches $p = 0.2$. The early-fusion row is the exception, lying below all five others by 0.039 to 0.056 (p from 0.016 to 0.078, uncorrected, two of the five nominally significant).

(2) **OT regularization provides no statistically significant gain under the 200-epoch protocol.** The fold-level difference between the dual encoder with and without OT is 0.0091 when the OT term is the corrected differentiable loss (95% CI -0.033 to $+0.051$, $p = 0.58$), and 0.0080 on the weight-shared two-stream encoder (95% CI -0.019 to $+0.035$, $p = 0.45$). Neither interval excludes zero, and the wider of the two extends to $+0.051$ Dice, so positive effects of up to approximately 0.05 Dice cannot be excluded at this sample size. That is what the data support, and not the stronger claim that the transport term is inert: five folds do not resolve effects of a few hundredths of a Dice. Note first that the two columns we report for the dual encoder differ only in the OT gradient: the OT term of that configuration was detached from the encoder, so it contributed no gradient at all. Correcting that defect does not change the outcome.

(3) **Cross-fold variance does not separate the configurations.** The direction of the dispersion difference depends on which single-encoder control is used: the weight-shared two-stream encoder has the *smaller* inter-fold SD (0.068 with OT, 0.077 without) than the dual encoder (0.075 to 0.081), while the early-fusion single encoder has the largest (0.099). We therefore report no variance claim.

Figure 2 presents these comparisons panel by panel: fold-level mean Dice, the per-fold values whose rankings reverse, per-cancer Dice, and the parameter and FLOP budget.

4.2. Convergence Dynamics: No Early Lead, and a Bounded Endpoint Difference

Figure 3 presents the validation Dice trajectories over the full 200 epochs, averaged over the five folds. The two rows to read are the dual encoder carrying the original, non-differentiable OT term against the same encoder without OT, labeled *Dual encoder, no OT* and *Dual encoder, + OT (original)* in Table 1. On that pair the OT arm is behind at every checkpoint we logged: at epoch 20 it reads 0.5582 against 0.5621 without OT, at epoch 50 it reads 0.6884 against 0.7016 (this 0.0132 gap is the largest of the three checkpoints we quote here), and at epoch 200 it reads 0.7728 against 0.7764. We do not rest the argument on that comparison, for the reason given above: a detached term contributes no gradient, so the original arm and the no-OT arm optimize the same objective and differ only in their random initialization, which makes the few thousandths between them a re-training difference rather than an effect of the transport term. What that pair does establish is that the sign of the early difference is the opposite of the one a convergence-acceleration account predicts; it cannot establish whether the transport term accelerates convergence. The corrected arm provides the relevant comparison. Its five-fold trajectory stays within 0.021 Dice of the no-OT arm at every epoch after the tenth, which is no clear evidence of a substantial convergence-speed advantage at the resolution these curves afford, and its endpoint difference is the 0.0091 recorded above, whose interval extends to $+0.051$ Dice.

A separate single-fold, 20-epoch pilot of an earlier OT variant shows how sensitive a short readout is. In that pilot the OT arm finished at 0.4992 against 0.4368 for the no-OT arm—a lead of 0.062, in the opposite direction from the five-fold curves—and it was run under a different schedule and on one fold. That lead does not persist at 200 epochs: on that fold the same two configurations finish at 0.6407 and 0.6406 by per-case Dice, and on the five-fold patch-level curves the OT arm is behind at every checkpoint by at most 0.013. The pilot did, however,

read at 77.9% of its own converged value on the same fold, which is the quantitative form of the concern this paper raises: a 20-epoch readout is a partial-progress measurement, and a lead it produces need not survive convergence.

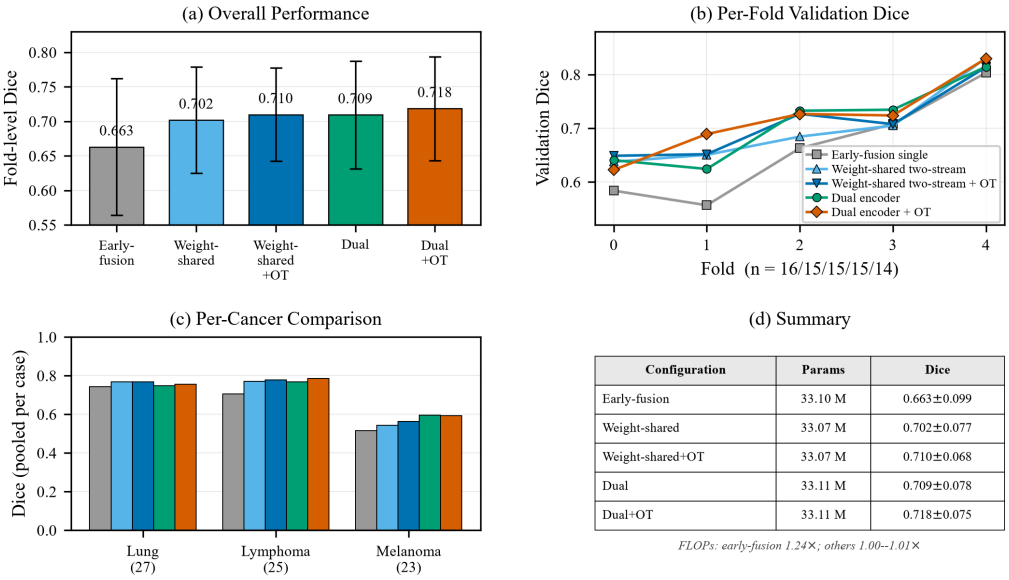

Figure 2. Summary of the 200-epoch comparison across five of the six configurations of Table 1, omitting the dual encoder with the original non-differentiable OT term. (**a**) Fold-level mean Dice, with the SD taken over the five folds. (**b**) Per-fold validation Dice; the ranking of the configurations reverses between folds. (**c**) Per-cancer Dice, pooled per case. (**d**) Parameter counts and fold-level mean Dice. Every value is computed from the same artifacts as Tables 1 and 2. The two matched-capacity control configurations are specified in Section 3.2; neither is the original 22 M baseline encoder, and their parameter counts are given in panel (**d**).

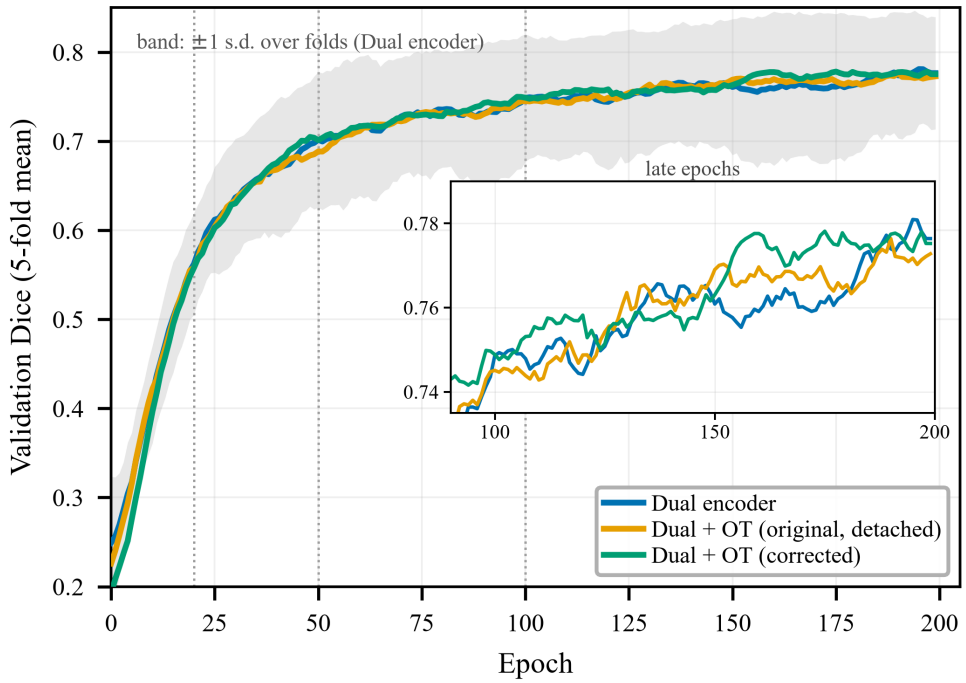

Figure 3. Five-fold mean validation Dice over the full 200-epoch schedule for the three dual-encoder configurations; the vertical axis is the patch-level Dice computed on the validation patches during training, not the per-case Dice reported in Table SI—for the dual encoder without OT the two read 0.776 and 0.709 on the same runs. The shaded band is one SD over the folds for the dual encoder without OT. Against the original, non-differentiable OT term the largest separation is 0.0132 Dice at epoch 50, with the OT arm behind. The corrected, differentiable OT term is not separated either: after the tenth epoch it stays within 0.021 Dice of the no-OT arm at every epoch. Neither separation is significant (inset, late epochs).

