WHAT MAKES HIGH-MAGNIFICATION KNOWLEDGE TRANSFERABLE? A STUDY OF CROSS-RESOLUTION DISTILLATION IN WHOLE-SLIDE IMAGING

Anonymous authors
Paper under double-blind review

ABSTRACT

Cross-resolution knowledge distillation aims to improve low-magnification whole-slide analysis by transferring high-magnification representations, yet the conditions for useful transfer remain unclear. We develop a decomposition-based analysis of teacher access, representation loss, and model excess, motivating three questions: whether **(a) teacher targets help the task, (b) low-magnification students can predict them,** and **(c) slide models benefit from those predictions.** We investigate them through controlled experiments across ten pathology cohorts spanning classification, grading, and survival prediction. In the main comparison, providing teacher regional means alongside native low-magnification features improves downstream performance in all ten cohorts. Direct prediction achieves lower reconstruction error than residual prediction, yet the predicted features underrepresent variation in the teacher targets. Moreover, better reconstruction does not consistently improve downstream scores, and retaining native features changes performance even when the predicted teacher features are held fixed. Together, these findings expose a gap between reconstructing teacher representations and realizing their downstream value. They challenge the sufficiency of reconstruction error as a measure of cross-resolution transfer and provide a diagnostic framework for examining where that transfer breaks down. Future distillation designs must account for both what students can predict and how slide models use those predictions. Code is provided in supplementary material and will be publicly released after review.

1 INTRODUCTION

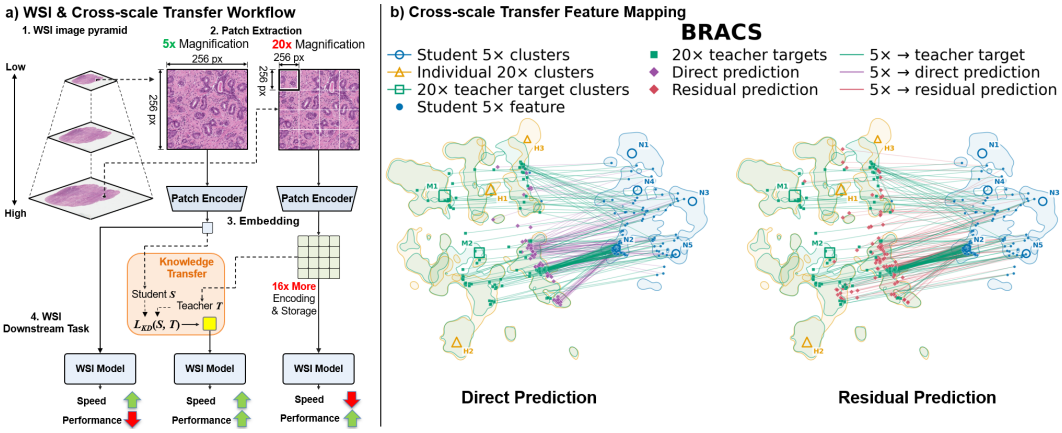


Figure 1: **Cross-resolution transfer in WSI representation learning.** (a) General WSI workflow for improving slide-level prediction through high-magnification supervision while retaining a low-magnification token budget. (b) Direct transfers predict the teacher target from student input directly, while residual transfers predict the difference between the student input and teacher target. In either case, the predictions fail to cover the full diversity of their high-magnification teacher targets.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

It is believed that knowledge transfer across resolutions can improve low-magnification task models in computational pathology (Su et al., 2025; Shin & Jeong, 2026), but the conditions that make high-magnification supervision transferable remain poorly understood. Slide models aggregate large bags of patch features for diagnosis and prognosis (Ilse et al., 2018; Xu et al., 2024; Wang et al., 2024; Hosseini et al., 2024; Ding et al., 2025; Vorontsov et al., 2026). For the same tissue area and patch dimensions, $20\times$ processing requires sixteen times as many patches as $5\times$. Cross-resolution transfer seeks to use this detailed supervision during training while retaining low-magnification inputs at deployment, as illustrated in Figure 1a. Choosing what to transfer is part of the problem. Adapting teacher guidance to student capacity can improve distillation (Qian et al., 2025), while stronger teachers may not lead to comparable student gains (Lee et al., 2025). A regional mean averages high-magnification features within a low-magnification field, while spatial targets retain variation within that field. The low-magnification input may not determine all of this detail, reducing the prediction diversity (Figure 1b). Even an accurate prediction may provide little task benefit if it replaces useful native features or is difficult for the slide model to use.

Recent systems align representations across magnifications (Han et al., 2026), combine regional and spatial teacher guidance (Su et al., 2025), or transfer patch and slide supervision to a low-magnification student (Shin & Jeong, 2026). These studies evaluate complete transfer systems and their components. Task gains alone do not establish whether (a) the teacher target is useful, (b) low-magnification input can predict it, and (c) the slide model benefits from that prediction. In our study, we start from mathematics to careful experimental designs to answer these questions. We study these questions for high-resolution to low-resolution transfer across five classification and grading and five survival cohorts. Formal analysis and matched experiments connect observed teacher targets, student predictions and their use by the slide model. Our contributions are as follows.

Task risk decomposition. We formulate cross-resolution transfer through a decomposition of task risk. The formulation separates the value of teacher information from information lost in the student representation and the task model's ability to use what remains.

Controlled experiments across ten cohorts. We study these distinctions across ten classification, grading and survival cohorts with five realizations. Our controls let us examine teacher content, prediction and delivery separately for clean answers to each question.

Empirical findings on useful transfer. We find that observed regional teacher means can complement native features, yet better reconstruction does not consistently improve task scores. Predictions often fail to cover full teacher target diversity. Further, retaining native features changes performance. Therefore, the way predictions are used becomes an essential part of the transfer evaluation.

2 Related Work

Knowledge transfer across representations. Knowledge distillation transfers teacher predictions, features or attention (Hinton et al., 2015; Romero et al., 2015; Zagoruyko & Komodakis, 2017). Recent methods adapt guidance to student capacity (Qian et al., 2025) or preserve useful student features during teacher matching (Zhang et al., 2025). Residual error distillation trains an assistant to predict differences between teacher and student features (Gao et al., 2021). Our direct and residual predictors share the same frozen native input and regional target. Cross-resolution transfer additionally changes the information available at deployment, connecting it to learning with privileged information (Vapnik & Vashist, 2009; Lopez-Paz et al., 2016).

Cross-resolution learning. Earlier histology studies transfer high-resolution feature, output and class-specific attention guidance to lower-resolution classifiers (DiPalma et al., 2021; del Amor et al., 2023). Recent approaches align magnification-specific representations (Han et al., 2026), distill regional means and local features (Su et al., 2025), or combine aligned patch supervision with slide predictions and attention (Shin & Jeong, 2026). Their component ablations assess particular systems. Our controls evaluate observed teacher targets separately from student predictions, prediction beyond a training-derived constant, and native retention with the prediction unchanged.

Reconstruction and useful information. Closer agreement with teacher predictions need not improve generalization (Stanton et al., 2021; Tiapkin et al., 2025). Usable information and conditional probing distinguish what a representation contains from what a restricted model can exploit, including beyond a baseline (Xu et al., 2020; Hewitt et al., 2021). Work on spatial token compression within

tiles evaluates both reconstruction and slide-task performance (Li et al., 2025). These distinctions motivate our separate tests of teacher prediction and task benefit when native features are retained.

3 Problem Formulation and Analytical Rationale

Cross-resolution knowledge transfer involves choosing a teacher target, learning from low-magnification input and delivering the resulting representation to a task-specific model. We distinguish these choices through a decomposition of population risk. It separates the benefit of teacher access from information lost in the representation and the task model's ability to use what remains.

Information access and task risk. We denote task outcome as Y, complete low magnification input as B, including raw images or feature embeddings. For each region i, a summary function τ maps the aligned high-magnification features H_i to a target t_i, and $T = (t_i)_i$ collects the teacher targets. A student f_θ with parameters θ produces Q from B, either a teacher prediction or an adapted encoder output. To describe how outputs reach the task model, we use a delivery function D that combines or replaces native features to form representation Z. The task model g maps Z to its prediction $\widehat{Y}$,

$$t_i = \tau(H_i), \qquad Q = f_\theta(B), \qquad Z = D(B, Q) \text{ or } Z = D(Q), \qquad \widehat{Y} = g(Z). \tag{1}$$

The analysis concerns an independent evaluation population after training is complete. Training data, transformations and parameters remain fixed, so Z is a deterministic function of B. Let U denote the input supplied to a task model. It can be the native input B, the student representation Z or the joint input (B, T). A nonnegative loss ℓ compares a task prediction with outcome Y, and $\mathbb{E}$ averages over the evaluation population. We write the population risk of a trained model g using U as $\mathcal{R}(g; U) = \mathbb{E}[\ell(g(U), Y)]$. To distinguish the information in an input from a particular model's ability to use it, let a range over all measurable prediction rules using U with the allowed task outputs. The infimum risk $\rho(U) = \inf_a \mathbb{E}[\ell(a(U), Y)]$ is the lower limit permitted by U without restricting network architecture. We write $\rho(B, T)$ for the joint input (B, T). For a trained model g using Z, the excess risk relative to $\rho(Z)$ is $\varepsilon_g = \mathcal{R}(g; Z) - \rho(Z)$. This gap includes the effects of architecture, finite training data, optimization and model selection.

Theorem 1. *Suppose the defined risks are finite and all functions are measurable. Then*

$$\mathcal{R}(g; Z) - \rho(B, T) = \underbrace{\rho(B) - \rho(B, T)}_{\textit{teacher access}} + \underbrace{\rho(Z) - \rho(B)}_{\textit{representation loss}} + \underbrace{\varepsilon_g}_{\textit{model excess}}. \tag{2}$$

Each term is nonnegative. If B can be recovered from Z, representation loss is zero. Let Z_P and Z_I be two deterministic augmentations retaining the same complete B, with trained task models g_P and g_I and respective excess risks ε_{g_P} and ε_{g_I}. Then

$$\mathcal{R}(g_P; Z_P) - \mathcal{R}(g_I; Z_I) = \varepsilon_{g_P} - \varepsilon_{g_I}. \tag{3}$$

The teacher access measures the value of observing T in addition to B, while representation loss measures the task information discarded when B becomes Z. Model excess is the remaining gap between the trained task model and the best rule using Z. Through these terms, we distinguish an informative teacher from a useful student. The distinction matters because an observed teacher target may add information about the evaluation case, whereas a deterministic student prediction can only transform information already present in B. Such a transformation can still make learning easier for a finite model (Xu et al., 2020). If both representations retain the complete native input, their representation loss is zero. Using Equation 3, we can then attribute their difference in population loss to model excess. Removing the native block does not by itself imply that relevant information has been lost, since the student output may preserve it through a replacement or residual sum. The decomposition applies to classification log-loss and discrete survival NLL with common observation weights. Reconstruction error and aggregate task metrics describe different quantities and are evaluated separately. Likewise, a model receiving only T does not implement the joint reference (B, T). We use these distinctions to decide which empirical comparisons can inform each term, with further details in Appendix A.1.

Mean supervision for a shared regional output. Individual teacher features contain variation that their regional mean removes. Whether this variation gives the student a distinct training signal depends on the learning objective. For our prediction setting in Equation 1, let $q_i = f_\theta(B)_i$ be one output shared across all teacher vectors in region i. With this shared output, we compare squared error against the individual teacher features with squared error against their mean.

Lemma 1. *For region i, let $H_i = (h_{ij})_{j=1}^{K_i}$ contain $K_i \geq 1$ fixed teacher vectors in $\mathbb{R}^d$; d being the feature dimension. Define their mean $m_i = K_i^{-1} \sum_j h_{ij}$. Every shared prediction $q_i \in \mathbb{R}^d$ satisfies*

$$\frac{1}{K_i} \sum_{j=1}^{K_i} \|q_i - h_{ij}\|_2^2 = \|q_i - m_i\|_2^2 + \frac{1}{K_i} \sum_{j=1}^{K_i} \|h_{ij} - m_i\|_2^2. \tag{4}$$

The final term is independent of q_i and contributes no predictor gradient.

In our study, $K_i = 16$, so uniform squared error against all sixteen features gives the same gradient as supervision from their mean. Sampling one feature uniformly preserves the expected gradient, although its variability and the resulting training trajectory may change. An objective can provide a different expected training signal by using separate outputs or inputs that depend on the teacher target, which depart from the lemma's assumptions. We therefore interpret the equivalence as a statement about supervision, leaving the task information contained in the teacher collection as a separate question. Appendix A.2 gives the proof and specifies its gradient conditions.

Direct and residual prediction under squared-error. Once a target t_i in Equation 1 is chosen, the student can predict it directly or learn a correction to the native feature. Let x_i denote the native feature and q_i the student prediction, with $x_i, q_i, t_i \in \mathbb{R}^d$ and B containing x_i. To compare these two prediction forms, we use the conditional mean $\mu_i(B) = \mathbb{E}[t_i \mid B]$, which averages possible teacher targets given the available input. This expectation differs from the regional mean m_i in Lemma 1, which averages the teacher vectors observed in one region.

Proposition 1. *Suppose t_i, x_i and q_i have finite second moments, with x_i and q_i measurable from B. Then*

$$\mathbb{E}\|t_i - q_i\|_2^2 = \mathbb{E}\|t_i - \mu_i(B)\|_2^2 + \mathbb{E}\|\mu_i(B) - q_i\|_2^2. \tag{5}$$

Over all square integrable functions of B, direct prediction is minimized uniquely almost surely by $q_i = \mu_i(B)$. A residual function r_i defines prediction $q_i = x_i + r_i(B)$ and has optimum $r_i(B) = \mu_i(B) - x_i$, giving the same output.

The first term in Equation 5 is target variation that the specified input B cannot determine, while the second measures how far the predictor remains from the conditional mean. Although direct and residual learning use the same information and share the same unrestricted optimum, differences in capacity, initialization, and optimization can still lead to different reconstruction errors. The regional mean averages teacher features across patch locations and remains unchanged when those locations are rearranged. Spatial variation describes how the features differ across positions within the region, capturing additional structure. A teacher summary can keep some of this structure through additional spatial components. This variation across patch locations is distinct from uncertainty about a target given B. We apply the decomposition separately to the regional mean and each spatial component. Appendix A.3 gives the proof.

4 Study Design

Theorem 1 separates the information provided by a teacher, the representation learned by a student, and the application of that representation by a downstream task model. We design the study to identify the points where these stages no longer reinforce each other. First, we use teacher targets to define which outputs are useful for prediction. Next, we conduct reconstruction tests to evaluate whether low-magnification inputs can predict these targets more effectively than taking an average. Finally, we perform downstream comparisons to isolate the contribution of feature prediction itself and the presence of native features.

We develop experiments in two stages as shown in Figure 2 using CAMELYON16/17 (Ehteshami Bejnordi et al., 2017; Bándi et al., 2019), BRACS(Brancati et al., 2022), PANDA(Bulten et al., 2022), and the private colorectal polyp cohort Private-CRC for classification, and TCGA-(KIRC, KIRP, LUAD, STAD, and UCEC) (National Cancer Institute, n.d.) for survival analysis. We use Atlas-Patch (Alagha et al., 2026) to extract non-overlapping patches of 256×256 pixels at $5\times$ and $20\times$ magnification. We choose UNI-v2 (Chen et al., 2024) as the patch encoder for most experiments. For summary prediction, we encode the same $5\times$ patches with frozen DINOv3 (Siméoni et al., 2025). Figure 3(a) summarizes the tissue and task coverage, while (b) shows how effecient the number of input token is saved using features at low-magnification. See Appendix B for details.