Table 2. Per-cancer Dice at 200 epochs. Values are pooled per-case means; SD is taken over the five fold means. The folds are not stratified—fold 0 holds a single lymphoma case—so no per-cancer cell supports a significance claim.

Cancer (n)	Early-Fusion	Weight-Shared Two-Stream	Weight-Shared Two-Stream + OT	Dual	Dual + OT
Melanoma (23)	0.515	0.542	0.563	0.594	0.594
Lymphoma (25)	0.705	0.770	0.777	0.767	0.786
Lung cancer (27)	0.742	0.767	0.767	0.748	0.755
Fold-level SD (melanoma)	0.254	0.185	0.212	0.216	0.154
Fold-level SD (lymphoma)	0.068	0.050	0.058	0.050	0.059
Fold-level SD (lung)	0.087	0.071	0.067	0.086	0.075

Pilot screening protocol. Because pilot runs are cheap and are genuinely useful for triage, we state explicitly how the 20-epoch numbers in this paper should and should not be used. The 20-epoch pilot reproduces between 68% and 78% of the converged Dice on the same fold, and, as the paragraph above shows, the lead it gives the OT arm over the no-OT arm does not persist to convergence: 0.062 Dice at 20 epochs against 0.0001 at 200. We therefore use short pilots only to eliminate configurations that fail outright—a crash, a collapsed prediction, a clear regression—and never to rank configurations that survive. Every comparative claim in this paper rests on the 200-epoch runs.

4.3. Per-Cancer Analysis

Table 2 presents the per-cancer Dice. We report the pooled per-case mean for each cancer type, together with the fold-level SD, because the five folds are not stratified (Section 3.5) and fold 0 contains a single lymphoma case. Table 2 reports the per-cancer values measured on the final artifacts. Figure 4 shows the same values with the fold-level error bars.

(1) **Melanoma is the hardest cancer type and the noisiest.** Melanoma has by far the lowest Dice of the three (0.515 to 0.594) and by far the largest fold-to-fold spread (0.154 to 0.254, against 0.050 to 0.068 for lymphoma and 0.067 to 0.087 for lung cancer). It typically presents diffuse PET uptake and ambiguous CT boundaries, which is consistent with both observations. Given that spread, the melanoma columns should not be read as a ranking: the 0.079 gap between the early-fusion and dual configurations is about a third of the melanoma fold-level SD, and the 0.052 gap between the weight-shared two-stream and dual configurations is a fifth of it.

(2) **Lymphoma is where the parameter-matched early-fusion control loses most.** Its Dice drops to 0.705 from 0.767 to 0.786 in the other four configurations, a gap of about 0.07, and lymphoma's fold-level spread is the smallest of the three cancer types (0.05 to 0.07), so this is the one per-cancer contrast that is large relative to its own dispersion. It is still a descriptive observation on 25 cases and we make no significance claim. Lymphoma frequently involves multiple lymph nodes with strong morphological and metabolic heterogeneity, which is a plausible reason for a fused single-stream representation to be at a disadvantage, but we have not tested that explanation.

(3) **Lung cancer separates the configurations least.** All five configurations land between 0.742 and 0.767, a range of 0.025, narrower than the fold-level SD of any of them. Lung cancer typically has clear anatomical boundaries on CT, which is a plausible reason for the insensitivity, but on this cohort we can say only that the metric does not distinguish the configurations.

(4) **OT yields no gain in any cancer type.** The OT term changes per-cancer Dice by at most 0.021 in either direction, in every case by less than the corresponding fold-level SD, which is consistent with the aggregate finding.

4.4. Exploratory Analysis: Pre-training and Regularization Interaction

Table 3 gives the MAE arms measured at 200 epochs but on a single fold, and two properties of those arms matter for reading them. All three are built on the 22.17 M single encoder, and no dual-encoder MAE arm was run. They are also fold-0 measurements only, so they are not comparable to the 20-epoch numbers. We report this as an exploratory single-fold analysis rather than as one of the paper's primary findings.

Three observations. (1) On fold 0 the MAE and OT combination is 0.008 below MAE alone, which is not a degradation we can distinguish from run-to-run variability—these arms are single-seed, and two independent runs of the same architecture elsewhere in this study differ by 0.0101 Dice; (2) The ordering is not even stable across the two pre-training orders: applying OT before MAE instead of after it reads 0.639, that is, 0.022 above MAE alone, again within noise but in the opposite direction; (3) The 20-epoch reading of the same interaction has a per-cancer decomposition on fold 0—melanoma falling from 0.347 to 0.207, lung rising from 0.509 to 0.587, one lymphoma

case—which is to say that its aggregate penalty was one cancer type falling while another rose, on 9 and 6 cases respectively. Subgroups of that size cannot support any statistical claim, and we offer no causal explanation for the effect. The three mechanistic hypotheses are retained only as candidate explanations for a small single-fold observation, and are listed in Section 5.2. Figure 5 shows these fold-0 arms at 200 epochs.

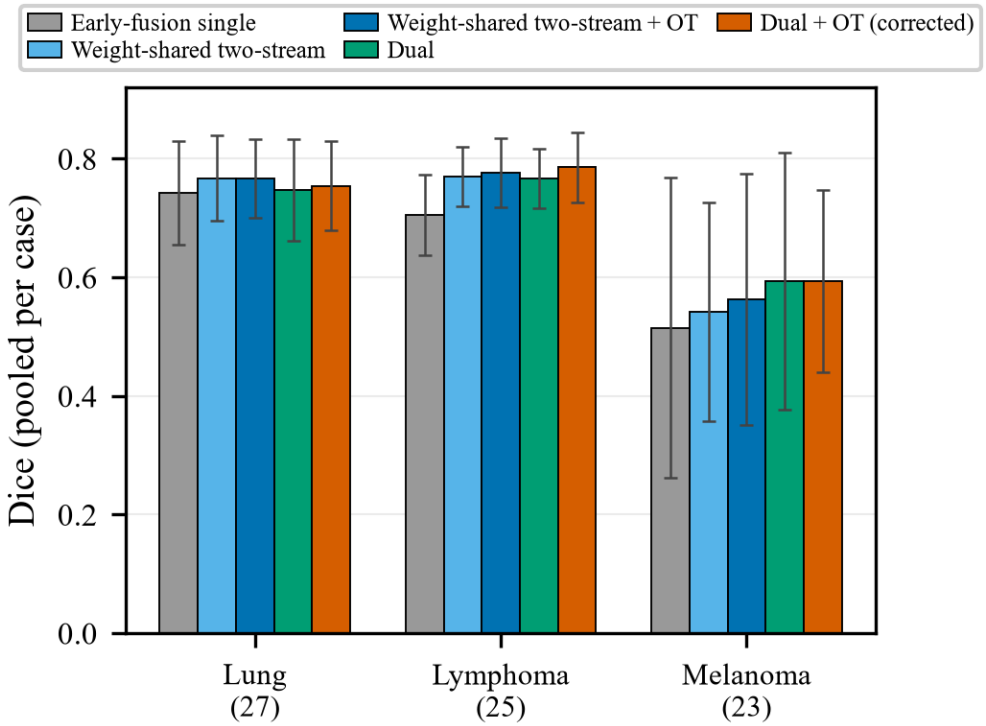

Figure 4. Per-cancer Dice across the five 200-epoch configurations of Table 2. Melanoma is both the lowest-scoring and the most dispersed of the three cancer types. Error bars are the SD over the five fold means, the same dispersion the table reports beside each cancer type.

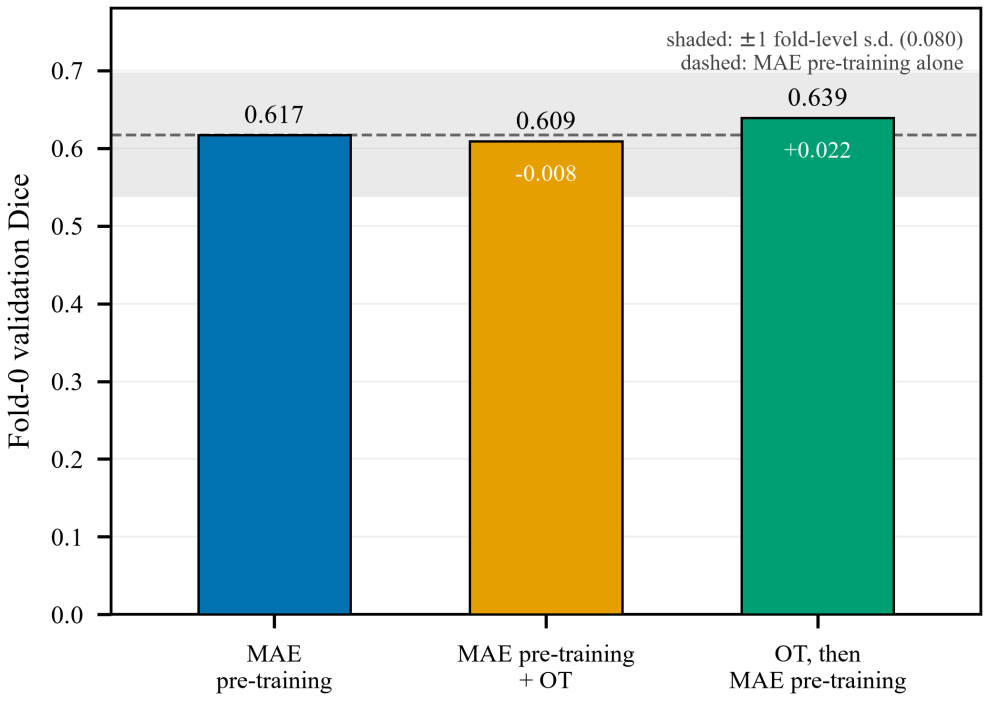

Figure 5. Single-fold (fold 0) MAE and OT interaction at 200 epochs. The MAE and OT combination is 0.008 below MAE alone and the reverse ordering is 0.022 above it, so the data do not establish an antagonistic interaction.

Table 3. MAE pre-training on fold 0 at 200 epochs. All rows are the 22.17 M single-encoder (early-fusion) network. No dual-encoder MAE arm and no no-pre-training cell at this capacity was run, so the pre-training effect is read within the MAE rows. Fold 0 contains 16 cases (6 lung, 1 lymphoma, 9 melanoma).

Configuration (22.17 M Single Encoder)	Fold-0 Dice	vs. MAE Alone
MAE pre-training	0.617	—
MAE pre-training + OT	0.609	−0.008
OT then MAE pre-training	0.639	+0.022

No per-case MAE comparison was run in this study, so no per-case evidence bears on the interaction: the per-cancer figure contains no MAE arm, and the per-case figure compares the dual encoder against the dual encoder with OT, likewise without MAE. We therefore make no claim of a systematic effect.

4.5. Multi-Metric and Lesion-Level Analysis

Table 4 repeats the comparison on eight further metrics. The reason is that Dice aggregates over voxels and is dominated by the largest structures in a whole-body scan, so a configuration can recover additional small lesions while its Dice moves imperceptibly.

Table 4. Multi-metric comparison at 200 epochs. Values are fold-level means ± SD over the five folds ($n = 5$; per-fold case counts 16/15/15/15/14). Sensitivity excludes empty-reference cases and HD95 excludes cases with no defined distance. HD95 is in millimetres and is set by distant spurious components rather than by imprecise boundaries, so it is not a boundary-precision measure.

Metric	Early-Fusion	Weight-Shared	Weight-Shared + OT	Dual	Dual + OT
Dice	0.663 ± 0.099	0.702 ± 0.077	0.710 ± 0.068	0.709 ± 0.078	0.718 ± 0.075
IoU	0.550 ± 0.088	0.587 ± 0.076	0.594 ± 0.065	0.592 ± 0.076	0.603 ± 0.075
HD95 (mm)	127.4 ± 48.1	105.3 ± 50.9	126.5 ± 28.4	91.0 ± 46.4	98.8 ± 50.9
Precision	0.734 ± 0.067	0.780 ± 0.042	0.770 ± 0.054	0.769 ± 0.055	0.785 ± 0.046
Sensitivity	0.714 ± 0.083	0.738 ± 0.075	0.743 ± 0.070	0.753 ± 0.058	0.742 ± 0.079
Lesion recall (any)	0.722 ± 0.018	0.725 ± 0.052	0.732 ± 0.056	0.740 ± 0.051	0.730 ± 0.057
Lesion recall (50%)	0.605 ± 0.053	0.606 ± 0.076	0.605 ± 0.108	0.621 ± 0.062	0.632 ± 0.094
Lesion precision	0.588 ± 0.120	0.632 ± 0.096	0.622 ± 0.112	0.648 ± 0.083	0.640 ± 0.122
False-positive lesions/case	5.51 ± 2.58	4.86 ± 2.70	4.93 ± 2.20	4.49 ± 1.41	4.22 ± 1.94

Voxel overlap does not separate these configurations; lesion detection shows nominal differences only. On Dice, all four comparisons among the matched-capacity configurations—the dual encoder against the weight-shared two-stream encoder in each OT condition, and the OT effect within each of the two—lie within 0.0091 of one another and none is significant (p from 0.45 to 0.61). At the lesion level nominal encoder-related differences appear, though not on the same metric in both OT conditions: without OT the any-overlap lesion recall difference between the dual and weight-shared two-stream encoders is $+0.0156$ ($p = 0.042$), and with OT the 50%-overlap difference is $+0.0271$ ($p = 0.021$). Separately, the parameter-matched early-fusion encoder recovers fewer lesions precisely than the weight-shared two-stream encoder does (-0.0443, $p = 0.034$). None of these three survives correction for the ninety tests run in this section, as stated below.

We ran 90 tests across this section—all ten pairwise comparisons among the five configurations of Table 4, on each of the nine metrics in that table. Ten of the ninety are nominally significant at 0.05, against about four or five expected by chance at this family size; none survives the corresponding Bonferroni threshold of 0.00056, and the tests are not independent, since the pairwise comparisons share arms and the metrics are correlated. We therefore report the lesion-level effects as nominal trends rather than as established findings. They are nonetheless the concrete form of the limitation: a study that reported only Dice would have concluded that these five configurations are equivalent. Figure 6 gives the per-case view of the same comparison.