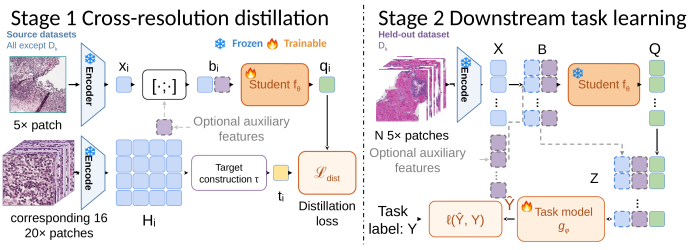


Figure 2: **Cross-resolution feature prediction and evaluation pipeline. Stage 1** learns to predict 20× teacher summaries from 5× features on datasets without task labels. **Stage 2** evaluates the frozen predictor on an unseen dataset, using the predicted features with optionally concatenated native 5× features for supervised slide-level learning.

In Stage 1, we train a predictor to estimate teacher representations from 20× patches using features from the corresponding 5× regions, keeping patch encoders frozen. We use a two-layer MLP with GELU activations, which is trained on paired features excluding the dataset used for the downstream task. In Stage 2, we freeze the predictor and train the downstream model using an unseen dataset to evaluate the performance. TransMIL (Shao et al., 2021) is used as the primary downstream model. We report the AUC for CAM16, macro F_1 for CAM17, private-CRC and BRACS, quadratically weighted Cohen's κ for PANDA, and the C-index for survival tasks. All experiments use the same five random seeds for dataset split, model initialization, and training. See Appendix B for details.

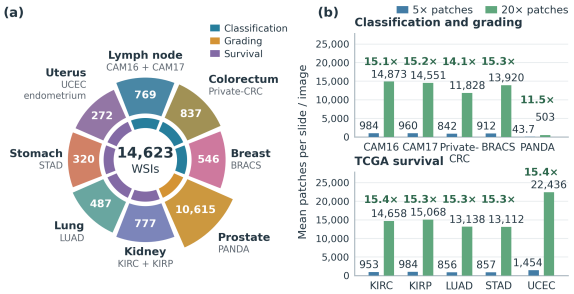


Figure 3: **Data coverage and processing scale. (a)** We use 14,623 slides across ten datasets covering classification, grading, and survival prediction. **(b)** We compare the average number of tokens at 5× and 20×.

Evaluating teacher content separately from token count. In Equation 1, τ determines which teacher information is available for transfer. For each 5× region, we average its sixteen aligned 20× feature vectors, following XMAG (Su et al., 2025). We compare the resulting regional means M with native 5× features X, so both inputs contain one token per region. We also construct M^{16} by repeating each M sixteen times to compare M^{16} with M to test the effect of repetition without adding feature information. In addition, we compare individual 20× features H with M^{16} to evaluate the value of the variation within the region at the same token count. To test whether a compact summary can retain useful spatial variation beyond the mean, we adapt Fourier Compressor (Wang et al., 2026). We apply a discrete cosine transform to the 4×4 grid of independently encoded 20× patch features and retain four low-frequency components. The constant component is rescaled to match the regional mean, while the other three (spatial components) demonstrate how features vary across the region. (Appendix C.2)

The teacher access term in Theorem 1 examines whether teacher summaries improve prediction when added to native 5× features. To investigate this, we concatenate the native UNI2-h (Chen et al., 2024) and DINOv3(Siméoni et al., 2025) features and then append either a teacher mean, a mean with spatial components, or zeros. We use zero padding to keep a fixed input width. This comparison evaluates whether the joint native and teacher inputs could benefit downstream tasks. (Appendix C.5)

Predicting regional means and spatial variation. For f_θ in Equation 1, we compare two ways to predict the regional teacher mean m_i from native features. Let x_i be the native UNI2-h vector. Direct prediction produces $q_i = f_\theta(B)_i$, while residual prediction learns $m_i - x_i$ through a network r_ϕ with parameters ϕ, giving $q_i = x_i + r_\phi(B)_i$. Both approaches use the same MLP architecture with a GELU activation (Hendrycks & Gimpel, 2016), taking x_i and its slide mean as input. The final layer maps the hidden activations through trainable weights and a bias, producing either the direct estimate or the residual correction. Proposition 1 establishes a common unrestricted optimum under squared error. To examine how finite networks approach this optimum, we compare both prediction forms under zero and random initialization of the final layer. Zero initialization sets its weights and bias to zero, so direct prediction initially returns $q_i = 0$ and residual prediction returns $q_i = x_i$. Instead, random initialization starts with an untrained direct estimate or an untrained correction to x_i.

To examine whether prediction enhances both regional content and spatial variation, we train a separate MLP with GELU activation to jointly predict regional mean and spatial components. We compare this model to a constant predictor that returns the average training target for each component, regardless of input. This comparison tests whether incorporating a region's low-magnification features improves prediction beyond the training average. By reporting mean and spatial errors separately, we identify which components contribute to reconstruction progress. Our evaluation includes datasets excluded from predictor training. See Appendix D.4 for details.

Separating native feature preservation from distillation. We examine whether native features remain useful alongside learned teacher predictions. By keeping each prediction q_i, we train task models on either $[x_i; q_i]$ or $[0; q_i]$. In this scenario, only direct access to x_i changes, and the predictions and their reconstruction erros remain identical. We repeat this comparison for direct and residual predictors under both final-layer initializations. Retaining the complete native bag guarantees zero representation loss in Theorem 1. Because the residual output already includes x_i in its sum, omitting a separate native block does not remove all of that information. Appendix E.5 provides details.

To examine model excess, we distinguish the benefit of distillation from the benefit of appending an untrained network's output. For summary augmentation, we retain the native bag $B = (b_i)_i$, where $b_i = [x_i; u_i]$ concatenates the native UNI2-h feature x_i and DINOv3 feature u_i. Let t_i denote the teacher summary, q_i the trained predictor's output, and q_i^I its output at the exact initialization used before distillation. We compare four inputs of equal width:

$$A_0 = ([b_i; 0])_i, \quad A_P = ([b_i; q_i])_i, \quad A_M = ([b_i; t_i])_i, \quad A_I = ([b_i; q_i^I])_i. \tag{6}$$

Each input trains a separate task model. Teacher summaries evaluate task performance with joint access to (B, T), without assuming that its task model attains the optimal risk $\rho(B, T)$. To assess learned spatial outputs, we keep the trained mean constant and append trained, initial or zero spatial components. See Appendix C.6.

Changing teacher supervision during encoder adaptation. Lemma 1 shows why averaging squared errors against individual features does not change the gradient of one shared output. To examine supervision that can distinguish these targets, we use conditional flow matching (Lipman et al., 2023). A network conditioned on the $5\times$ encoder feature learns the velocity along a path from noise to a teacher target. The target determines its path input, violating the lemma's assumption of a shared output. Paths toward regional means or sampled individual features allow us to assess how teacher variation shapes the encoder. Both approaches use a common mean KD objective and flow weight .1, underusing low-rank adaptation (LoRA) with rank eight for UNI2-h (Hu et al., 2022). Each of the three distillation partitions uses six cohorts and excludes its evaluation cohort. In Stage 2, we discard both training heads and use only adapted encoder features. With raw images as B in Equation 1, the encoder can alter both representation loss and model excess in Theorem 1. See Appendix E.6.

5 Results

Teacher information can enhance a task model, but accurate reconstruction does not guarantee that student predictions offer the same benefit. We present results in response to three questions. Section 5.1 identifies which teacher targets are useful, Section 5.2 compares how low magnification predicts them, and Section 5.3 examines whether the task model benefits from those predictions.

Table 1: **Downstream performance with teacher summaries added to native $5\times$ features.** Inputs retain b_i, the projected mean $\widetilde{m}_i$, and spatial component s_i. Teacher inputs use $20\times$ features at evaluation.

Dataset / metric	Native $5\times$ $[b_i; 0_{768}]$	Native $5\times$ + $20\times$ mean $[b_i; \widetilde{m}_i; o_{576}]$	Native $5\times$ + $20\times$ mean, spatial $[b_i; \widetilde{m}_i; s_i]$
CAM16 / AUC	0.8455	0.8621	**0.8932**
CAM17 / F_1	0.5197	**0.5594**	0.5198
Private-CRC / F_1	0.8305	**0.8410**	0.8305
PANDA / κ	0.8386	0.8446	**0.8456**
BRACS / F_1	0.4211	**0.4257**	0.4094
KIRC / C-index	0.6823	0.6892	**0.6907**
KIRP / C-index	0.7352	**0.7782**	0.7563
LUAD / C-index	0.5381	**0.5795**	0.5495
STAD / C-index	0.5104	**0.5112**	0.4947
UCEC / C-index	0.6499	**0.6793**	0.6672

5.1 Is a High Magnification Target Useful?

Regional teacher mean supplements the student representation. Adding the regional $20\times$ mean to native $5\times$ features increases the mean primary task score as

shown in Table 1. These results support the teacher mean as a useful addition to the low-magnification representation under this training procedure. The joint input (B, T) follows the teacher access setting in Theorem 1. The task model retains B and receives an observation from the teacher, so it does not remove useful native information. A reduction in task loss can reflect both the additional teacher information and the model's ability to use the new representation. The empirical gain confirms the regional mean could be a valuable reference target for transfer.

Adding spatial components does not consistently improve on the teacher mean alone. Mean primary scores decrease in seven of the ten datasets. While additional observations cannot worsen the unrestricted optimum in Theorem 1, the trained model's excess risk may still increase. This does not imply that spatial information is intrinsically harmful. An unrestricted predictor could ignore the extra components, whereas a trained model can generalize worse when they are supplied. Appendix C.5 provides further details.

Repeating tokens changes performance without adding information. Table 2 shows that individual teacher features generally improve classification over native input.

However, regional means, which are the most commonly used teacher targets in prior cross-scale WSI knowledge transfer works (Su et al., 2025; Yang et al., 2026; Bontempo et al., 2023), provide less consistent gains if not paired with native $5\times$ features as in Table 1. Repeating each mean sixteen times also changes task performance, although it adds no feature information.

Although repeated and single means allow the same unrestricted predictions under the information argument in Theorem 1, it provides different sequence lengths and positions to the downstream model. At the same token count, individual features H do not consistently outperform repeated means M^{16}. Thus, neither a longer sequence nor greater within-region variation guarantees a better task representation. Frequency analysis motivates compact summaries of spatial variation, but their task effects remain mixed in Appendix C. Ablation on more MIL architectures is presented in Appendix C.8.

Table 2: **Teacher content and token count have distinct task effects.** TransMIL receives native $5\times$ features, a regional $20\times$ mean, sixteen copies of that mean, or the sixteen individual features from the same tissue regions. Teacher inputs require $20\times$ images at evaluation.

	$5\times$ input	$20\times$ teacher construction		
Dataset / metric	Native $5\times$ X	Regional mean M	Repeated mean M^{16}	Individual $20\times$ H
Tokens per region	1	1	16	16
CAM16 / AUC	0.8430	0.8993	0.8872	**0.9047**
CAM17 / F_1	0.5054	0.4856	0.4904	**0.5418**
Private-CRC / F_1	0.8273	0.8146	0.8290	**0.8456**
PANDA / κ	0.8535	0.8660	0.8706	**0.8726**
BRACS / F_1	0.3747	0.3731	**0.3849**	0.3767
KIRC / C-index	0.7127	0.7105	0.6951	**0.7144**
KIRP / C-index	0.7793	**0.7802**	0.7735	0.7745
LUAD / C-index	0.5494	**0.5640**	0.5523	0.5444
STAD / C-index	0.5197	0.5402	**0.5640**	0.5533
UCEC / C-index	0.6383	0.6158	**0.6602**	0.6589

Next, we evaluate how well the student predicts teacher targets from $5\times$ inputs. We compare direct and residual prediction, and examine prediction of the mean and spatial components separately.

5.2 Can Low-Mag Predict the Target?

Direct prediction reduces the gap to the target. Direct prediction has lower MSE than residual prediction on every dataset in Table 3. The same ordering also holds under random final-layer initialization in Appendix D.1. Although the two forms share the same optimum over unrestricted functions (Proposition 1), their finite networks differ in parameterization and optimization. A residual connection does not guarantee better reconstruction of the teacher target. A lower error than the native feature is not sufficient evidence of success on the downstream task. The training target mean remains more accurate on Private-CRC, PANDA and UCEC, with nearly equal average error on LUAD. By contrast, all four predictors improve on the constant on separate evaluation groups from the nine datasets used for Stage 1 training.

Table 3: **Direct and residual learning of the regional teacher mean.** Each column estimates the target m_i, and native uses the unchanged $5\times$ feature x_i. The constant $\bar{m}_{\text{train}}$ is the target mean estimated only from distillation training groups. Direct and residual predictions use the networks defined in Section 4, with zero initialization of their final layers.

Dataset	Native $5\times$ x_i	Constant $\bar{m}_{\text{train}}$	Direct $f_\theta(B)_i$	Residual $x_i + r_\phi(B)_i$
CAM16	0.1759	0.0875	**0.0667**	0.0935
CAM17	0.1939	0.0912	**0.0776**	0.1064
Private-CRC	0.1876	**0.1051**	0.1331	0.1675
PANDA	0.1554	**0.0837**	0.0988	0.1231
BRACS	0.1882	0.1099	**0.0936**	0.1102
KIRC	0.1697	0.1218	**0.1084**	0.1182
KIRP	0.1549	0.1099	**0.0968**	0.1040
LUAD	0.2074	0.0954	**0.0953**	0.1333
STAD	0.1828	0.0927	**0.0866**	0.1080
UCEC	0.1857	**0.1118**	0.1183	0.1261

Figure 4 shows different prediction patterns across four test slides. We can see that student outputs can overlap high-magnification feature groups without reproducing their corresponding regional teacher targets, while predictions do not cover the diversity of the teacher targets, showing that some information is not predictable directly from the student. The teacher means overlap the individual 20× reference distribution, while student predictions cover these target groups unevenly. Both prediction forms overlap teacher features while retaining errors for their matched regional targets.

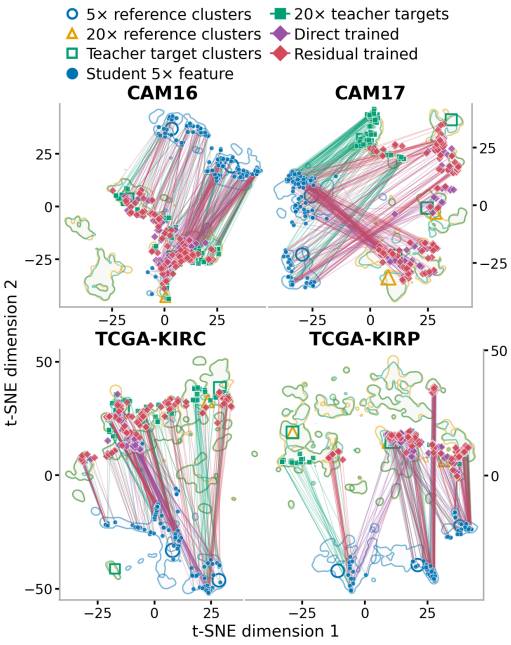


Figure 4: **Regional correspondence across four datasets.** Both prediction forms are shown for one test slide per dataset. Teacher targets are regional means. Lines connect native inputs to their targets and trained outputs. Each dataset has independent t-SNE coordinates. Appendix D.3 gives sampling details, original-coordinate measurements and the remaining six datasets.