False-positive lesions per case. The five configurations produce 4.2 to 5.5 false-positive lesions per case, and the differences track the Dice ordering: the early-fusion encoder has both the lowest lesion precision and the highest false-positive count, while the dual and weight-shared two-stream encoders are within 0.8 of one another.

The Hausdorff distances must not be read as boundary accuracy. The fold-level HD95 values are 91 to 127 mm, which is implausible as a boundary error for lesions of clinical interest. Across the five configurations

compared in this subsection the per-case values run from 2.0 mm to 1090.8 mm, with a mean of 110.6 mm; 375 per-case predictions are involved, 368 of them with a defined distance, and 67 of those 368 exceed 200 mm. The large values are not boundary imprecision. They are set by small spurious predicted components lying far from any reference lesion, in cases that contain few reference lesions: the 67 values above 200 mm carry 8.2 false-positive lesions per case on average against 4.1 for the remaining 301, and 58 of them have more predicted components than the reference has lesions. A case whose prediction or reference is empty has no defined distance at all, and is excluded rather than counted as large. HD95 and Dice therefore encode much of the same information on this dataset, and we report HD95 for completeness rather than as independent evidence.

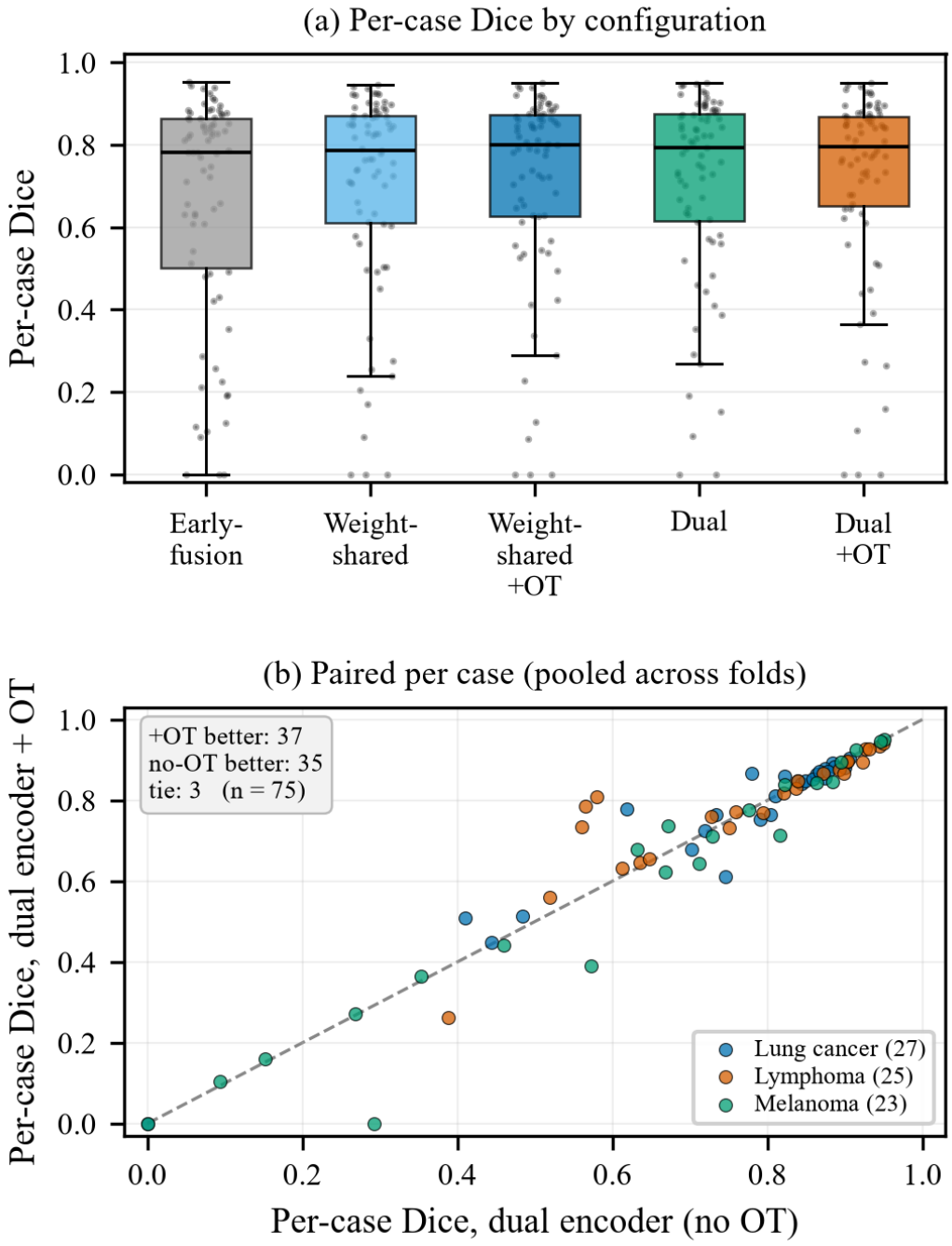

Figure 6. Per-case paired analysis. (**a**) Distribution of per-case Dice for the five configurations of Table 2, with individual cases overlaid; the boxes are markedly wider than the fold-level SD, which is what pooling 75 non-independent predictions does. (**b**) Paired per-case comparison of the dual encoder with and without OT, complementing the fold-level test in the text. The vertical axis is the corrected, differentiable OT arm—the *Dual encoder, + OT (corrected)* row of Table I—not the original detached one. Of the 75 pairs the corrected OT arm is higher on 37, lower on 35 and tied on 3, so the points scatter about the identity line rather than lying on it.

4.6. Qualitative Visualization

Figure 7 shows four difficult cases selected to illustrate the failure modes discussed above. Because the quantitative differences are small, we use these panels to show where the residual differences sit rather than as independent evidence. The micro-lesion panel is the relevant one: the early-fusion single-encoder prediction contains 5.5 false-positive lesions per case on average against 4.5 for the dual encoder, so at this scale of lesion the two configurations differ mainly in which small structures they fire on, not in whether the dominant lesion is found. The OT and no-OT predictions are visually indistinguishable, which is consistent with their fold-level difference of 0.009.

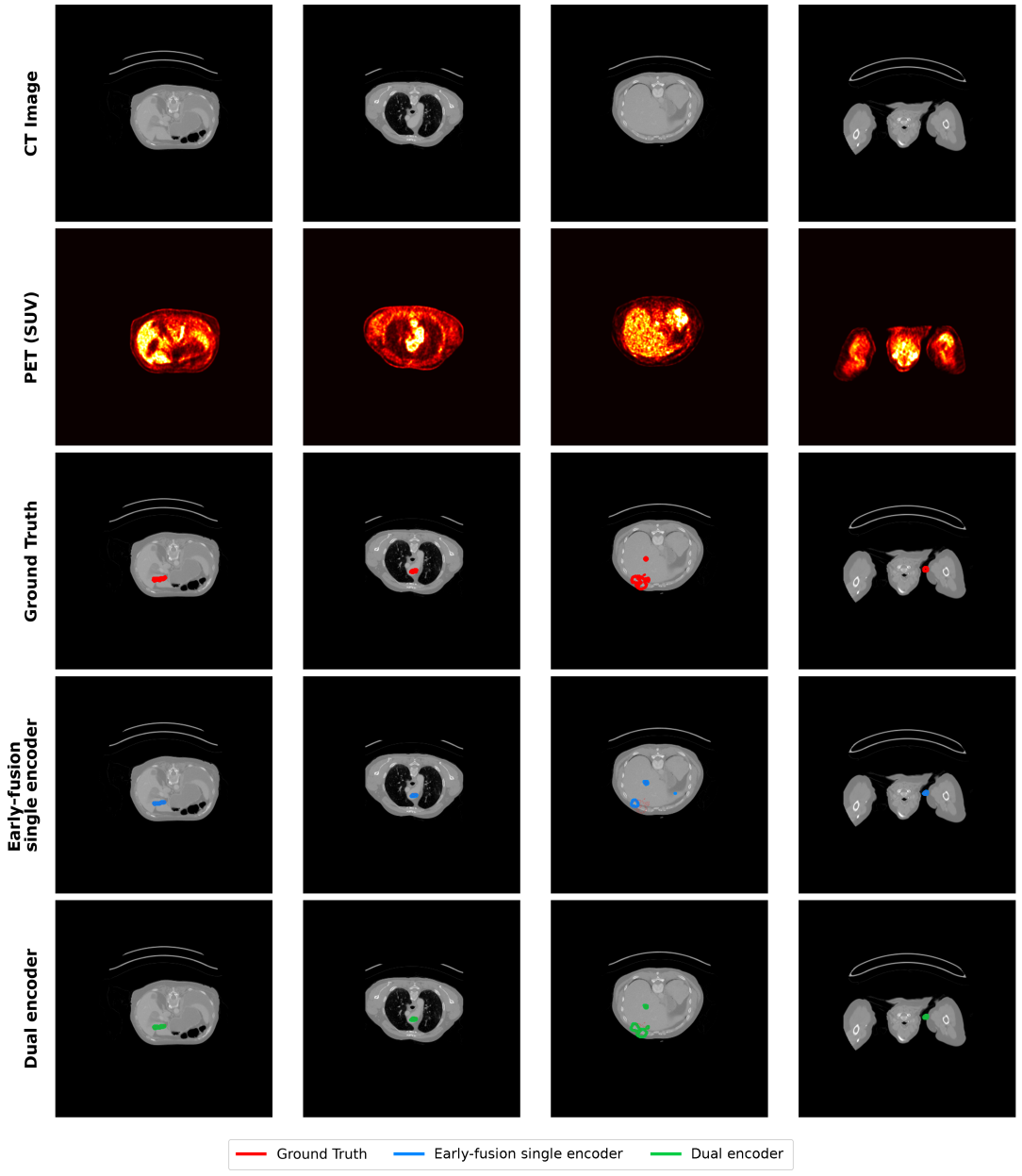

Figure 7. Qualitative segmentation comparison on four representative difficult cases. From top to bottom: CT, PET (SUV), the ground-truth delineation, the early-fusion single-encoder prediction and the dual-encoder prediction. These panels illustrate failure modes; they are not evidence of a performance difference, which on voxel Dice is not distinguishable between the configurations.

4.7. Per-Fold Reproducibility

Figure 8 displays the per-fold validation Dice across the five splits. There is no consistent winner: the dual encoder is ahead of the weight-shared two-stream encoder on three folds and behind on two. This is the per-fold form of the aggregate result in Table 1 and is the clearest single view of why the aggregate difference of 0.0076 is not meaningful.

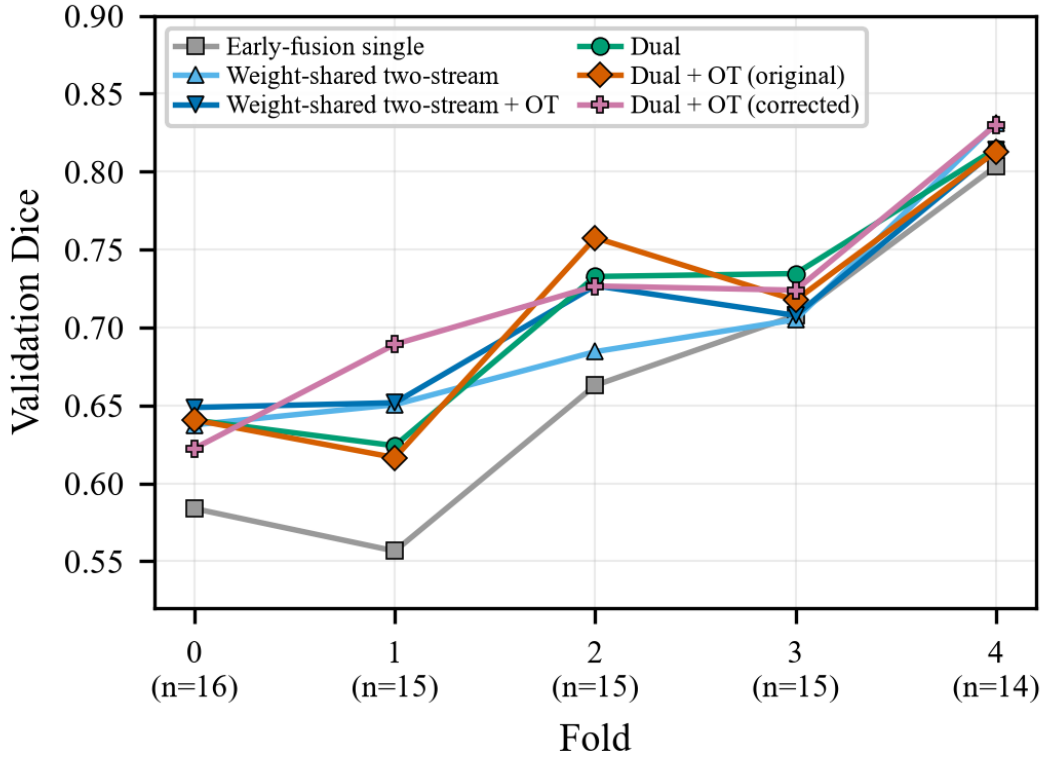

Figure 8. Per-fold validation Dice across the five splits, for the six configurations of Table 1. The rankings reverse between folds, so no configuration dominates across partitions. Fold sizes are 16, 15, 15, 15 and 14 cases, and the folds are not stratified by cancer type.

5. Discussion

5.1. Why the Architecture Effect Is Smaller Than Reported

The finding that requires explanation is a narrow one. Against single-encoder controls of matched capacity the dual encoder does not separate from a weight-shared two-stream encoder (0.0076 Dice, $p = 0.61$), while it is ahead of a parameter-matched early-fusion encoder by 0.047 Dice; OT contributes no late-training gain in either OT condition. The measurements therefore do not support a simple dual- versus single-encoder explanation. The surviving difference is between the two fusion positions, and that comparison changes the computational cost at the same time, so the available experiments do not identify whether it is associated with the fusion position, with the FLOP budget, or with both. The four interpretations below are candidates rather than established mechanisms. Two of the interpretations below are not supported by those controls, and we label them accordingly rather than leaving them as an unexamined list.

(1) **Information bottleneck theory: not supported by the controls.** A single shared encoder requires CT and PET to share one set of convolutional kernels, which plausibly creates a cross-modal information bottleneck [3,5]. The weight-shared two-stream control is exactly such an encoder: it pushes both modalities through the same weights. If the bottleneck were the operative constraint, that arm should fall behind the dual encoder as far as the early-fusion arm does. It does not. The weight-shared two-stream control stays within 0.0076 Dice of the dual encoder while consuming 1.01 times the FLOPs, whereas the early-fusion arm—matched in parameters but carrying 1.24 times the FLOPs—is the one that falls behind. The bottleneck is therefore not the operative constraint in these experiments. What the two single-encoder controls show is that the arm fusing at the input is the one that pays a cost, and that arm differs from the weight-shared two-stream control in fusion position and in computational cost at once, so we cannot separate the two.

(2) **Sufficiency of decoder-level implicit alignment.** The segmentation supervision task itself drives the shared decoder to map CT and PET features into the same tumor segmentation mask space, a process that accomplishes cross-modal fusion without requiring explicit alignment constraints [30]. This interpretation is the one the controls strengthen rather than weaken: an explicit alignment term applied on top of a decoder that is already supervised to fuse the two modalities changes fold-level Dice by 0.009 or less in both architectures, which is what one would expect if the supervision had already done the work.