In the KIRC and KIRP slides, residual outputs extend into teacher neighbourhoods covered less extensively by direct predictions. Nevertheless, direct prediction has lower reconstruction error on all four slides in the original coordinates (Table 18). Visual coverage and accuracy against corresponding regional targets describe different aspects. The quantitative analysis in Appendix D.2 measures variation in the original coordinates across five realizations. Direct outputs have lower centered variance than teacher means, while residual outputs have higher variance in every dataset. For both forms, distances between predicted features remain positively correlated with distances between their teacher targets. These measurements distinguish the spread of predictions from their accuracy against matched targets. Proposition 1 explains why squared error prediction need not reproduce the full variation of observed targets. Its optimum averages possible targets given the available input, rather than reproducing every possible outcome. The two prediction forms share this unrestricted optimum, but their finite architectures differ.

Reconstruction progress concentrates in the regional mean. A separate predictor estimates regional mean and spatial components. Figure 5 compares its MSE with a constant predictor that returns the average training target for each component. The network usually improves on the constant for the regional mean, but the spatial components remain less accurate than their constant baselines. This pattern also occurs on independent evaluation groups from the nine distillation datasets, not only on the excluded dataset. Thus, improvements in overall reconstruction error may primarily reflect better estimation of the regional mean, while spatial prediction remains limited.

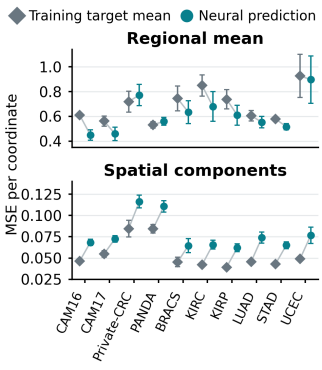


Figure 5: **Reconstruction gains in the regional mean.** The network is compared with the training target, which returns a constant for inputs.

Proposition 1 explains why the two components need separate interpretation. Error may persist as the input does not determine the target, or the network has not learned its conditional mean. The spatial results reveal a distinct learning challenge that should be measured directly rather than inferred from total error. Appendices D.6 and D.7 examine correction and spatial inputs, while the downstream spatial study uses separate heads retaining all feature channels.

5.3 Does Downstream Model Benefit from Prediction?

Finally, we evaluate whether a useful prediction leads to a useful task representation. We first keep the prediction unchanged and vary native retention, then separate the effects of distillation from those of adding a network transformation. Encoder adaptation further compares how the teacher target influences the deployed representation.

Native features can help when teacher predictions are unchanged. Figure 6 shows different task outcomes for the same teacher prediction. Providing native features separately gives more consistent gains in primary scores for direct prediction, while NLL often improves for both direct and residual prediction. Reconstruction accuracy remains identical within comparisons. The task difference stems from how the model receives the prediction and native information, not from a better teacher estimate.

The retained input $[x_i; q_i]$ (Section 4) allows recovery of the complete native bag, resulting in zero representation loss under Theorem 1. Removing the separate native block does not establish that all native information is lost, since some may remain in q_i, including through the residual connection. The task difference can therefore reflect both information preservation and how effectively the trained model uses its input. It identifies native feature retention as a consequential design choice without attributing its effect entirely to representation loss.

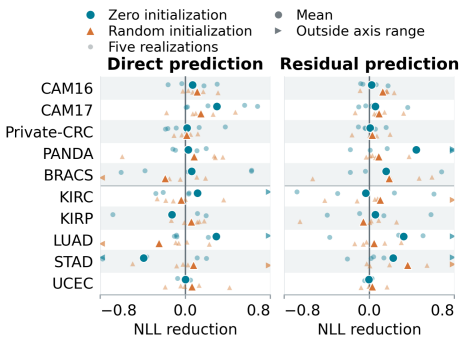


Figure 6: **Retaining native features can reduce task loss without changing teacher predictions.** NLL for $[0; q_i]$ minus NLL for $[x_i; q_i]$ is positive when retention helps. Small marks show five corresponding realization differences and large marks their means. Colours distinguish final-layer initialization before distillation.

The auxiliary teacher target changes downstream encoder performance. Table 4 shows that using individual teacher features or their regional mean as targets leads to distinct downstream results, even when the encoder adaptation procedure remains unchanged.

Lemma 1 demonstrates that the gradients of these targets are equivalent only when a squared error loss supervises a single shared output. Under the flow objective, the choice of teacher target also determines the network's path input, allowing it to change the training signal. Both models keep the same mean KD objective and deploy only encoder features, as Section 4 specifies.

This experiment uses raw images as B in Equation 1. The adapted encoder need not preserve the complete image, so both representation loss and model excess can change. Unlike native augmentation, these results cannot attribute the task effect solely to model excess. They show that the supervision objective determines whether individual teacher variation supplies a distinct learning signal, and that its task value must be evaluated for the resulting encoder. Supporting comparisons in Appendices E.3 and E.4 examine prediction delivery without adapting the encoder.

Table 4: **The auxiliary teacher target changes encoder performance.** Both flow models retain mean KD and supply only encoder features to TransMIL. Table 26 includes mean KD without flow.

		Mean KD + flow toward	
Dataset / metric	Native 5× Frozen UNI2-h	Regional 20× mean	Individual 20× feature
CAM16 / AUC	0.8808	**0.8827**	0.8749
CAM17 / F_1	0.5025	0.5077	**0.5211**
Private-CRC / F_1	0.8411	0.8225	**0.8492**
PANDA / κ	**0.8499**	0.8496	0.8495
BRACS / F_1	**0.4589**	0.4204	0.4347
KIRC / C-index	0.7321	**0.7378**	0.7069
KIRP / C-index	**0.7921**	0.6861	0.7744
LUAD / C-index	0.5307	0.5537	**0.5783**
STAD / C-index	0.5532	0.5912	**0.5953**
UCEC / C-index	**0.6561**	0.5924	0.5976

6 Discussion and Conclusion

Across ten cohorts, observed regional teacher means improve average primary scores when added to native features, but student predictions do not inherit that value automatically. In separate reconstruction tests, direct prediction has lower average error than residual prediction while its outputs vary less than teacher targets. Better reconstruction does not consistently improve task scores. Native retention changes performance at fixed predictions. Keeping the native input removes representation loss, but finite models can still differ in how they use it. Encoder adaptation also reveals no consistently preferred auxiliary teacher target. For future methods, we recommend separate tests of teacher usefulness, prediction beyond a constant estimated from training targets and downstream benefit.

The study covers one magnification pair and selected model families. Target coordinates, sampled regions and input widths differ across experiments, limiting comparisons between them. Incomplete patient linkage and sparse survival events further constrain inference. Broader studies could vary these design choices within a common protocol and examine which cohorts and cases benefit.

AI USE STATEMENT

The authors used GPT-6 Astra to assist in manuscript polishing, grammar checks, code implementations, and math proof reviews. The authors are responsible for the experiments, citations, mathematical arguments and interpretations.

REPRODUCIBILITY STATEMENT

Sections 3 and 4 define the models, training stages and comparisons. The appendices specify the protocols, seed roles, evaluation populations and uncertainty estimates. The supplementary materials includes our code, numerical exports and scripts for reproduction. Appendix B.7 states what is available and what is still needed to reproduce all model fits and intervals.

REFERENCES

Ahmed Alagha, Christopher Leclerc, Yousef Kotp, Omar Metwally, Calvin Moras, Peter Rentopoulos, Ghodsiyeh Rostami, Bich Ngoc Nguyen, Jumanah Baig, Abdelhakim Khellaf, Vincent Quoc-Huy Trinh, Rabeb Mizouni, Hadi Otrok, Jamal Bentahar, and Mahdi S. Hosseini. AtlasPatch: Scalable foundation model-based tissue detection and patch extraction for computational pathology. *arXiv preprint arXiv:2602.03998*, 2026. URL https://arxiv.org/abs/2602.03998.

Gianpaolo Bontempo, Angelo Porrello, Federico Bolelli, Simone Calderara, and Elisa Ficarra. Dasmil: Distilling across scales for mil classification of histological wsis. In Hayit Greenspan, Anant Madabhushi, Parvin Mousavi, Septimiu Salcudean, James Duncan, Tanveer Syeda-Mahmood, and Russell Taylor (eds.), *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*, pp. 248–258, Cham, 2023. Springer Nature Switzerland. ISBN 978-3-031-43907-0.

Nadia Brancati, Anna Maria Anniciello, Pushpak Pati, Daniel Riccio, Giosuè Scognamiglio, Guillaume Jaume, Giuseppe De Pietro, Maurizio Di Bonito, Antonio Foncubierta, Gerardo Botti, Maria Gabrani, Florinda Feroce, and Maria Frucci. BRACS: A dataset for BReAst carcinoma subtyping in H&E histology images. *Database*, 2022:baac093, October 2022. ISSN 1758-0463. doi: 10.1093/database/baac093. URL https://doi.org/10.1093/database/baac093.

Wouter Bulten, Kimmo Kartasalo, Po-Hsuan Cameron Chen, Peter Ström, Hans Pinckaers, Kunal Nagpal, Yuannan Cai, David F. Steiner, Hester van Boven, Robert Vink, Christina Hulsbergen-van de Kaa, Jeroen van der Laak, Mahul B. Amin, Andrew J. Evans, Theodorus van der Kwast, Robert Allan, Peter A. Humphrey, Henrik Grönberg, Hemamali Samaratunga, Brett Delahunt, Toyonori Tsuzuki, Tomi Häkkinen, Lars Egevad, Maggie Demkin, Sohier Dane, Fraser Tan, Masi Valkonen, Greg S. Corrado, Lily Peng, Craig H. Mermel, Pekka Ruusuvuori, Geert Litjens, Martin Eklund, the PANDA challenge consortium, Américo Brilhante, Aslı Çakır, Xavier Farré, Katerina Geronatsiou, Vincent Molinié, Guilherme Pereira, Paromita Roy, Günter Saile, Paulo G. O. Salles, Ewout Schaafsma, Joëlle Tschui, Jorge Billoch-Lima, Emíio M. Pereira, Ming Zhou, Shujun He, Sejun Song, Qing Sun, Hiroshi Yoshihara, Taiki Yamaguchi, Kosaku Ono, Tao Shen, Jianyi Ji, Arnaud Roussel, Kairong Zhou, Tianrui Chai, Nina Weng, Dmitry Grechka, Maxim V. Shugaev, Raphael Kiminya, Vassili Kovalev, Dmitry Voynov, Valery Malyshev, Elizabeth Lapo, Manuel Campos, Noriaki Ota, Shinsuke Yamaoka, Yusuke Fujimoto, Kentaro Yoshioka, Joni Juvonen, Mikko Tukiainen, Antti Karlsson, Rui Guo, Chia-Lun Hsieh, Igor Zubarev, Habib S. T. Bukhar, Wenyuan Li, Jiayun Li, William Speier, Corey Arnold, Kyungdoc Kim, Byeonguk Bae, Yeong Won Kim, Hong-Seok Lee, and Jeonghyuk Park. Artificial intelligence for diagnosis and Gleason grading of prostate cancer: the PANDA challenge. *Nature Medicine*, 28(1):154–163, Jan 2022. ISSN 1546-170X. doi: 10.1038/s41591-021-01620-2. URL https://doi.org/10.1038/s41591-021-01620-2.

Péter Bándi, Oscar Geessink, Quirine Manson, Marcory Van Dijk, Maschenka Balkenhol, Meyke Hermsen, Babak Ehteshami Bejnordi, Byungjae Lee, Kyunghyun Paeng, Aoxiao Zhong, Quanzheng Li, Farhad Ghazvinian Zanjani, Svitlana Zinger, Keisuke Fukuta, Daisuke Komura, Vlado Ovtcharov, Shenghua Cheng, Shaoqun Zeng, Jeppe Thagaard, Anders B. Dahl, Huangjing Lin, Hao Chen, Ludwig Jacobsson, Martin Hedlund, Melih Çetin, Eren Halıcı, Hunter Jackson, Richard Chen, Fabian Both, Jörg Franke, Heidi Küsters-Vandevelde, Willem Vreuls, Peter Bult,

Bram van Ginneken, Jeroen van der Laak, and Geert Litjens. From detection of individual metastases to classification of lymph node status at the patient level: The CAMELYON17 challenge. *IEEE Transactions on Medical Imaging*, 38(2):550–560, 2019. doi: 10.1109/TMI.2018.2867350. URL https://doi.org/10.1109/TMI.2018.2867350.

Richard J. Chen, Tong Ding, Ming Y. Lu, Drew F. K. Williamson, Guillaume Jaume, Andrew H. Song, Bowen Chen, Andrew Zhang, Daniel Shao, Muhammad Shaban, Mane Williams, Lukas Oldenburg, Luca L. Weishaupt, Judy J. Wang, Anurag Vaidya, Long Phi Le, Georg Gerber, Sharifa Sahai, Walt Williams, and Faisal Mahmood. Towards a general-purpose foundation model for computational pathology. *Nature Medicine*, 30(3):850–862, 2024. doi: 10.1038/s41591-024-02857-3. URL https://doi.org/10.1038/s41591-024-02857-3.

Rocío del Amor, Julio Silva-Rodríguez, Adrián Colomer, and Valery Naranjo. Attention to detail: Inter-resolution knowledge distillation. In *2023 31st European Signal Processing Conference (EUSIPCO)*, pp. 985–989, 2023. doi: 10.23919/EUSIPCO58844.2023.10289941. URL https://doi.org/10.23919/EUSIPCO58844.2023.10289941.

Tong Ding, Sophia J. Wagner, Andrew H. Song, Richard J. Chen, Ming Y. Lu, Andrew Zhang, Anurag J. Vaidya, Guillaume Jaume, Muhammad Shaban, Ahrong Kim, et al. A multimodal whole-slide foundation model for pathology. *Nature Medicine*, 31:3749–3761, 2025. doi: 10.1038/s41591-025-03982-3. URL https://doi.org/10.1038/s41591-025-03982-3.

Joseph DiPalma, Arief A. Suriawinata, Laura J. Tafe, Lorenzo Torresani, and Saeed Hassanpour. Resolution-based distillation for efficient histology image classification. *Artificial Intelligence in Medicine*, 119:102136, 2021. doi: 10.1016/j.artmed.2021.102136. URL https://doi.org/10.1016/j.artmed.2021.102136.

Babak Ehteshami Bejnordi, Mitko Veta, Paul J. van Diest, Bram van Ginneken, Nico Karssemeijer, Geert Litjens, Jeroen A. W. M. van der Laak, and the CAMELYON16 Consortium. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *JAMA*, 318(22):2199–2210, 12 2017. ISSN 0098-7484. doi: 10.1001/jama.2017.14585. URL https://doi.org/10.1001/jama.2017.14585.

Mengya Gao, Yujun Wang, and Liang Wan. Residual error based knowledge distillation. *Neurocomputing*, 433:154–161, 2021. doi: 10.1016/j.neucom.2020.10.113. URL https://doi.org/10.1016/j.neucom.2020.10.113.

Chu Han, Bingchao Zhao, Tianpeng Deng, Jingqi Huang, Fangfang Liu, Jiatai Lin, Shanshan Lyu, Longfei Wang, Cheng Lu, Changhong Liang, Hannah Y. Wen, Zhenwei Shi, Xiaojing Guo, and Zaiyi Liu. A magnification alignment framework enables computation- and communication-efficient computational pathology. *Cancer Research*, 86(12):3060–3073, 2026. doi: 10.1158/0008-5472.CAN-25-3255. URL https://doi.org/10.1158/0008-5472.CAN-25-3255.

Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (GELUs). *arXiv preprint arXiv:1606.08415*, 2016. URL https://arxiv.org/abs/1606.08415.

John Hewitt, Kawin Ethayarajh, Percy Liang, and Christopher Manning. Conditional probing: measuring usable information beyond a baseline. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 1626–1639, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.122. URL https://aclanthology.org/2021.emnlp-main.122/.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. URL https://arxiv.org/abs/1503.02531. Presented at the NIPS 2014 Deep Learning Workshop.