(3) **Priority of modality-specific representations, qualified.** Dedicated feature extractors for each modality and representations adapted to each modality's distribution [5,6] remain a reasonable design principle, but the evidence here does not rank it above weight sharing: the weight-shared two-stream encoder, which gives up modality-specific encoders entirely, matches the dual encoder. What the controls bear on is not whether the encoders are dedicated—the weight-shared two-stream arm gives them up entirely and still matches the dual encoder—but where fusion happens, and the arm that fuses at the input is also the arm that computes more, so the two cannot be separated here.

(4) **Phase-dependent regularization: no early advantage is present.** The five-fold curves show no early advantage for the OT term: the original, non-differentiable OT configuration is behind at every checkpoint we logged, including the earliest (0.5582 against 0.5621 at epoch 20), and the corrected differentiable term does not separate from its own no-OT arm either, so there is no early advantage for a phase-dependent account to explain. The late-training part of the argument—that the term becomes redundant—remains consistent with the measurements, but on its own it is a much weaker statement, and we do not present it as an explanation the data support.

What the controls do support. Of the four interpretations, two survive contact with the matched-capacity controls. The decoder's supervision already performs the cross-modal fusion, so an explicit alignment term adds nothing measurable (item 2); and among the single-encoder designs, the one that keeps the original input-level fusion pays a real cost—0.047 Dice and 0.044 lesion precision—for carrying both modalities through every layer. That cost is accompanied by the FLOP budget rather than by the parameter count, since the early-fusion arm and the weight-shared two-stream arm have the same parameters (33.10 M and 33.07 M) and differ in that the former computes 1.24 times as much—though the two are not separable in this design, because the excess FLOPs are themselves a consequence of fusing at the input. We state this as the reading the measurements favor, not as a demonstrated mechanism: a FLOP-matched control would be needed to close the argument, and we do not have one.

5.2. Candidate Mechanisms for the MAE–OT Interaction

On the evidence now available there is no antagonistic effect to explain. At 200 epochs the combination of MAE pre-training and OT is 0.008 below MAE alone on fold 0, and reversing the order of the two reads 0.022 above it, so the two arms bracket zero. The three mechanisms below are stated as candidates for a larger study to test, not as explanations of a demonstrated effect; they are unverified rather than refuted. The 0.008 difference is smaller than the 0.0101 Dice gap we measured between two independent runs of the same architecture in this study, so it is not a difference we can distinguish from re-training variability; the arms are single-seed and their pre-training length was not varied, as stated in Section 5.4.

(1) **Redundant constraint hypothesis.** MAE masked reconstruction pre-training has already enabled the encoder to learn cross-modal base features with relatively matched distributions. Superimposing additional OT enforced alignment constraints downstream may only perturb these already-aligned representations rather than improving them.

(2) **Gradient conflict hypothesis.** The optimization gradients corresponding to the parameter initialization conferred by MAE pre-training exhibit deviation from—or even oppose—the gradient direction of the OT alignment loss. When the two are jointly back-propagated, gradient interference may prevent both objectives from being optimized effectively.

(3) **Hierarchical matching mismatch hypothesis.** MAE pre-training predominantly focuses on learning shallow, low-level visual texture features, whereas OT regularization acts on high-level semantic tumor feature alignment [26,29]. The two operate at incompatible levels of the feature hierarchy, and forcing alignment at deeper layers may undo the beneficial shallow representations learned during pre-training.

The present experiments cannot test any of the three, because the effect they would explain is not present in the data: all three MAE arms are single-fold, and the two orderings differ in sign. The practical statement we can support is the narrow one: on fold 0, at 200 epochs, adding OT on top of MAE pre-training did not improve Dice, while reversing the order of the two improved it by 0.022. That is a single-fold observation and we do not extend it to a recommendation.

5.3. Per-Cancer Behavior of the Configurations

The per-cancer numbers do not support a simple dual-versus-single-encoder pattern. Measured against the matched-capacity controls, the dual encoder is ahead of the weight-shared two-stream encoder only on melanoma (0.594 against 0.542 and 0.563), behind it on lung cancer (0.748 to 0.755 against 0.767), and level with it on

lymphoma, so the ordering that appears is between the fusion positions rather than between one encoder and two. The early-fusion single encoder is the weakest configuration on lymphoma (0.705 against 0.767 to 0.786 for the other four) and on melanoma (0.515 against 0.542 to 0.594), and lung cancer separates the configurations least of all (0.742 to 0.767). We offer the three following readings as candidates, not as established subgroup effects; with 23, 25 and 27 cases and unstratified folds, no per-cancer contrast in this study can carry a significance claim.

(1) **Where modality fusion happens may matter more for some cancer types than others.** Melanoma and lymphoma show high PET metabolic heterogeneity and ambiguous CT boundaries, and they are the two cancer types on which the input-fused single encoder falls furthest behind; lung cancer, whose CT boundaries are clearer, is where the configurations are hardest to tell apart. The direction is consistent with the complementarity of the two modalities varying by cancer type, but a per-cancer interaction of this size on 23 to 27 cases is well inside what these folds can resolve, and we have not tested it.

(2) **No deployment recommendation follows from these results.** A clinical recommendation would rest on per-cancer differences of 0.025 to 0.081 Dice on 23 to 27 cases each, at a fold-level dispersion of 0.05 to 0.25, from a single cohort, with no external validation and no reader study. Nothing in this study supports a statement about which architecture to deploy for which cancer.

(3) **Per-cancer reporting, with the caveat that makes it usable.** Reporting per-cancer numbers alongside the aggregate is worth doing—an aggregate Dice of 0.72 hides that melanoma sits at 0.52 to 0.59 while lung cancer sits near 0.75—but a per-cancer table invites over-reading unless it is accompanied by the fold composition and the per-cancer dispersion, which is why both are given in Table 2, and in full in the supplementary material (Section S-II). The relevant caution is not only that aggregate metrics obscure subgroups; it is that subgroup metrics on unstratified folds can manufacture differences that the aggregate correctly declines to report.

5.4. Limitations and Future Directions

Several limitations of this study should be noted. First, all experiments are based on the AutoPET single-center dataset (75 cases, 3 tumor types). Generalizability to other PET tracers, anatomical sites, and multi-center data requires validation.Tumor-site-specific PET-CT benchmarks on other anatomical sites exist, such as the HECKTOR challenge for head and neck [32], and we have no evidence that the comparison reported here transfers to them. The 75 cases were drawn to balance the three cancer types rather than sampled from the case mix of the source collection, so the cohort-level Dice describes this deliberately balanced cohort and not the distribution of the source data, and each per-cancer number rests on 23 to 27 cases. Second, capacity is controlled only partially, and the residual confound does not run in one direction only. Both single-encoder controls match the dual encoder in parameter count (33.07 M and 33.10 M against 33.11 M), and the weight-shared two-stream control also matches it to within 1% in FLOPs (1.01 times), so the comparison to that control is clean, and the difference it measures is 0.0076 Dice with a 95% interval of -0.031 to $+0.046$—not significant, and consistent with an effect of up to 0.05 Dice rather than showing it to be zero. The early-fusion control is the one that consumes 1.24 times the FLOPs, so the 0.047 Dice by which it trails cannot be attributed to the parameter count—but it also cannot be attributed unambiguously to fusion position, because a FLOP-matched early-fusion model was not built. What the controls show is that independent modality-specific encoder parameters do not explain the architectural advantage; whether the FLOP budget does remains open. Third, this study examines OT as a representative explicit alignment scheme; whether other cross-modal alignment strategies (contrastive learning, mutual information maximization) behave similarly at matched training budgets remains an open question, and we no longer assert the answer. Fourth, the fixed nnU-Net hyperparameter configuration means that OT loss weight scheduling was not subjected to exhaustive grid search, leaving room for potential optimization. The transport estimator itself was compared against sliced Wasserstein and entropic OT on fold 0 only, where the four arms span 0.021 Dice and each arm was trained once, so the comparison carries no re-run dispersion of its own. That screen can exclude gross estimator failure but cannot rank the estimators, and we draw no relative-accuracy conclusion from it (supplementary material, Section S-I). Fifth, the MAE pre-training arms are single-fold and use all 75 cases for pre-training without a fold-wise split, so those results are an upper bound rather than a clean cross-validated estimate, as stated in Section 3.4. Two further limits apply to those arms, and we state them rather than leave them implied. Each MAE configuration was trained once, so we have no seed-level error bar and cannot separate the pre-training effect from run-to-run variability; for scale, two independent runs of the same architecture elsewhere in this study differ by 0.0101 Dice on fold 0 (0.6305 against 0.6406), which is larger than the 0.008 gap between the MAE and MAE+OT arms. Pre-training length was not varied either—24,000 optimization steps is the only setting we trained—so the sensitivity of the interaction to the pre-training budget is untested and we make no claim about it. Both gaps are listed under future work.