Mahdi S. Hosseini, Babak Ehteshami Bejnordi, Vincent Quoc-Huy Trinh, Lyndon Chan, Danial Hasan, Xingwen Li, Stephen Yang, Taehyo Kim, Haochen Zhang, Theodore Wu, Kajanan Chinniah, Sina Maghsoudlou, Ryan Zhang, Jiadai Zhu, Samir Khaki, Andrei Buin, Fatemeh Chaji, Ala Salehi, Bich Ngoc Nguyen, Dimitris Samaras, and Konstantinos N. Plataniotis. Computational pathology:

A survey review and the way forward. *Journal of Pathology Informatics*, 15:100357, 2024. ISSN 2153-3539. doi: 10.1016/j.jpi.2023.100357. URL https://www.sciencedirect.com/science/article/pii/S2153353923001712.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Maximilian Ilse, Jakub Tomczak, and Max Welling. Attention-based deep multiple instance learning. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2127–2136, 2018. URL https://proceedings.mlr.press/v80/ilse18a.html.

Nikolaus Kriegeskorte, Marieke Mur, and Peter Bandettini. Representational similarity analysis—connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, 2:4, 2008. doi: 10.3389/neuro.06.004.2008.

Jungsoo Lee, Debasmit Das, Munawar Hayat, Sungha Choi, Kyuwoong Hwang, and Fatih Porikli. CustomKD: Customizing large vision foundation for edge model improvement via knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 25176–25186, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Lee_CustomKD_Customizing_Large_Vision_Foundation_for_Edge_Model_Improvement_via_CVPR_2025_paper.html.

Bin Li, Yin Li, and Kevin W. Eliceiri. Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14318–14328, 2021. URL https://openaccess.thecvf.com/content/CVPR2021/html/Li_Dual-Stream_Multiple_Instance_Learning_Network_for_Whole_Slide_Image_Classification_CVPR_2021_paper.html.

Honglin Li, Zhongyi Shui, Yunlong Zhang, Chenglu Zhu, and Lin F. Yang. PathVQ: Reforming computational pathology foundation model for whole slide image analysis via vector quantization. In *Advances in Neural Information Processing Systems 38*, 2025. doi: 10.52202/085713-2086. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/59f278de1619bdb6b53fd04e8e0976e0-Abstract-Conference.html.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=PqvMRDCJT9t.

David Lopez-Paz, Léon Bottou, Bernhard Schölkopf, and Vladimir Vapnik. Unifying distillation and privileged information. In *International Conference on Learning Representations*, 2016. URL https://arxiv.org/abs/1511.03643.

National Cancer Institute. The Cancer Genome Atlas Program (TCGA). https://www.cancer.gov/ccg/research/genome-sequencing/tcga, n.d. Accessed 21 March 2026.

Pavlin G. Poličar, Martin Stražar, and Blaž Zupan. openTSNE: A modular Python library for t-SNE dimensionality reduction and embedding. *Journal of Statistical Software*, 109(3):1–30, 2024. doi: 10.18637/jss.v109.i03. URL https://www.jstatsoft.org/index.php/jss/article/view/v109i03.

Chengyao Qian, Trung Le, and Mehrtash Harandi. A good teacher adapts their knowledge for distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1239–1248, 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Qian_A_Good_Teacher_Adapts_Their_Knowledge_for_Distillation_ICCV_2025_paper.html.

Zequn Qin, Pengyi Zhang, Fei Wu, and Xi Li. FcaNet: Frequency channel attention networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 783–792, 2021. URL https://openaccess.thecvf.com/content/ICCV2021/html/Qin_FcaNet_Frequency_Channel_Attention_Networks_ICCV_2021_paper.html.

Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems*, volume 20, 2007. URL https://proceedings.neurips.cc/paper/2007/hash/013a006f03dbc5392effeb8f18fda755-Abstract.html.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. SAM 2: Segment anything in images and videos. In *International Conference on Learning Representations*, volume 2025, pp. 28085–28128, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/45c1f6a8cbf2da59ebf2c802b4f742cd-Abstract-Conference.html.

Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. FitNets: Hints for thin deep nets. In *International Conference on Learning Representations*, 2015. URL https://arxiv.org/abs/1412.6550.

Zhuchen Shao, Hao Bian, Yang Chen, Yifeng Wang, Jian Zhang, Xiangyang Ji, and Yongbing Zhang. TransMIL: Transformer based correlated multiple instance learning for whole slide image classification. In *Advances in Neural Information Processing Systems*, volume 34, pp. 2136–2147, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/10c272d06794d3e5785d5e7c5356e9ff-Abstract.html.

Yonghan Shin and Won-Ki Jeong. LRMIL: Efficient Low-Resolution Multiple Instance Learning via High-Resolution Knowledge Distillation for Whole Slide Image Classification . In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2026*, volume LNCS 16891. Springer Nature Switzerland, September 2026.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. DINOv3. *arXiv preprint arXiv:2508.10104*, 2025. URL https://arxiv.org/abs/2508.10104.

Samuel Stanton, Pavel Izmailov, Polina Kirichenko, Alexander A. Alemi, and Andrew Gordon Wilson. Does knowledge distillation really work? In *Advances in Neural Information Processing Systems*, volume 34, pp. 6906–6919, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/376c6b9ff3bedbbea56751a84fffc10c-Abstract.html.

Ziyu Su, Abdul Rehman Akbar, Usama Sajjad, Anil V. Parwani, and Muhammad Khalid Khan Niazi. Streamline pathology foundation model by cross-magnification distillation. *arXiv preprint arXiv:2509.23097*, 2025. URL https://arxiv.org/abs/2509.23097.

Wenhao Tang, Fengtao Zhou, Sheng Huang, Xiang Zhu, Yi Zhang, and Bo Liu. Feature re-embedding: Towards foundation model-level performance in computational pathology. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11343–11352, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Tang_Feature_Re-Embedding_Towards_Foundation_Model-Level_Performance_in_Computational_Pathology_CVPR_2024_paper.html.

Daniil Tiapkin, Daniele Calandriello, Johan Ferret, Sarah Perrin, Nino Vieillard, Alexandre Rame, and Mathieu Blondel. On teacher hacking in language model distillation. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 59552–59569. PMLR, 2025. URL https://proceedings.mlr.press/v267/tiapkin25a.html.

Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008. URL https://jmlr.org/papers/v9/vandermaaten08a.html.

Vladimir Vapnik and Akshay Vashist. A new learning paradigm: Learning using privileged information. *Neural Networks*, 22(5–6):544–557, 2009. doi: 10.1016/j.neunet.2009.06.042. URL https://doi.org/10.1016/j.neunet.2009.06.042.

Eugene Vorontsov, George Shaikovski, Adam Casson, Julian Viret, Eric Zimmermann, et al. End-to-end multimodal pathology foundation model with clinical dialogue. *Nature Medicine*, 32(9): 3203–3213, 2026. doi: 10.1038/s41591-026-04521-4. URL https://doi.org/10.1038/s41591-026-04521-4.

Huanyu Wang, Jushi Kai, Haoli Bai, Lu Hou, Bo Jiang, Ziwei He, and Zhouhan Lin. Fourier Compressor: Frequency-domain visual token compression for vision-language models. In Paolo Favaro, Zuzana Kukelova, Atsuto Maki, Anna Rohrbach, Konrad Schindler, and Federico Tombari (eds.), *Computer Vision – ECCV 2026*, volume 17009 of *Lecture Notes in Computer Science*, pp. 284–300, Cham, 2026. Springer Nature Switzerland. ISBN 978-3-032-37035-8. doi: 10.1007/978-3-032-37035-8_16. URL https://doi.org/10.1007/978-3-032-37035-8_16.

Xiyue Wang, Junhan Zhao, Eliana Marostica, Wei Yuan, Jietian Jin, Jiayu Zhang, Ruijiang Li, Hongping Tang, Kanran Wang, Yu Li, Fang Wang, Yulong Peng, Junyou Zhu, Jing Zhang, Christopher R. Jackson, Jun Zhang, Deborah Dillon, Nancy U. Lin, Lynette Sholl, Thomas Denize, David Meredith, Keith L. Ligon, Sabina Signoretti, Shuji Ogino, Jeffrey A. Golden, MacLean P. Nasrallah, Xiao Han, Sen Yang, and Kun-Hsing Yu. A pathology foundation model for cancer diagnosis and prognosis prediction. *Nature*, 634:970–978, 2024. doi: 10.1038/s41586-024-07894-z. URL https://www.nature.com/articles/s41586-024-07894-z.

Hanwen Xu, Naoto Usuyama, Jaspreet Bagga, Sheng Zhang, Rajesh Rao, Tristan Naumann, Cliff Wong, Zelalem Gero, Javier González, Yu Gu, Yanbo Xu, Mu Wei, Wenhui Wang, Shuming Ma, Furu Wei, Jianwei Yang, Chunyuan Li, Jianfeng Gao, Jaylen Rosemon, Tucker Bower, Soohee Lee, Roshanthi Weerasinghe, Bill J. Wright, Ari Robicsek, Brian Piening, Carlo Bifulco, Sheng Wang, and Hoifung Poon. A whole-slide foundation model for digital pathology from real-world data. *Nature*, 630(8015):181–188, 2024. doi: 10.1038/s41586-024-07441-w. URL https://doi.org/10.1038/s41586-024-07441-w.

Yilun Xu, Shengjia Zhao, Jiaming Song, Russell Stewart, and Stefano Ermon. A theory of usable information under computational constraints. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=r1eBeyHFDH.

Lihe Yang, Zhen Zhao, and Hengshuang Zhao. SigLIP-HD by fine-to-coarse supervision. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=XeLrfKEOZS.

Sergey Zagoruyko and Nikos Komodakis. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=Sks9_ajex.

Weijia Zhang, Yuehao Liu, Wu Ran, and Chao Ma. Cross-architecture distillation made simple with redundancy suppression. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 23256–23266, 2025. doi: 10.1109/ICCV51701.2025.02159. URL https://openaccess.thecvf.com/content/ICCV2025/html/Zhang_Cross-Architecture_Distillation_Made_Simple_with_Redundancy_Suppression_ICCV_2025_paper.html.

APPENDIX CONTENTS

A Definitions and Proofs **17**

A.1 Information access and task risk 17

A.2 Mean supervision for a shared regional output 18

A.3 Direct and residual prediction under squared error 19

B Data and comparison design **21**

B.1 Datasets and task partitions 21

B.2 Physical fields and frozen features 22

B.3 Token order and recipient processing 23

B.4 Patch counts and representation size 23

B.5 Cross-resolution distillation and downstream fitting 23

B.6 Comparison designs 25

B.7 Variation across realizations and reproducible scope 26

C Teacher Targets and Their Task Utility **28**

C.1 The mean summary in the training diagram 28

C.2 The four spatial components 28

C.3 Spatial frequency of the teacher features 29

C.4 Separating regional content from token count 29

C.5 Teacher information alongside the same native input 30

C.6 Spatial utility when all feature channels are retained 32

C.7 An alternative way to deliver spatial summaries 32

C.8 Regional means across task models 32

D Neural Prediction of Teacher Targets **36**

D.1 Regional mean prediction on independent test fields 36

D.2 Feature distributions before and after distillation 37

D.3 Teacher correspondence in a common feature map 40

D.4 Predicting projected mean and spatial components 47

D.5 Reconstruction errors by component 47

D.6 Independent neural error correction 47

D.7 Spatial access to the low magnification image 50

E Downstream Use and Attribution to Distillation **52**

E.1 Learning the nonconstant spatial component 52

E.2 General augmentation controls 52

E.3 Removing spatial predictions while retaining the same mean 52

E.4 Direct and residual prediction across ten datasets 53

E.5 Native retention with an unchanged teacher prediction 54
E.6 Individual features or their mean as auxiliary supervision 58

A Definitions and Proofs

This appendix gives the proofs behind the three study questions. The risk decomposition distinguishes teacher information, information lost during delivery and the ability of a trained task model to use its input. The squared error identities clarify what a shared regional output learns and what prediction can recover.

A.1 Information access and task risk

Definitions and facts. Consider a fixed joint evaluation distribution of the outcome Y, the complete deployment input B and the teacher target T. The outcome may be a class label or a survival observation containing time and censoring. Evaluation takes place after the training data, transformations and model parameters have been fixed, on a population independent of the training procedure. For a randomized learning procedure, we condition on its realized training data and randomness. The resulting prediction $Q = f_\theta(B)$ then uses only the declared deployment input.

Let U denote the input available to the task model, such as the native input B, the delivered representation Z or the joint input (B, T). All representations and functions are measurable. A common action space specifies the allowed predictions, such as class probabilities or a discrete survival distribution. A prediction rule a maps U to this space. We take $\mathcal{A}_U$ to be the nonempty set of all such measurable rules.

The nonnegative loss ℓ compares a task prediction with outcome Y. The expectation $\mathbb{E}$ averages over the fixed evaluation population, with input and outcome drawn from the same example. Every risk comparison uses the same population, loss, action space and observation weighting. For a trained task rule g, the population risk is $\mathcal{R}(g; U) = \mathbb{E}[\ell(g(U), Y)]$. To describe what the input permits without restricting the model architecture, we define the unrestricted risk as the infimum over all allowed rules,

$$\rho(U) = \inf_{a \in \mathcal{A}_U} \mathbb{E}[\ell(a(U), Y)]. \tag{7}$$

The notation $\rho(B, T)$ abbreviates $\rho((B, T))$, the infimum of risk with access to both inputs. This infimum need not be attained by a rule. For the decomposition, assume that $\rho(B, T)$, $\rho(B)$, $\rho(Z)$ and $\mathcal{R}(g; Z)$ are finite, so the differences are defined. These unrestricted optima describe the information available in a representation, whereas a finite set of trained models may achieve higher risks.

Let V be another representation, obtained through a measurable map c with $V = c(U)$ almost surely. Every rule $a \in \mathcal{A}_V$ can be used with U by composing it with c. The resulting rule $a \circ c \in \mathcal{A}_U$ makes the same predictions almost surely and therefore has the same risk. Thus the rules available from U include those available from V. Taking infima, we obtain

$$\rho(U) \leq \rho(V). \tag{8}$$

This step relies on containment of the admissible rules, so an optimal rule or a conditional density is not needed. If each representation can be recovered measurably from the other almost surely, applying the inequality in both directions gives equal optimal risks.

Proof of Theorem 1. The joint input (B, T) contains B. Under either delivery form in Equation 1, Z is a measurable function of B. Applying Equation 8 to these two containments gives

$$\rho(B, T) \leq \rho(B) \leq \rho(Z). \tag{9}$$

The fitted model g is one admissible rule using Z, so the definition of the infimum also gives

$$\rho(Z) \leq \mathcal{R}(g; Z), \qquad \varepsilon_g = \mathcal{R}(g; Z) - \rho(Z) \geq 0. \tag{10}$$

Adding and subtracting the finite quantities $\rho(B)$ and $\rho(Z)$, we obtain

$$\begin{aligned} \mathcal{R}(g; Z) - \rho(B, T) = {} & \big[\rho(B) - \rho(B, T)\big] \\ & + \big[\rho(Z) - \rho(B)\big] + \big[\mathcal{R}(g; Z) - \rho(Z)\big]. \end{aligned} \tag{11}$$

The first two terms are nonnegative by Equation 9, and the final term is ε_g. This proves the decomposition and the sign of every term.