A limitation applies to every statistical statement in this paper, and not only to the MAE arms. The tests rest on

five folds, and failure to reject the null hypothesis should not be read as evidence of equivalence: the non-significant results reported here indicate an absence of detected differences at the resolution these five folds afford, not that the true effects are zero. Each comparison is therefore quoted with the interval that says how large an effect the data fail to exclude.

Future work should pursue: (1) multi-center external validation of the architectural comparison; (2) a FLOP-matched early-fusion control, which is the measurement that would separate the fusion position from the computational budget; (3) systematic comparison of alternative alignment strategies at matched training budgets; and (4) multi-seed replication of the MAE arms together with a sweep of pre-training lengths, since the interaction is currently read from one run per configuration at a single pre-training budget and neither quantity has been varied.

5.5. Methodological Recommendations

Three practices make the comparisons in this area more legible, and we offer them as suggestions rather than as standards. All three are drawn from a single cohort of 75 cases under one training framework, and none has been tested as a policy.

(1) **Report at, or near, convergence, and say how convergence was judged.** A fixed epoch count is not the same thing as convergence, and announcing one is not the same as demonstrating it: state the criterion, show that the arms meet it, and report the trajectories. Short pilots remain useful for screening configurations that fail outright; on our data they reproduce 68 to 78 percent of the converged value and their ranking of two arms disagreed with the ranking at 200 epochs, so a pilot should not be used to rank arms that both survive it.

(2) **Report convergence trajectories.** When proposing new regularization or training strategies, authors should provide full training curves showing both the proposed method and the baseline, enabling readers to distinguish convergence acceleration from late-training improvement. Report the curves for the arms as trained, with the sign of any early difference stated explicitly and read from the logs rather than from the mechanism one expects: on our own curves the sign of the early difference is the opposite of the one a convergence-acceleration account predicts.

(3) **Conduct component-level ablation.** Novel methods combining architectural modifications with regularization terms should include ablations that isolate each component's contribution, measured at 200 epochs. Gains observed only at early epochs should be explicitly labeled as convergence effects. Where an architectural modification also changes parameter count or compute, include a control matched on the quantity that changed, and state which quantity the control does *not* match—as we were unable to do for the FLOP budget in this study.

6. Conclusions

This study revisited dual-encoder fusion and OT regularization for multimodal PET-CT segmentation under matched-capacity and extended-training controls. Against a parameter-matched and approximately compute-matched weight-shared two-stream encoder, the dual encoder showed no statistically significant Dice advantage. Although it achieved significantly higher Dice than a parameter-matched early-fusion control, that control required $1.24\times$ the FLOPs. Therefore, independent modality-specific encoder parameterization alone is not sufficient to explain the observed performance, and the remaining difference between late- and early-fusion designs cannot be attributed uniquely to fusion position or computational cost.

OT regularization yielded no statistically significant late-training Dice gain on either encoder, and the pre-training–OT interaction is reported here only as an exploratory 200-epoch fold-0 analysis. Voxel Dice did not distinguish the matched-capacity configurations, whereas lesion-level detection showed only nominal differences that did not survive multiple-comparison correction.

These findings suggest that conclusions about multimodal PET-CT segmentation depend strongly on control design, training duration, and evaluation metric. Within this cohort and framework, the evidence does not establish an independent benefit of modality-specific encoder parameters or additional OT regularization. Future work should use multi-center data, FLOP-matched fusion controls, and repeated-seed experiments to determine whether these observations generalize.

Supplementary Materials

The following supporting information can be downloaded at: https://media.sciltp.com/articles/others/261 0211637080467/AIM-26060273_SM.pdf, Table S-I: Comparison of transport estimators at a single fold; Table S-II: Per-fold cancer-type composition of the cross-validation split.

Author Contributions

M.L.: conceptualization, methodology, software, validation, formal analysis, data curation, writing—original draft preparation, visualization; Y.T.: conceptualization, methodology, software, investigation, writing—review and editing; X.Z.: validation, investigation, writing—review and editing; Y.Z.: data curation, visualization, writing—review and editing; M.W.: supervision, project administration, funding acquisition, writing—review and editing. All authors have read and agreed to the published version of the manuscript.

Funding

This work is supported by the General Program of the National Natural Science Foundation of China (Grant No. 82572327) and the Liaoning Provincial Natural Science Foundation (Ph.D. Startup Project) (Grant No. 2024-BS-258).

Institutional Review Board Statement

Ethical review and approval were waived for this study, due to the secondary analysis of a publicly available, fully anonymized historical dataset (TCIA FDG-PET-CT-Lesions collection). The original data collection and anonymized publication were approved by the local ethics committee and data protection officer at the University Hospital Tübingen, and this study was conducted according to the guidelines of the Declaration of Helsinki.

Informed Consent Statement

Patient consent was waived due to the use of publicly available, de-identified data approved by an ethics committee for public release, and because the PET/CT images displayed in this study contain no identifiable personal features.

Data Availability Statement

The imaging data analysed in this study are publicly available from The Cancer Imaging Archive (TCIA) as the FDG-PET-CT-Lesions (AutoPET) collection, https://doi.org/10.7937/gkr0-xv29. No new imaging data were generated. The source code is not publicly available and can be obtained from the corresponding author on reasonable request.

Acknowledgments

The authors thank the organizers of the AutoPET Challenge for providing the benchmark dataset used in this study.

Conflicts of Interest

The authors declare no conflict of interest.

Use of AI and AI-Assisted Technologies

During the preparation of this work, the authors used ChatGPT, DeepSeek, and Claude to assist with drafting the manuscript text, polishing the language, and preparing the figures (plotting code and layout). After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

References

1. Gatidis, S.; Früh, M.; Fabritius, M.P.; et al. Results from the autoPET Challenge on Fully Automated Lesion Segmentation in Oncologic PET/CT Imaging. *Nat. Mach. Intell.* **2024**, *6*, 1396–1405. https://doi.org/10.1038/s42256-024-00912-9.
2. Litjens, G.; Kooi, T.; Bejnordi, B.E.; et al. A Survey on Deep Learning in Medical Image Analysis. *Med. Image Anal.* **2017**, *42*, 60–88. https://doi.org/10.1016/j.media.2017.07.005.
3. Dolz, J.; Gopinath, K.; Yuan, J.; et al. HyperDense-Net: A Hyper-Densely Connected CNN for Multi-Modal Image Segmentation. *IEEE Trans. Med. Imaging* **2019**, *38*, 1116–1126. https://doi.org/10.1109/tmi.2018.2878669.
4. Kamnitsas, K.; Ledig, C.; Newcombe, V.F.J.; et al. Efficient Multi-Scale 3D CNN with Fully Connected CRF for Accurate Brain Lesion Segmentation. *Med. Image Anal.* **2017**, *36*, 61–78. https://doi.org/10.1016/j.media.2016.10.004.
5. Chartsias, A.; Joyce, T.; Giuffrida, M.V.; et al. Multimodal MR Synthesis via Modality-Invariant Latent Representation. *IEEE Trans. Med. Imaging* **2018**, *37*, 803–814. https://doi.org/10.1109/tmi.2017.2764326.