Now suppose a measurable map c recovers B from Z almost surely. The information comparison then works in both directions. Equation 8 gives $\rho(Z) \leq \rho(B)$, which together with Equation 9 yields $\rho(Z) = \rho(B)$. Representation loss is therefore zero. For two deterministic augmentations Z_P and Z_I that each preserve the same complete B, both optimal risks equal $\rho(B)$. Let g_P and g_I denote their trained task models, and assume both trained risks are finite. Their excess risks ε_{g_P} and ε_{g_I} are defined relative to the respective optima. The trained risks can therefore be written as

$$\mathcal{R}(g_P; Z_P) = \rho(B) + \varepsilon_{g_P}, \qquad \mathcal{R}(g_I; Z_I) = \rho(B) + \varepsilon_{g_I} \tag{12}$$

and, by subtracting these expressions, we obtain Equation 3. The equality holds even when the trained models have different parameters and neither attains its optimal risk. □

Interpretation of the three terms. Teacher access measures how much joint access to B and T can lower unrestricted risk. Its value depends on the chosen target T, which may summarize only part of the original high magnification images. A model receiving T alone has a different input and may also incur model excess. We therefore distinguish that model's score from the joint optimum $\rho(B, T)$.

Representation loss measures the task information lost when B is transformed into Z. Concatenating $Z = [B, f_\theta(B)]$ makes this term zero because the first block recovers B. An invertible change of coordinates of B has the same property. A residual sum or replacement may not recover the complete native input but may still retain everything relevant to the task. The absence of a demonstrated recovery map is therefore insufficient to establish positive representation loss.

Model excess measures the trained model's gap to the unrestricted risk for its representation. Architecture, finite data, estimation, optimization and model selection can all contribute. Equal information can support different performance in finite models (Xu et al., 2020), including when an initial predictor transforms B before any teacher training. With B retained, we compare the trained predictor with its exact initialization to examine what teacher training contributes beyond the initial transformation.

The decomposition applies to a loss defined for each observation. Discrete survival NLL has this form when predictions are survival distributions and Y contains time and censoring, with common time bins and observation weights. Losses that couple observations through risk sets require a joint formulation. AUC, C-index, macro F_1 and quadratic κ depend on comparisons or counts across observations and remain separate evaluation criteria. For this reason, we interpret test scores as estimates of trained model performance and do not identify them with the unrestricted terms of the decomposition.

Classification log loss as a special case. For classification, let H denote Shannon entropy and KL denote Kullback-Leibler divergence, both using natural logarithms. Consider a finite class label and allow every vector of class probabilities as a prediction. For ordinary unweighted log loss under the fixed evaluation distribution, the optimal risk is the conditional entropy $\rho(U) = H(Y \mid U)$. To verify this identity, we write $p(\cdot \mid U)$ for the true conditional class distribution. A prediction rule a returns class probabilities, with $a_Y(U)$ denoting the probability assigned to the observed class. Its conditional expected log loss is

$$\mathbb{E}[-\log a_Y(U) \mid U] = H\big(p(\cdot \mid U)\big) + \mathrm{KL}\big(p(\cdot \mid U) \,\|\, a(U)\big). \tag{13}$$

The divergence is nonnegative and vanishes at $a(U) = p(\cdot \mid U)$. Terms with zero true class probability contribute zero, using the convention $0 \log 0 = 0$. Assigning probability zero to a class with positive true probability gives infinite loss and divergence. Averaging the identity establishes the optimal risk. Let I denote conditional mutual information. Because Z is a function of B, conditioning on (B, Z) is equivalent to conditioning on B, so $H(Y \mid B, Z) = H(Y \mid B)$. The definitions of conditional mutual information then express the first two terms of Equation 2 as

$$\rho(B) - \rho(B, T) = I(Y; T \mid B), \qquad \rho(Z) - \rho(B) = I(Y; B \mid Z). \tag{14}$$

The general risk identity recovers the classification information decomposition while also accommodating survival losses without this entropy representation.

A.2 Mean supervision for a shared regional output

Definitions and facts. For vectors $u, v \in \mathbb{R}^d$, the Euclidean inner product is $u^\top v$ and $\|u\|_2^2 = u^\top u$. Deviations from a regional mean sum to zero. Together with the expansion $\|u + v\|_2^2 = \|u\|_2^2 + \|v\|_2^2 + 2u^\top v$, this lets us separate error relative to the mean from variation among the teacher vectors.

Proof of Lemma 1. Fix one region with $K \geq 1$ teacher vectors, omit its index, and set $m = K^{-1} \sum_{j=1}^{K} h_j$. The same prediction q is compared with every teacher vector. Writing $q - h_j = (q - m) + (m - h_j)$, we expand each squared norm and average to obtain

$$\frac{1}{K} \sum_{j=1}^{K} \|q - h_j\|_2^2 = \|q - m\|_2^2 + \frac{1}{K} \sum_{j=1}^{K} \|m - h_j\|_2^2 + 2(q - m)^\top \frac{1}{K} \sum_{j=1}^{K} (m - h_j). \tag{15}$$

The cross term vanishes because $K^{-1} \sum_j (m - h_j) = m - m = 0$, leaving Equation 4. The variation among teacher vectors contributes a term independent of q. Only $\|q - m\|_2^2$ depends on the prediction, so both objectives give the output gradient $2(q - m)$. For a differentiable predictor q_θ with teacher features fixed with respect to θ, the chain rule gives the same parameter gradient $2J_\theta^\top (q_\theta - m)$, where $J_\theta = \partial q_\theta / \partial \theta$ is its output Jacobian. □

Scope of the identity. The comparison uses fixed teacher features, uniform weights within each region, common coordinates and the same fixed region weights in both objectives. For a shared output q that does not depend on the sampled index J, uniform sampling gives $\mathbb{E}_J[2(q - h_J)] = 2(q - m)$. The expected parameter gradient also agrees when the output Jacobian is shared. Sampling can still change gradient variance and the finite training trajectory. Equal weighting of all patches gives regions with more teacher patches greater weight, which can change the aggregate objective.

Separate positional outputs depart from the assumption of a shared output, and other loss functions are not covered by this squared error identity. Jointly learning the teacher can introduce gradients through the variation term, so equality of student gradients with the teacher held fixed does not imply equality of the full parameter gradients. The flow objective in Equation 35 also falls outside the setting of a shared output. Its sampled target changes the path input z_s, allowing the network output to depend on which target was drawn. We therefore use the lemma to establish equivalence of supervision under its assumptions, leaving the task information contained in the teacher collection as a separate question.

A.3 Direct and residual prediction under squared error

Definitions and facts. A vector V is square-integrable when $\mathbb{E}\|V\|_2^2 < \infty$. Its conditional mean $\mu(B) = \mathbb{E}[V \mid B]$ averages possible target values for a given deployment input, coordinate by coordinate. This differs from averaging observed vectors within one region. Conditional Jensen's inequality gives $\mathbb{E}\|\mu(B)\|_2^2 \leq \mathbb{E}\|V\|_2^2$, so the conditional mean is also square integrable. For an integrable random variable W, the tower property $\mathbb{E}\mathbb{E}[W \mid B] = \mathbb{E}W$ allows expectations to be evaluated by first conditioning on B. A measurable function of B can be taken outside that conditional expectation when the product is integrable. For the products below, square-integrability and the Cauchy-Schwarz inequality ensure this condition holds.

Proof of Proposition 1. Suppress the region index and write $V = t_i$, $q = q_i$, $x = x_i$ and $\mu = \mathbb{E}[V \mid B]$. To separate the part of the error that depends on the prediction, we expand $V - q = (V - \mu) + (\mu - q)$ and take expectations,

$$\mathbb{E}\|V - q\|_2^2 = \mathbb{E}\|V - \mu\|_2^2 + \mathbb{E}\|\mu - q\|_2^2 + 2\mathbb{E}[(V - \mu)^\top (\mu - q)]. \tag{16}$$

The definition of the conditional mean gives $\mathbb{E}[V - \mu \mid B] = 0$. Both μ and q are functions of B, so their difference can be taken outside the conditional expectation. The cross term therefore vanishes,

$$\mathbb{E}[(V - \mu)^\top (\mu - q)] = \mathbb{E}\left[\mathbb{E}[V - \mu \mid B]^\top (\mu - q)\right] = 0. \tag{17}$$

Substituting this result gives

$$\mathbb{E}\|V - q\|_2^2 = \mathbb{E}\|V - \mu\|_2^2 + \mathbb{E}\|\mu - q\|_2^2. \tag{18}$$

The first term is independent of the prediction. The second is nonnegative and vanishes precisely when $q = \mu$ almost surely. Since μ is square integrable, this unique optimum belongs to the allowed prediction class.

For residual prediction, x is already a square integrable function of B. Every square integrable residual $r(B)$ therefore defines a square integrable output $q = x + r$. Conversely, every such output defines a square integrable residual $r = q - x$. This bijection means that the unrestricted direct and residual predictors can produce exactly the same outputs. The residual optimum is consequently $r(B) = \mu(B) - x$, giving the same optimal prediction $\mu(B)$. □

Interpretation of prediction error. The conditional mean can vary less across observations than the teacher target. To see why, we apply Equation 18 to the constant prediction $q = \mathbb{E}[V]$. The tower property gives $\mathbb{E}[\mu] = \mathbb{E}[V]$, yielding

$$\mathbb{E}\|V - \mathbb{E}[V]\|_2^2 = \mathbb{E}\|V - \mu\|_2^2 + \mathbb{E}\|\mu - \mathbb{E}[\mu]\|_2^2. \tag{19}$$

The first term is nonnegative, so the total centered variation of μ cannot exceed that of V. Reduced variation is therefore compatible with optimal prediction under squared error. This explains a property of the unknown conditional mean, but does not identify why a particular learned network produces less varied features.

Equation 18 applies to each aligned region and feature coordinate in a bag. It also extends to weighting by a measurable function $w(B)$ with $0 < w(B) < \infty$ almost surely, provided $\mathbb{E}[w(B)\|V\|_2^2]$ and $\mathbb{E}[w(B)\|q\|_2^2]$ are finite. Conditional Jensen's inequality ensures a finite weighted second moment for μ as well. Multiplying each squared error inside the expectations by $w(B)$ gives the corresponding weighted identity. The weight is fixed when conditioning on B, so the cross term still vanishes and the conditional mean remains the unique optimum almost surely. A weight that depends on the target beyond B can change that optimum.

Observed reconstruction errors estimate the sum of conditional target variation and predictor error under the evaluation distribution. Reconstruction error alone does not separate these contributions, which depend on the unknown conditional mean. In interpreting the experiments, we also distinguish both terms from representation loss and model excess in Theorem 1.

The shared unrestricted optimum leaves room for differences between neural predictors. A native skip around a bottleneck can change both the available outputs and the initial function, while training procedures can also affect the fitted result. Our experiments compare these finite predictors. Their errors depend on the evaluation population as well, since a population shift can change the conditional mean and conditional variation. Poor prediction by a trained model alone cannot establish an information limit in the available input B.

B DATA AND COMPARISON DESIGN

The prediction and transfer studies use cross-resolution distillation in Stage 1, followed by downstream task training in Stage 2. Teacher comparisons instead fit task models directly on observed features. This appendix describes the datasets, physical fields, extraction and training procedures for these studies, then explains their differences and the variation across training realizations.

B.1 DATASETS AND TASK PARTITIONS

The study includes five classification and grading datasets and five cancer-specific TCGA cohorts for overall survival. Each TCGA cancer cohort forms a separate evaluation dataset. Public data come from CAMELYON16, CAMELYON17, BRACS, PANDA and TCGA (Ehteshami Bejnordi et al., 2017; Bándi et al., 2019; Brancati et al., 2022; Bulten et al., 2022; National Cancer Institute, n.d.). Private-CRC is a private colorectal polyp dataset. Figure 3(a) summarizes their anatomical specimen sites, tasks and included slide/image counts. Table 5 distinguishes slides or images from the groups used for task splitting. It shows seed 42 because membership and counts can change across realizations. All reported evaluations use internal holdouts.

Table 5: Included datasets, observation counts and task partitions at seed 42. The Slides/images column gives the number of observations, while training, validation and test counts refer to the grouping unit in the preceding column. Slides or image identifiers do not establish patient independence. Higher values are better for every primary metric.

Dataset	Metric	Slides/images	Group	Train / val. / test
CAM16	AUC	270	Slide	194 / 22 / 54
CAM17	Macro F_1	499	Challenge group	70 / 10 / 20
Private-CRC	Macro F_1	837	Slide	604 / 67 / 166
BRACS	Macro F_1	546	Patient	135 / 17 / 37
PANDA	Quadratic κ	10,615	Image	7645 / 847 / 2123
TCGA-KIRC	Case C-index	494	Case	337 / 52 / 101
TCGA-KIRP	Case C-index	283	Case	176 / 28 / 55
TCGA-LUAD	Case C-index	487	Case	291 / 45 / 88
TCGA-STAD	Case C-index	320	Case	199 / 33 / 63
TCGA-UCEC	Case C-index	272	Case	172 / 26 / 50

Figure 3(a) combines CAM16 and CAM17 into 769 lymph node records and KIRC and KIRP into 777 kidney records. The remaining sites each correspond to one evaluation dataset. CAM16 and CAM17 sample lymph nodes from breast cancer patients, while BRACS samples breast tissue and UCEC samples uterine endometrium. These specimen sites describe the tissue sampled rather than distinct disease types or independent patient populations. For a site with n included records, the colored outer radius is $r = .80 + .20\ln(n/270)/\ln(10615/270)$. The normalization uses the smallest and largest individual dataset counts, retaining the same anchors when sites are grouped. Equal angular sectors and compressed radii do not give areas proportional to sample size or disease prevalence.

CAM16 predicts tumour presence. CAM17 distinguishes negative slides, isolated tumour cells, micrometastases and macrometastases rather than patient pN stage. Its challenge groups combine slides from different biological patients within a centre (Bándi et al., 2019). Private-CRC distinguishes hyperplastic polyp, sessile serrated lesion, tubular adenoma and tubulovillous adenoma, with 212, 201, 207 and 217 slides, respectively. Normal slides are excluded. BRACS predicts seven tissue and lesion categories. PANDA predicts ISUP categories 0 through 5, including benign tissue in category 0. The TCGA cohorts comprise kidney renal clear cell carcinoma (KIRC), kidney renal papillary cell carcinoma (KIRP), lung adenocarcinoma (LUAD), stomach adenocarcinoma (STAD) and uterine corpus endometrial carcinoma (UCEC) (National Cancer Institute, n.d.). Their task models use four survival bins derived from the task training partition.

CAM16 and Private-CRC partitions are stratified by slide label. CAM17 assigns 14, 2 and 4 challenge groups per centre to training, validation and testing. BRACS groups slides by patient and stratifies by each patient's multilabel vector. PANDA stratifies image identifiers by provider and grade. TCGA groups by case and stratifies by event and time quartile, except UCEC, which uses event alone.

The five holdouts overlap across realizations and are not independent folds. Task membership and selection rules match across models within each contrast.

The test sets differ substantially in the support available for each metric. One BRACS class has only three test patients at seed 42. Across realizations, CAM17 has 3 to 8 test slides with isolated tumour cells, and seed 44 has none in validation. At seed 42, the survival cohorts KIRC, KIRP, LUAD, STAD and UCEC contain 34, 9, 33, 27 and 7 test events and 2,083, 252, 1,550, 1,012 and 247 comparable pairs, respectively. Sparse class and event counts limit the precision of task comparisons.

B.2 Physical fields and frozen features

AtlasPatch selects tissue patches for separate pretrained feature encoders (Alagha et al., 2026). Its fine-tuned SAM2 segmenter (Ravi et al., 2025) predicts a tissue mask from a box spanning a $1.25\times$ thumbnail capped at 1024 pixels. Mask contours are mapped to level-0 slide coordinates. The extraction configuration disables the minimum tissue area filter and rejection of white or black patches. A patch is accepted when at least one of four probes, offset by a quarter patch width around its centre, lies in tissue and the centre is not strictly inside a detected hole.

Both magnifications use 256×256 pixel patches at a stride of 256 pixels. Extraction follows the recorded objective power and resizes pyramid reads to the requested width when needed. It does not enforce a common number of microns per pixel. Physical patch extent therefore depends on slide metadata, even when nominal magnification and pixel width agree.

A $5\times$ patch with level-0 origin (x, y) covers sixteen $20\times$ positions with origins $(x + cw, y + rw)$ for $r, c \in \{0, 1, 2, 3\}$, where w is the $20\times$ footprint. The parent footprint is $4w$, and fields extending outside the slide are excluded. New reads follow these positions from left to right within each row and then from top to bottom, independently of the global grid origins in the earlier caches.

The complete field study selects at most 128 eligible regions per slide without using task outcomes. It first samples one parent from each occupied cell of a 16×8 coverage grid, then fills the remaining places uniformly from eligible parents. All representations and task realizations use this selection. Reserve parents recorded in advance replace unreadable fields before bags are assembled. The resulting support retains all 837 Private-CRC slides and all 10,615 PANDA images. Native features come from earlier caches, while the aligned teacher fields use new reads. Requiring every teacher position distinguishes this support from unrestricted native $5\times$ bags.

The released UNI2-h checkpoint uses a modified ViT-H/14 and represents a patch by its 1,536-dimensional class token after the final LayerNorm (Chen et al., 2024; **?**). The earlier AtlasPatch extractor obtains its evaluation transform from the published configuration, using a 224-pixel bilinear resize, crop fraction one and ImageNet channel normalization. New extraction implements the same intended transform with an antialiased tensor resize. Before encoding, the reader applies OpenCV linear resizing when the pyramid read differs from 256 pixels.

Summary predictors also receive the 768-dimensional normalized class token from the published DINOv3 ViT-B/16 on the same $5\times$ field (Siméoni et al., 2025). DINOv3 uses the full 256-pixel patch. Both encoders normalize RGB values in $[0, 1]$ with means $(.485, .456, .406)$ and standard deviations $(.229, .224, .225)$, without random cropping or stain normalization. Concatenating the cached UNI2-h and DINOv3 vectors gives the 2,304-coordinate native base. Aligned $20\times$ UNI2-h vectors define the teacher targets. The two encoders remain frozen during summary prediction and task training.

Differences between extraction pipelines. We checked whether the earlier caches and new extraction produced the same features by re-encoding eighty physical regions. Each check uses one $5\times$ region from the first successfully processed slide of an extraction chunk. The relative L_2 discrepancy has median $.0195$ and range $[.0064, .1870]$. All ten datasets contribute between four and twenty checks. This deterministic sample shows that the pipelines can produce different vectors for the same region, while leaving the contributions of image reading, interpolation and precision unresolved.

These extraction differences qualify the comparison $M - X$, which changes both magnification content and the pipeline. By contrast, $H - M^{16}$ and $M^{16} - M$ reuse exactly the same $20\times$ vectors, as do spatial summaries and regional means. Their differences cannot arise from using cached rather than newly extracted teacher features. Trained predictors and their initial references likewise share

feature banks within each study. Isolating magnification itself in $M - X$ would require extracting both scales through a common pipeline and repeating the task comparison.

B.3 Token order and recipient processing

Complete-field bags preserve the saved order of selected regions. Within a region, the sixteen teacher vectors follow the 4×4 grid from left to right and top to bottom. The repeated mean occupies those same sixteen positions. Bags are not sorted globally by coordinates. A 2,048-token cap accommodates every regional bag of at most 128 tokens and every expanded bag of at most 2,048 tokens, so loaders preserve all rows and their order. Historical bags use a separate cap of 49,152 tokens and, when needed, deterministic selection spaced over coordinates.

TransMIL first projects each token to 512 coordinates. For a bag of N tokens, it sets $L = \lceil \sqrt{N} \rceil$ and repeats the first $L^2 - N$ vectors to form a square sequence. A class token and Nyström attention layer precede the pyramid position encoding generator, whose depthwise convolutions of sizes 7, 5 and 3 operate on the $L \times L$ sequence and add their outputs to its input. A second attention layer and the normalized class token produce the prediction. This square follows serialized order, since TransMIL does not receive slide coordinates. The inputs H and M^{16} consequently share token count, order, padding and positional processing. Comparing M^{16} with M changes these operations together with sequence length, so it tests their combined effect.

B.4 Patch counts and representation size

Native $5\times$ bags contain fewer patch vectors in every evaluation dataset, as Figure 3(b) shows using the patch-count presentation of XMAG (Su et al., 2025). We count stored UNI2-h feature rows for the same 14,623 slide or image records at both magnifications, covering all included task partitions. Every record has a readable, nonempty bag at both scales. The caches contain 4,247,420 $5\times$ rows and 62,321,475 $20\times$ rows before parent selection or task token caps. These native bags retain the patches accepted at each magnification without requiring identical tissue coverage or patchwise correspondence. Their empirical ratios differ from the exact sixteen children per parent imposed by the complete-field construction.

The two rows in Figure 3(b) use the same vertical scale. Each bar pair averages over the same observations at both magnifications. Reduction labels divide the $20\times$ mean by the $5\times$ mean, which equals the ratio of their total counts. This differs from averaging ratios calculated separately for each slide. Table 5 retains the exact observation counts. The figure describes the included study sample rather than the complete original dataset releases.

Reducing the number of task tokens does not reduce the number of feature coordinates by the same factor when augmentation increases token width. Figure 7 derives both quantities for one complete physical field. Regional means and native UNI2-h features each occupy one token, while individual teacher features and repeated means occupy sixteen. Summary augmentation also occupies one token, retaining the 2,304-coordinate native base and appending either 768 projected coordinates or four concatenated 1,536-coordinate vectors. Relative to individual teacher features, these two augmented inputs reduce the total number of feature entries by factors of 8 and approximately 2.91, respectively, while each reduces token count by a factor of sixteen. The mean and spatial controls within each augmentation study retain its full delivery width, including repeated means or zero coordinates where specified.

Observed teacher summaries still require $20\times$ extraction. Their distilled predictions use only $5\times$ images at deployment, with both UNI2-h and DINOv3 supplying the retained native base. Patch counts and input dimensions describe different parts of that computation. Encoder work, recipient projection and padding, and intermediate activations also contribute to runtime and memory use, so neither figure establishes a processing speedup.

B.5 Cross-resolution distillation and downstream fitting

For summary prediction, the LODO strategy excludes the entire downstream evaluation dataset from Stage 1 training, transformation estimation and predictor selection. Aligned magnifications from the other nine datasets train the predictor without task labels. This exclusion is repeated for each of the

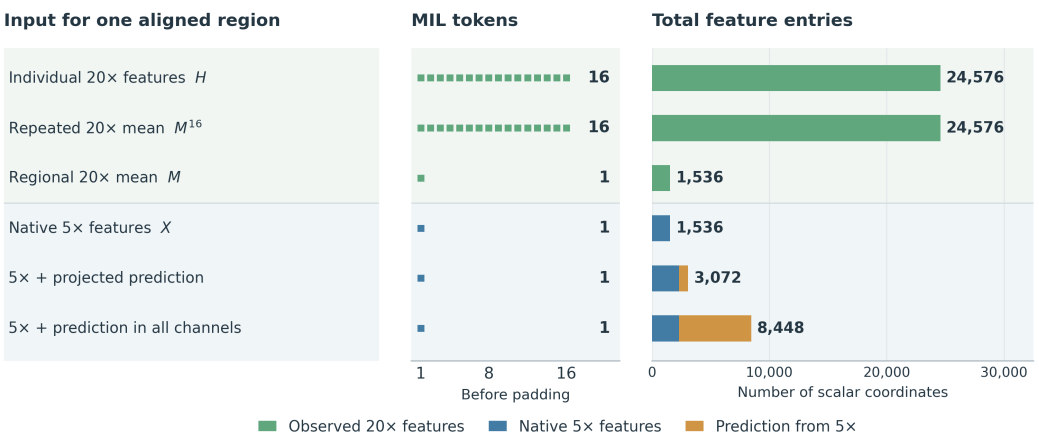


Figure 7: **Token count and feature width jointly determine input size.** The middle panel counts task tokens per complete region before recipient padding. The right panel counts their scalar coordinates, with the native base and appended prediction shown separately for augmentation. Individual teacher features H and repeated means M^{16} each use sixteen 1,536-coordinate tokens. The regional mean M and native $5\times$ vector X each use one. Projected augmentation and augmentation in all feature channels deliver one token of width 3,072 and 8,448, respectively. The four summaries that retain every feature channel are concatenated within that token. These are derived representation dimensions, without measurements of memory use or runtime.

ten evaluation datasets. When one TCGA cancer cohort is excluded, the other TCGA cancer cohorts remain eligible for distillation. All transformations and coordinate scales exclude the evaluated dataset, and distillation validation error selects the predictor checkpoint. Stage 2 freezes the encoders, transforms and predictor, then trains a new task model using the excluded dataset's training labels. Validation groups from that dataset select the task model, and separate test groups provide task scores. Stage 2 uses these within-dataset partitions rather than repeating the Stage 1 dataset exclusion. Test outcomes select neither stage. Figure 2 illustrates these roles. The separate encoder experiment in Appendix E.6 instead adapts UNI2-h during Stage 1 before freezing it for task training.

In Figure 2, the native vector $b_i = [x_i; u_i]$ combines UNI2-h and DINOv3 features. The mean target is $t_i = [a_i; \psi(a_i)]$, where a_i averages projected teacher features and the cosine map acts on that average. The diagram's prediction bag T corresponds to Q in Equation 1, so its deployed representation is $Z = [B; Q]$. Its DINOv3 feature bag U is distinct from the generic task input U in Theorem 1. Appendix C.1 specifies this mean summary, while raw and DCT targets use their own constructions.

Table 6: Task training settings for the primary TransMIL comparisons. Decay is the AdamW weight decay, and accumulation counts slide batches per optimizer update. CAM16 selects checkpoints by validation macro F_1 and reports test AUC. Other task models and encoder adaptation use the exceptions specified in their respective sections.

Dataset	Epochs	Learning rate	Decay	Accum.	Selection
CAM16	15	.0004	.05	16	Macro F_1
CAM17	15	.0002	.01	16	Macro F_1
Private-CRC	15	.0002	.01	16	Macro F_1
BRACS	15	.0002	.01	16	Macro F_1
PANDA	20	.0004	.05	16	Quadratic κ
KIRC	20	.0001	.001	1	Case C-index
KIRP	20	.0002	.00001	1	Case C-index
LUAD	20	.0001	.00001	1	Case C-index
STAD	20	.00005	.00001	1	Case C-index
UCEC	20	.00005	.00001	1	Case C-index

The mean predictor illustrated in Figure 2 uses a GELU network of width $2304 \to 512 \to 768$ and samples the permitted training population from the other nine datasets. Its separate preprocessing and selection panels contain at most 32 groups and 256 regions per group per distillation dataset. Direct and residual prediction uses sixteen fitting groups, eight validation groups and at most 32 regions per group. The component diagnostic divides each distillation dataset into eight base training groups, eight correction groups, four validation groups for checkpoint selection and four independent evaluation groups. Appendix D.4 specifies the diagnostic, and Appendix E.4 gives the direct and residual architectures. These distinct samples define the scope of each reconstruction comparison.

Unless a comparison states an exception, frozen-feature predictors train for 4,096 AdamW updates with batch size 1,024, learning rate .001, weight decay .0001 and gradient clipping at 1. Checkpoints are evaluated every 512 updates. Appendix D.6 gives the correction and denser evaluation schedules, while Appendix E.6 specifies encoder adaptation. Moments estimated from distillation training data weight datasets equally, groups equally within datasets and regions equally within groups. The default coordinate scale is $s_d = \max(\mathrm{SD}_d, \tau)$, where $\tau = \max(10^{-6}, 10^{-3}\,\mathrm{median}_{d:\mathrm{SD}_d>0}\,\mathrm{SD}_d)$. Models within a comparison share the specified transformations and delivery scales.

The primary TransMIL models use AdamW, a cosine schedule and warmup ratio .05, with task settings given in Table 6. Complete fields supply at most 128 regional tokens or 2,048 expanded tokens. Studies using the original bags cap input at 49,152 tokens. Differences in support prevent absolute scores across these studies from isolating a single factor.

B.6 Comparison designs

Tissue support differs across studies in three ways. Complete fields contain at most 128 selected parents per slide with all sixteen newly encoded teacher positions. Historical shared support retains each $5\times$ row with at least one aligned cached teacher vector, allowing the number of teacher patches to vary. Unrestricted native bags retain the original $5\times$ rows without requiring teacher observations during task evaluation. The latter two can reach the 49,152-token cap. We compute effects within studies that share support, width and distillation populations.

Teacher content and token count. Teacher replacement in Table 2 uses complete fields and 1,536-coordinate inputs. Native features X and teacher means M provide one token per parent, while repeated means M^{16} and individual features H provide sixteen. Each receives a separate TransMIL fit without standardization using distillation training data or an additional native block. Seeds 42 through 46 vary task splitting and initialization. Means and sample SDs describe those realizations. The additional task models in Appendix C.8 instead use historical shared support, distillation and split seed 42, and task seeds 1002 through 1006.

Teacher augmentation with native features retained. Teacher augmentation in Table 1 retains the 2,304-coordinate UNI2-h and DINOv3 base on complete fields. Four projected DCT blocks add 192 coordinates each. True teacher components, a true mean with spatial zeros, and all zeros share normalization estimated from distillation training data and the total width of 3,072. Seeds 42 through 46 vary distillation transformations and task realizations. The comparison reports primary task metrics and classification NLL. Appendix C.5 provides the full comparison.

Neural prediction of regional mean and spatial variation. The component diagnostic in Table 19 samples complete fields separately from task bags. Each distillation dataset contributes eight base training groups, eight correction groups, four validation groups for checkpoint selection and four independent evaluation groups, with at most 32 fields per group. Eight diagnostic validation groups come from the excluded dataset. A $2304 \to 512 \to 768$ GELU network predicts four 192-coordinate blocks. Moments of the native $5\times$ features in the base training groups determine the transformations, centering the mean block and scaling all blocks. Five distillation realizations quantify reconstruction without fitting task models. Correction keeps the base predictor frozen and learns a second head on its own groups.

Spatial summaries with all channels retained. The full-channel study in Appendix C.6 trains separate heads for the 1,536-coordinate mean and 4,608 nonconstant coordinates. Each distillation dataset supplies sixteen fitting groups and eight selection groups, with at most 32 fields per group.

Moments of native $5\times$ features from the distillation training groups normalize both targets, and those groups determine a common scale for all decoded summaries. Complete fields support task evaluation. Each token has the 2,304-coordinate native base plus four concatenated vectors of 1,536 coordinates, totaling 8,448. Five realizations jointly vary distillation, task splitting and task training. Primary metrics and NLL describe separate aspects of task performance. The alternative compression in Appendix C.7 shares support and total width but has its own calibration on distillation training data and three teacher representations.

Direct and residual prediction. Direct and residual predictors in Figure 16 learn the same regional mean from complete fields in the distillation datasets. Each distillation dataset contributes sixteen fitting groups and eight selection groups. The head receives one native UNI2-h vector and the slide mean, for 3,072 coordinates with common standardization and target error scales. Its 1,536-coordinate output replaces the native vector in unrestricted task bags. The residual skip changes access to native information despite identical output width. Five realizations jointly vary distillation, task splitting and task training.

The independent analysis in Appendix D.1 reuses these selected heads on complete fields from task test groups, excluding their distillation training and validation groups. Its constant reference is the training target mean estimated from the original distillation training sample. Permutation moves native features, slide context and the residual skip together. The earlier training-field figure and this test analysis retain separate evaluation populations.