6. Fu, X.; Bi, L.; Kumar, A.; et al. Multimodal Spatial Attention Module for Targeting Multimodal PET-CT Lung Tumor Segmentation. *IEEE J. Biomed. Health Inform.* **2021**, *25*, 3507–3516. https://doi.org/10.1109/jbhi.2021.3059453.
7. Chen, T.; Kornblith, S.; Norouzi, M.; et al. A Simple Framework for Contrastive Learning of Visual Representations. In Proceedings of the 37th International Conference on Machine Learning, Online, 13–18 July 2020; pp. 1597–1607.
8. Courty, N.; Flamary, R.; Tuia, D.; et al. Optimal Transport for Domain Adaptation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 1853–1865. https://doi.org/10.1109/tpami.2016.2615921.
9. Huang, J.; Chen, S.; Liang, X.; et al. C^2MAOT: Cross-Modal Complementary Masked Autoencoder with Optimal Transport for Cancer Segmentation in PET-CT Images. In *Medical Image Computing and Computer Assisted Intervention—MICCAI 2025, Proceedings of the 28th International Conference, Daejeon, South Korea, 23–27 September 2025*; Springer: Cham, Switzerland, 2025; Volume 15960, pp. 87–97. https://doi.org/10.1007/978-3-032-04927-8_9.
10. Gatidis, S.; Kuestner, T. A Whole-Body FDG-PET/CT Dataset with Manually Annotated Tumor Lesions (FDG-PET-CT-Lesions). Available online: https://www.cancerimagingarchive.net/collection/fdg-pet-ct-lesions/ (accessed on 5 October 2026).
11. Liang, X.; Huang, J.; Han, L.; et al. DpDNet: An Dual-Prompt-Driven Network for Universal PET-CT Segmentation. In *Medical Image Computing and Computer Assisted Intervention—MICCAI 2025, Proceedings of the 28th International Conference, Daejeon, South Korea, 23–27 September 2025*; Springer: Cham, Switzerland, 2025; pp. 163–172. https://doi.org/10.1007/978-3-032-04978-0_16.
12. Neha, F.; Shukla, D.K. Radiomics in Medical Imaging: Methods, Applications, and Challenges. *J. Imaging* **2026**, *12*, 220. https://doi.org/10.3390/jimaging12060220.
13. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015, Proceedings of the 18th International Conference, Munich, Germany, 5–9 October 2015*; Springer: Cham, Switzerland, 2015; pp. 234–241. https://doi.org/10.1007/978-3-319-24574-4_28.
14. Çiçek, Ö.; Abdulkadir, A.; Lienkamp, S.S.; et al. 3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation. In *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2016, Proceedings of the 19th International Conference, Athens, Greece, 17–21 October 2016*; Springer: Cham, Switzerland, 2016; pp. 424–432. https://doi.org/10.1007/978-3-319-46723-8_49.
15. Milletari, F.; Navab, N.; Ahmadi, S.A. V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation. In Proceedings of the 2016 Fourth International Conference on 3D Vision (3DV), Stanford, CA, USA, 25–28 October 2016; pp. 565–571. https://doi.org/10.1109/3dv.2016.79.
16. Neha, F.; Bhati, D.; Shukla, D.K.; et al. An Analytics-Driven Review of U-Net for Medical Image Segmentation. *Healthc. Anal.* **2025**, *8*, 100416. https://doi.org/10.1016/j.health.2025.100416.
17. Hatamizadeh, A.; Tang, Y.; Nath, V.; et al. UNETR: Transformers for 3D Medical Image Segmentation. In Proceedings of the 2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), Waikoloa, HI, USA, 3–8 January 2022; pp. 1748–1758. https://doi.org/10.1109/wacv51458.2022.00181.
18. Cao, H.; Wang, Y.; Chen, J.; et al. Swin-Unet: Unet-like Pure Transformer for Medical Image Segmentation. In Proceedings of the Computer Vision—ECCV 2022, Tel Aviv, Israel, 23–27 October 2022; pp. 205–218. https://doi.org/10.1007/978-3-031-25066-8_9.
19. Chen, J.; Lu, Y.; Yu, Q.; et al. TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation. *arXiv* **2021**, arXiv:2102.04306.
20. Ganin, Y.; Ustinova, E.; Ajakan, H.; et al. Domain-Adversarial Training of Neural Networks. *J. Mach. Learn. Res.* **2016**, *17*, 1–35.
21. Tzeng, E.; Hoffman, J.; Saenko, K.; et al. Adversarial Discriminative Domain Adaptation. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 2962–2971. https://doi.org/10.1109/cvpr.2017.316.
22. Arjovsky, M.; Chintala, S.; Bottou, L. Wasserstein GAN. *arXiv* **2017**, arXiv:1701.07875.
23. Gulrajani, I.; Ahmed, F.; Arjovsky, M.; et al. Improved Training of Wasserstein GANs. In Proceedings of the 31st International Conference on Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; Volume 30, pp. 5767–5777.
24. Peyré, G.; Cuturi, M. Computational Optimal Transport with Applications to Data Sciences. *Found. Trends Mach. Learn.* **2019**, *11*, 355–607. https://doi.org/10.1561/2200000073.

25. Frogner, C.; Zhang, C.; Mobahi, H.; et al. Learning with a Wasserstein Loss. In Proceedings of the 29th International Conference on Neural Information Processing Systems, Montreal, QC, Canada, 7–12 December 2015; Volume 28, pp. 2053–2061.
26. He, K.; Chen, X.; Xie, S.; et al. Masked Autoencoders Are Scalable Vision Learners. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 15979–15988. https://doi.org/10.1109/cvpr52688.2022.01553.
27. Zhou, Z.; Sodha, V.; Pang, J.; et al. Models Genesis. *Med. Image Anal.* **2021**, *67*, 101840. https://doi.org/10.1016/j.media.2020.101840.
28. Azizi, S.; Culp, L.; Freyberg, J.; et al. Robust and Data-Efficient Generalization of Self-Supervised Machine Learning for Diagnostic Imaging. *Nat. Biomed. Eng.* **2023**, *7*, 756–779. https://doi.org/10.1038/s41551-023-01049-7.
29. Tang, Y.; Yang, D.; Li, W.; et al. Self-Supervised Pre-Training of Swin Transformers for 3D Medical Image Analysis. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 20698–20708. https://doi.org/10.1109/cvpr52688.2022.02007.
30. Isensee, F.; Jaeger, P.F.; Kohl, S.A.A.; et al. nnU-Net: A Self-Configuring Method for Deep Learning-Based Biomedical Image Segmentation. *Nat. Methods* **2021**, *18*, 203–211. https://doi.org/10.1038/s41592-020-01008-z.
31. Cuturi, M. Sinkhorn Distances: Lightspeed Computation of Optimal Transport. In Proceedings of the 27th International Conference on Neural Information Processing Systems, Lake Tahoe, NV, USA, 5–10 December 2013; Volume 26, pp. 2292–2300.
32. Andrearczyk, V.; Oreiller, V.; Abobakr, M.; et al. Overview of the HECKTOR Challenge at MICCAI 2022: Automatic Head and Neck Tumor Segmentation and Outcome Prediction in PET/CT. In *Head and Neck Tumor Segmentation and Outcome Prediction, Proceedings of the 3D Head and Neck Tumor Segmentation in PET/CT Challenge, Singapore, 22 September 2022*; Springer: Cham, Switzerland, 2023; pp. 1–30. https://doi.org/10.1007/978-3-031-27420-6_1.