Native retention independently of prediction. Native retention in Figure 6 holds the prediction q_i unchanged and supplies either $[x_i; q_i]$ or $[0; q_i]$ at width 3,072. All native rows, task assignments, initial task weights and training settings match. Four hundred task fits cross direct or residual prediction, predictor final-layer initialization and retention over ten datasets and five realizations. The main figure shows NLL changes within each realization. Figure 17 adds the primary scores, and the appendix tables report absolute results for each input. Appendix E.5 explains recovery of the complete native input from the retained version.

Spatial access at predictor input. The input study in Appendix D.7 supplies either repeated global DINOv3 features or its local grid to a predictor of the same 768-coordinate regional mean. Predictions are appended to the 2,304-coordinate global base on complete fields. The distillation training and validation groups, normalization and checkpoint selection match, and five joint realizations provide task scores for each input. The target contains no nonconstant spatial blocks. Because the global base need not recover every local feature used by the predictor, it does not establish the complete-input recovery required to cancel representation loss in the theorem.

B.7 Variation across realizations and reproducible scope

The five realizations vary the factors specified by each study. Teacher content comparisons vary task splits and initialization over seeds 42 through 46. Comparisons against initial predictions and the spatial study also vary Stage 1 training. The comparison across task models holds distillation and split seed 42 constant and varies task initialization over seeds 1002 through 1006. Reported sample SDs describe the corresponding variation, without separating the contribution of each factor.

We compute each metric within a realization and average the five values with equal weight. Classification retains slide or image observations. Survival C-index first averages slide risks within each case before assigning the case outcome. Spatial-study NLL instead retains slide weights and each realization's training-derived time bins. Absolute NLL values are compared only within the same dataset and evaluation protocol. Task predictions are not pooled across overlapping test sets.

The tables report absolute performance for each input rather than differences alone. Each comparison names its reference and metric direction. The five holdouts can overlap, so their SD is not a measure of uncertainty from five independent samples. Descriptive ordering does not establish a statistically reliable population improvement or equivalence. Saved comparison records remain available for subsequent analysis.

Model training for the comparisons displayed in the main Results sectionAll experiments presented in the main test uses NVIDIA RTX 6000 Ada GPUs with PyTorch 2.6 and CUDA 12.4. The

corresponding appendix analyses, including neural correction, teacher-summary compression and comparisons of predictor inputs, reuse or extend these NVIDIA experiments. Training generally uses, generally using BF16 forward passes and, FP32 losses, with and no TF32 matrix multiplication disabled.

The experiments retaining all spatial channels in Appendices C.6 and E.1 and the comparison across MIL architectures in Appendix C.8 use AMD MI350X GPUsll spatial channels use AMD MI350X GPUs with PyTorch 2.11 and ROCm 7.2. Theirose task models use FP32, while the predictors in the full-channellatter study use BF16 forward passes and FP32 losses. Feature-geometry measurements and correspondence maps use CPU inference and analysis of saved NVIDIA-trained predictors.

These are distinct execution settings. The project repository contains numerical exports and rendering scripts that reproduce the displays. The compact manuscript package does not contain every configuration and case prediction needed to rerun all model fits.

C TEACHER TARGETS AND THEIR TASK UTILITY

Before predicting teacher features, we test which parts help the task. This appendix defines regional means and spatial summaries, then compares their utility. Replacing native features tests a teacher representation on its own. Adding teacher features to the native input tests whether they contribute useful complementary information.

C.1 THE MEAN SUMMARY IN THE TRAINING DIAGRAM

Figure 2 illustrates distillation with a 768-coordinate summary of the regional mean. Let $P \in \mathbb{R}^{384\times1536}$ have independent Gaussian entries of variance $1/384$, and define $a_i = K_i^{-1}\sum_j Ph_{ij}$. A random feature map follows Rahimi & Recht (2007),

$$\psi(v) = \sqrt{\frac{2}{384}}\cos\left(\Omega\frac{v}{\max(\|v\|_2, 10^{-6})} + \beta\right), \qquad \Omega_{jk} \sim \mathcal{N}(0,1), \quad \beta_j \sim \mathrm{Unif}[0, 2\pi). \tag{20}$$

The diagram uses $t_i = [a_i; \psi(a_i)]$, with the random map held constant. Both blocks are functions of the same regional mean, so the second block adds no independent teacher observation. This target explains the example in the diagram. The main teacher comparison instead uses the DCT components defined below, and the raw mean prediction study uses all 1,536 original coordinates.

C.2 THE FOUR SPATIAL COMPONENTS

Fourier Compressor reduces a set of visual tokens by applying a spatial DCT, retaining low frequency components and decoding them with an inverse DCT (Wang et al., 2026). We use this construction to summarize sixteen independently encoded $20\times$ patches within one $5\times$ region. For their 4×4 feature field V, let $C = \mathrm{DCT}_{\mathrm{ortho}}(V)/4$, where the transform is an orthonormal two-dimensional DCT of type II. With $c_r = \cos(\pi(2r+1)/8)$ for $r = 0, 1, 2, 3$, the four retained vectors are

$$C_{00} = \frac{1}{16}\sum_{r=0}^{3}\sum_{s=0}^{3} V_{rs}, \qquad C_{01} = \frac{\sqrt{2}}{16}\sum_{r=0}^{3}\sum_{s=0}^{3} c_s V_{rs}, \tag{21}$$

$$C_{10} = \frac{\sqrt{2}}{16}\sum_{r=0}^{3}\sum_{s=0}^{3} c_r V_{rs}, \qquad C_{11} = \frac{1}{8}\sum_{r=0}^{3}\sum_{s=0}^{3} c_r c_s V_{rs}. \tag{22}$$

The constant component C_{00} equals the regional mean. The other three components describe broad horizontal, vertical and joint changes across the feature field. These are spatial frequencies of embeddings, rather than image pixels. We call the retained set four low frequency components (LF4). FcaNet also uses DCT components, for channel attention (Qin et al., 2021). We adapt the compression method to the transfer question by projecting the features, normalizing them with distillation training data and predicting them from $5\times$ inputs.

The projected diagnostic in Appendix D.4 uses $V_{rs} = P_{192}h_{rs}$, where P_{192} contains the first 192 rows of the projection bank. Concatenating four components gives 768 outputs. Moments of native $5\times$ features in the base training groups center the constant block and supply the common scales. The separate study in Appendix C.6 retains all 1,536 channels and decodes the four normalized components with an orthonormal inverse DCT on their 2×2 coefficient grid,

$$G = 2\,\mathrm{IDCT}_{2,\mathrm{ortho}}(\widetilde{C}). \tag{23}$$

Decoding a mean-only target gives four copies of its normalized mean because of the factor of two. The four vectors summarize smooth variation across the field and are neither observed patches nor quadrant averages. Flattening them gives 6,144 coordinates. Every representation in this study combines them with the 2,304-coordinate native base $b_i = [x_i; u_i]$ using

$$\alpha = \sqrt{\frac{\mathbb{E}_{\mathrm{train}}\|b_i\|^2}{2\mathbb{E}_{\mathrm{train}}\|\,\mathrm{vec}(G_i)\|^2}}, \tag{24}$$

where $\mathbb{E}_{\mathrm{train}}$ averages over the distillation training data, and the teacher reference determines α on that fitting population. The projected diagnostic measures reconstruction error. The study retaining all channels also tests the decoded predictions in a task model. Their target widths, distillation samples and output transformations differ, so they provide separate evidence rather than an isolated comparison of channel count.

C.3 SPATIAL FREQUENCY OF THE TEACHER FEATURES

To describe the variation removed by regional pooling, we examine the spatial frequencies of the teacher feature field. The magnitude heatmap follows Figure 1 of Fourier Compressor (Wang et al., 2026). Squared energy and permutation of patch positions distinguish the mean from organized nonconstant variation. We sample 1,000 unique complete regions per dataset from saved pools of 1,208 to 1,280 eligible regions. The resulting 10,000 regions contain 160,000 feature vectors from 381 recorded groups and 593 slides. They describe the sampled feature banks, without representing a uniform patient sample or the exact downstream test population.

Original 1,536-coordinate features and their 192-coordinate projections use the same fields. Absolute coefficient magnitudes are averaged over channels and regions, then equally across datasets before applying the display logarithm. Energy fractions divide aggregate squared coefficients, rather than averaging a ratio from each region. In the original features, DC retains 62.56% of total energy and LF4 retains 75.98%. The three nonconstant LF4 modes retain 35.84% of the energy remaining after DC is removed, compared with 20.06% after permutation. The corresponding projected fractions are 62.72%, 76.09% and 35.88%. These summaries use one descriptive sample per dataset, without model fitting or repeated training seeds.

The regional mean contains most feature energy, while the retained spatial components capture a substantial share of the variation that pooling removes. Figure 8 shows that their share of nonconstant energy decreases after patch permutation. Under a uniform permutation, each of the fifteen nonconstant orthonormal modes has the same expected energy, giving a 20 percent reference for the three retained modes. The saved permutation approximates this reference. It describes spatial organization in the embeddings and motivates testing the retained components for task utility, which energy alone cannot establish.

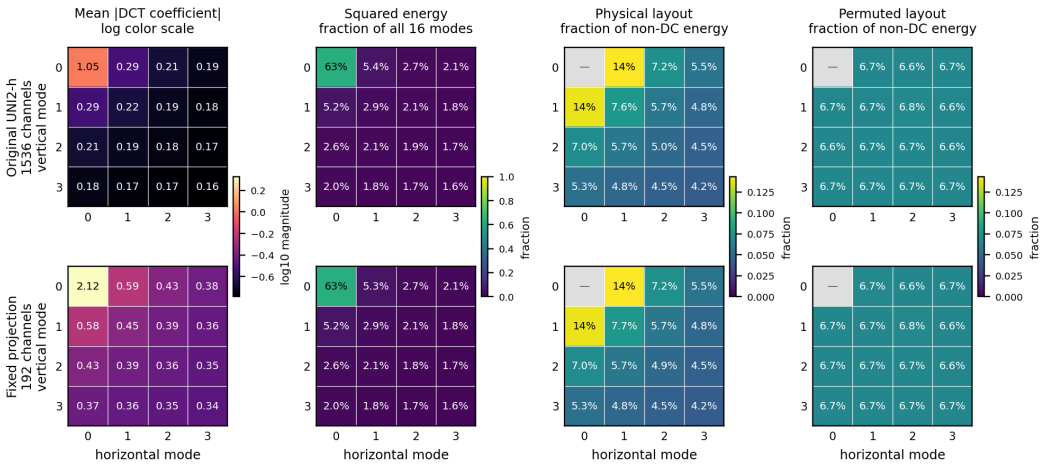


Figure 8: **Broad spatial patterns capture variation beyond the regional mean.** Each row summarizes the same 1,000 complete regions from each of ten datasets, with equal dataset weight, giving 10,000 regions and 160,000 patch vectors. The top row uses all 1,536 feature channels and the bottom row their 192-coordinate projection. Axes index horizontal and vertical DCT modes, not tissue positions. Zero is constant and larger indices represent faster changes. From left to right, panels show mean absolute coefficient magnitude, fractions of total squared energy, and fractions of nonconstant energy before and after patch permutation. Brighter colors indicate larger values within each scale. Magnitude colors use a base-ten log while cell labels retain the original values. Gray marks the removed DC component. The retained LF4 modes occupy the upper-left 2×2 block. These descriptive feature summaries motivate testing spatial targets for task benefit.

C.4 SEPARATING REGIONAL CONTENT FROM TOKEN COUNT

To separate teacher content from repetition, we train task models on four representations of the same complete fields. Native input X has one $5\times$ vector per region. The teacher mean M replaces it

with the average of sixteen aligned $20\times$ vectors, while M^{16} repeats that average in sixteen slots. Individual features H occupy the same slots in physical order. Comparing H with M^{16} holds token count and width constant. Every representation receives a separate task fit with shared initialization and selection rules. Table 2 reports their absolute performance.

Table 7: **Teacher content and token count have distinct task effects.** TransMIL receives native $5\times$ features, a regional $20\times$ mean, sixteen copies of that mean, or the sixteen individual features from the same tissue regions. Teacher inputs require $20\times$ images at evaluation and cover these regions rather than the full slide. Each score comes from a separate TransMIL fit without distillation. Full results for Table 2 are reported as mean $\pm$ sample SD across five realizations.

	$5\times$ input	$20\times$ teacher construction		
Dataset / metric	Native $5\times$ X	Regional mean M	Repeated mean M^{16}	Individual $20\times$ H
Tokens per region	1	1	16	16
CAM16 / AUC	0.8430 ± 0.0215	0.8993 ± 0.0118	0.8872 ± 0.0540	$\mathbf{0.9047} \pm 0.0360$
CAM17 / F_1	0.5054 ± 0.0612	0.4856 ± 0.0476	0.4904 ± 0.0456	$\mathbf{0.5418} \pm 0.0593$
Private-CRC / F_1	0.8273 ± 0.0332	0.8146 ± 0.0232	0.8290 ± 0.0365	$\mathbf{0.8456} \pm 0.0419$
PANDA / κ	0.8535 ± 0.0125	0.8660 ± 0.0159	0.8706 ± 0.0069	$\mathbf{0.8726} \pm 0.0108$
BRACS / F_1	0.3747 ± 0.0532	0.3731 ± 0.0462	$\mathbf{0.3849} \pm 0.0454$	0.3767 ± 0.1097
KIRC / C-index	0.7127 ± 0.0293	0.7105 ± 0.0418	0.6951 ± 0.0592	$\mathbf{0.7144} \pm 0.0470$
KIRP / C-index	0.7793 ± 0.0995	$\mathbf{0.7802} \pm 0.1074$	0.7735 ± 0.1319	0.7745 ± 0.1178
LUAD / C-index	0.5494 ± 0.0539	$\mathbf{0.5640} \pm 0.0511$	0.5523 ± 0.0461	0.5444 ± 0.0415
STAD / C-index	0.5197 ± 0.0572	0.5402 ± 0.0317	$\mathbf{0.5640} \pm 0.0683$	0.5533 ± 0.0692
UCEC / C-index	0.6383 ± 0.1067	0.6158 ± 0.1350	$\mathbf{0.6602} \pm 0.0645$	0.6589 ± 0.1348

Teacher means do not consistently improve on native features when they replace them. CAM16 and PANDA improve in every realization, while other datasets show mixed directions. Individual teacher features can also help beyond repeated means, but this effect varies across realizations. CAM17 illustrates why the reference matters. Individual features improve on repeated means even though the mean itself performs worse than native input. Useful variation within a teacher representation need not make that representation a better replacement for native features.

Repeating the mean changes task scores without adding feature information. Recovering a single mean from its copies, or generating the copies from that mean, gives equal unrestricted risks under Theorem 1. The observed changes reflect how finite models process the representation. Sequence layout, attention approximation, normalization and optimization remain possible contributors that this experiment changes together. Appendix C.8 examines regional means with additional task models on the historical shared support.

C.5 Teacher information alongside the same native input

To test complementary teacher information, we retain the UNI2-h and DINOv3 base $b_i = [x_i; u_i]$ from Section 4 and append teacher components. Each $20\times$ feature is projected to 192 coordinates before the transform in Equation 22. The four blocks contain the regional mean and three spatial components, normalized using common statistics from distillation training data. The task model receives 3,072 coordinates per region. This construction follows the joint native and teacher input in Theorem 1.

Using the symbols from Section 4, the task inputs are $[b_i; 0_{768}]$, $[b_i; \widetilde{m}_i; 0_{576}]$ and $[b_i; \widetilde{m}_i; s_i]$. The vector b_i contains native $5\times$ features, $\widetilde{m}_i$ is the normalized projected $20\times$ regional mean, and s_i concatenates its three normalized spatial components. Each representation has a separate TransMIL fit, with shared fields, task assignments, initial weights, optimization and selection within a realization. Bags contain at most 128 regions. The five realizations jointly vary distillation transformations, task splitting and training. These results reuse 150 completed reference fits and require teacher observations during evaluation.

A benefit from this augmentation shows that the teacher helps the trained recipient beyond its native input. Under the common loss and population of the theorem, that benefit combines the teacher information contribution with a change in model excess. The experiment measures complementary utility for this training procedure, while leaving those contributions unseparated.

Task loss. Classification NLL uses saved slide or image probabilities, floored at 10^{-7} and renormalized before averaging log loss. Survival NLL is the discrete censored likelihood averaged over slides, with four time bins derived from training cases. An event contributes survival through earlier bins and the hazard in its event bin. A censored observation contributes survival through its censoring bin. Models within a realization share bin boundaries and evaluated slides. Case C-index instead averages slide risks within a case and evaluates their ordering. NLL is reported separately for each dataset because task definitions and training-derived time bins differ. The means and sample SDs summarize the five saved realizations and do not establish a population ranking.

Table 8: Task NLL for teacher summaries appended to the same native input. Lower is better. The reference is native $5\times$ with zero padding. The direction control retains the teacher mean and component norms.

Dataset	Native $5\times$ $[b_i; 0_{768}]$	Native $5\times$ + mean $[b_i; \widetilde{m}_i; 0_{576}]$	Native $5\times$ + mean and spatial $[b_i; \widetilde{m}_i; s_i]$	Native $5\times$ + mean and direction control $[b_i; \widetilde{m}_i; s_i^{\mathrm{rnd}}]$
CAM16	0.6110 ± 0.1605	0.6717 ± 0.3244	$\mathbf{0.5663}\pm0.2296$	0.5702 ± 0.2110
CAM17	$\mathbf{0.8392}\pm0.4304$	0.8605 ± 0.3460	1.0058 ± 0.5968	0.9509 ± 0.5255
Private-CRC	0.6230 ± 0.0551	$\mathbf{0.5563}\pm0.0356$	0.6503 ± 0.1546	0.6924 ± 0.2528
PANDA	1.4584 ± 0.5817	$\mathbf{1.1328}\pm0.4491$	1.3559 ± 0.6075	1.4657 ± 0.6855
BRACS	1.8613 ± 0.5405	$\mathbf{1.7576}\pm0.5130$	1.8878 ± 0.6489	1.9374 ± 0.4963
KIRC	2.0562 ± 0.4878	$\mathbf{1.6173}\pm0.4347$	1.7888 ± 0.5891	1.6763 ± 0.4250
KIRP	1.2932 ± 0.2403	1.2128 ± 0.2770	$\mathbf{1.0900}\pm0.1797$	1.2335 ± 0.2100
LUAD	2.7670 ± 0.6057	2.5482 ± 0.4887	$\mathbf{2.4012}\pm0.4809$	2.5907 ± 0.6864
STAD	2.4675 ± 0.3891	2.5468 ± 0.5305	2.5089 ± 0.4825	$\mathbf{2.3861}\pm0.4141$
UCEC	1.0027 ± 0.1632	0.9424 ± 0.1180	0.9429 ± 0.1190	$\mathbf{0.9311}\pm0.1241$

The regional teacher mean raises the average primary score in every dataset, as Table 1 shows, but its NLL effect is less consistent. Adding spatial components also has mixed effects on NLL. These results concern the same trained models. They show why probabilistic loss and discrimination or grading metrics provide distinct evidence of task utility, as the scope of Theorem 1 requires.

Spatial direction at the same component magnitude. We next test whether the direction of each spatial vector matters beyond its magnitude. Fifty additional task fits retain the true mean and replace each spatial block with a Gaussian direction scaled to its original norm, written s_i^{rnd}. Fields, input width, task initialization and training remain unchanged. Table 9 reports primary scores alongside the same native and teacher references. The original directions help some datasets and reduce performance on others. Every CAM16 realization favors the original directions, but this pattern does not generalize across datasets. The replacement retains teacher-derived norms and alters covariance together with direction. It tests sensitivity to the supplied vectors rather than an isolated biological property.

Table 9: Primary task scores for teacher summaries and their spatial direction control. Higher is better. The reference is native $5\times$ with zero padding. The task metrics follow Table 1.

Dataset	Native $5\times$ $[b_i; 0_{768}]$	Native $5\times$ + mean $[b_i; \widetilde{m}_i; 0_{576}]$	Native $5\times$ + mean and spatial $[b_i; \widetilde{m}_i; s_i]$	Native $5\times$ + mean and direction control $[b_i; \widetilde{m}_i; s_i^{\mathrm{rnd}}]$
CAM16	0.8455 ± 0.0485	0.8621 ± 0.0600	$\mathbf{0.8932}\pm0.0285$	0.8662 ± 0.0348
CAM17	0.5197 ± 0.0336	$\mathbf{0.5594}\pm0.0457$	0.5198 ± 0.0414	0.5297 ± 0.0346
Private-CRC	0.8305 ± 0.0276	$\mathbf{0.8410}\pm0.0272$	0.8305 ± 0.0233	0.8263 ± 0.0342
PANDA	0.8386 ± 0.0067	0.8446 ± 0.0124	$\mathbf{0.8456}\pm0.0114$	0.8357 ± 0.0089
BRACS	0.4211 ± 0.0710	0.4257 ± 0.0783	0.4094 ± 0.0704	$\mathbf{0.4381}\pm0.0722$
KIRC	0.6823 ± 0.0548	0.6892 ± 0.0470	0.6907 ± 0.0435	$\mathbf{0.6990}\pm0.0550$
KIRP	0.7352 ± 0.1486	0.7782 ± 0.0956	0.7563 ± 0.1345	$\mathbf{0.8324}\pm0.0926$
LUAD	0.5381 ± 0.0614	$\mathbf{0.5795}\pm0.0276$	0.5495 ± 0.0214	0.5378 ± 0.0327
STAD	0.5104 ± 0.0188	0.5112 ± 0.0209	0.4947 ± 0.0497	$\mathbf{0.5169}\pm0.0456$
UCEC	0.6499 ± 0.1086	$\mathbf{0.6793}\pm0.1251$	0.6672 ± 0.1257	0.6525 ± 0.1023

C.6 SPATIAL UTILITY WHEN ALL FEATURE CHANNELS ARE RETAINED

This study tests teacher utility and distillation without projecting away feature channels. Separate $2304 \to 512 \to 1536$ and $2304 \to 512 \to 4608$ networks predict the regional mean and non-constant components, giving 100 distillation fits across ten datasets and five realizations. Six task representations retain the same native base and use 8,448 coordinates, giving 300 task fits.

For this comparison, $Z = [b_i; 0_{6144}]$ is native $5\times$ input with zero augmentation. M appends four copies of the teacher regional mean, and S appends the teacher mean and spatial components after decoding. Their neural predictions are $\widehat{M}$ and $\widehat{S}$. The input I_S combines the trained mean prediction with the exact initial spatial head. These symbols identify complete native-plus-summary inputs in Table 10. They are distinct from the standalone teacher bags in Table 2.

The three neural representations use the same trained mean head. Only the spatial output differs, through omission, the trained head or its initial state. Comparing trained and initial spatial outputs tests the contribution of distillation. Comparing trained outputs with omission tests whether supplying them helps beyond the predicted mean. Teacher spatial components improve the average primary score over the teacher mean in several datasets, with adverse directions in others. Neural spatial predictions also have mixed effects. These comparisons do not establish a general primary-score advantage for retaining spatial variation. Appendix E.1 develops the NLL findings using the same inputs.

Reconstruction by the downstream predictors. To connect the task results to their Stage 1 training, Table 11 reports reconstruction by the same mean and spatial heads. Each trains and selects checkpoints using its own component MSE. Every mean head improves on the training target mean in the selection groups, while every spatial head has higher error relative to that constant. Both improve on their own initialization, and all choose step 512, the first eligible checkpoint. Stage 1 training can reduce spatial error without overcoming the constant reference.

These groups select the checkpoints, so their errors describe model selection rather than independent generalization. Independent component errors are unavailable for these heads. The projected diagnostic in Section 5.2 evaluates different predictors and cannot supply that missing evidence.

C.7 AN ALTERNATIVE WAY TO DELIVER SPATIAL SUMMARIES

We also compare three summaries that retain every feature channel. The first decodes the low frequency components, the second repeats the regional mean four times, and the third decodes components after permuting the sixteen teacher positions. Concatenating the four 1,536-coordinate vectors with the 2,304-coordinate native base gives one 8,448-coordinate token per region. All three use the same normalization estimated from distillation training data, delivery scale, fields and task training within each realization, for 150 fits across ten datasets. A region's permutation is unchanged throughout the experiment and preserves its mean, but changes the retained frequency energy. This study has its own calibration on distillation training data, so comparison with Appendix C.6 does not isolate a single design choice.

Table 12 shows that spatial summaries improve average task performance in some datasets and reduce it in others. Permuting the teacher positions can also change that ordering. The results do not establish a common benefit from preserving spatial order. The frequency pattern motivates a compact spatial target, but the downstream comparison is still needed to determine whether its retained variation helps the task.

C.8 REGIONAL MEANS ACROSS TASK MODELS

To test whether regional mean utility persists across task models, we compare TransMIL (Shao et al., 2021), gated ABMIL (Ilse et al., 2018) and RRT-MIL (Tang et al., 2024) on all ten datasets and DSMIL (Li et al., 2021) on the five classification datasets. All receive frozen features. DSMIL combines instance and bag losses equally, uses mutually exclusive cross entropy and evaluates bag logits. Survival models train with discrete time NLL, then average slide risks within each case for C-index. The comparison covers complete training procedures whose differences extend beyond architecture.

Table 10: Task performance when all feature channels are retained. Every input contains the native base, with different teacher or neural summaries in the remaining coordinates. The reference is $Z = [b_i; 0_{6144}]$. The upper and lower panels report primary scores and NLL separately.

Primary metric, higher is better

Dataset	Native 5× Z	Teacher mean M	Teacher spatial S	Predicted mean $\widehat{M}$	Predicted spatial $\widehat{S}$	Initial spatial I_S
CAM16	0.8722 ±0.0275	**0.8955** ±0.0126	0.8653 ±0.0389	0.8341 ±0.0620	0.8835 ±0.0272	0.8463 ±0.0226
CAM17	**0.5261** ±0.0418	0.5111 ±0.0505	0.5242 ±0.0357	0.4968 ±0.0308	0.5042 ±0.0407	0.5092 ±0.0425
Private-CRC	0.8284 ±0.0179	0.8097 ±0.0475	0.8313 ±0.0318	0.8209 ±0.0286	0.8258 ±0.0411	**0.8345** ±0.0274
PANDA	0.8467 ±0.0131	0.8566 ±0.0069	**0.8608** ±0.0089	0.8460 ±0.0119	0.8446 ±0.0139	0.8385 ±0.0160
BRACS	0.4276 ±0.0577	0.3966 ±0.0875	**0.4414** ±0.0755	0.4398 ±0.0827	0.4224 ±0.0871	0.4147 ±0.0525
KIRC	0.6986 ±0.0340	0.6945 ±0.0614	0.6982 ±0.0640	0.6976 ±0.0242	0.6801 ±0.0490	**0.7123** ±0.0358
KIRP	0.8073 ±0.1143	0.7684 ±0.1740	**0.8546** ±0.0371	0.8007 ±0.1032	0.8086 ±0.0982	0.7732 ±0.1654
LUAD	0.5660 ±0.0362	0.5423 ±0.0707	0.5635 ±0.0268	0.5440 ±0.0359	0.5332 ±0.0548	**0.5729** ±0.0334
STAD	0.5164 ±0.0626	**0.5279** ±0.0293	0.5037 ±0.0597	0.5054 ±0.0506	0.5193 ±0.0430	0.5179 ±0.0436
UCEC	0.6218 ±0.0866	0.6282 ±0.0799	0.6198 ±0.0800	0.6424 ±0.1200	0.6378 ±0.0750	**0.6516** ±0.0658

NLL, lower is better

Dataset	Native 5× Z	Teacher mean M	Teacher spatial S	Predicted mean $\widehat{M}$	Predicted spatial $\widehat{S}$	Initial spatial I_S
CAM16	**0.5407** ±0.1000	0.5841 ±0.2263	0.6975 ±0.2300	0.6398 ±0.0860	0.6897 ±0.2603	0.6895 ±0.1239
CAM17	0.7237 ±0.2901	0.7563 ±0.2721	0.8016 ±0.2640	1.1302 ±0.5135	0.9913 ±0.4640	**0.7099** ±0.1403
Private-CRC	0.6161 ±0.1126	0.6279 ±0.0905	0.5790 ±0.0802	**0.5591** ±0.0641	0.6064 ±0.1424	0.5768 ±0.1226
PANDA	1.0584 ±0.4364	0.9727 ±0.1695	1.3085 ±0.7090	0.9212 ±0.0914	**0.8832** ±0.0301	0.9300 ±0.0329
BRACS	2.1090 ±0.5462	1.4396 ±0.3561	2.0671 ±0.5976	1.4789 ±0.2412	**1.4129** ±0.2799	1.7211 ±0.5466
KIRC	**1.6016** ±0.1369	1.9291 ±0.2473	1.8049 ±0.4113	1.7670 ±0.1869	1.7589 ±0.2831	2.1189 ±0.4656
KIRP	1.2996 ±0.1336	1.1313 ±0.0762	**1.0616** ±0.1776	1.2368 ±0.1122	1.1533 ±0.2994	1.3178 ±0.2141
LUAD	**1.9082** ±0.2802	2.2922 ±0.5052	2.3800 ±0.5268	2.4893 ±0.7290	1.9505 ±0.3098	2.2323 ±0.6362
STAD	2.4946 ±0.5711	2.6962 ±0.5397	2.3705 ±0.5052	**2.3427** ±0.4329	2.5259 ±0.4128	2.4605 ±0.3984
UCEC	1.0279 ±0.1880	1.0393 ±0.2126	1.0164 ±0.1590	**0.9815** ±0.0727	0.9907 ±0.1317	1.0041 ±0.1715

The distillation realization and task split use seed 42. Hyperparameter selection uses task seed 1001 and four candidates that cross learning rate multipliers $\{0.5, 1\}$ with weight decay multipliers $\{0.1, 1\}$. The mean validation metric across representations selects one configuration per dataset and recipient. Evaluation then uses this configuration with five new task initializations, numbered 1002 through 1006. Initial weights match across input representations within each recipient and seed. Training lasts 15 epochs for CAM16, CAM17, Private-CRC and BRACS, and 20 for PANDA and survival. All inputs use the same 49,152-token limit and the dataset's validation selection rule.

Table 13 uses the historical shared support, where every region has at least one aligned $20\times$ feature. Patch counts may vary, unlike the complete fields in Table 2. Native and mean inputs are compared within each task model. All four models favor the regional mean on average for CAM16 and PANDA, while the direction varies elsewhere. These results show that the mean can help multiple recipients

Table 11: Reconstruction by the mean and spatial heads used downstream. MSE is evaluated on the groups used for checkpoint selection, with the constant equal to the training target mean for that component. Lower is better. Mean and spatial errors use their own scales.

	Regional mean		Nonconstant components	
Excluded dataset	Predictor	Constant	Predictor	Constant
CAM16	**0.3538**±0.0079	0.6935±0.0160	0.0730±0.0018	**0.0563**±0.0012
CAM17	**0.3461**±0.0107	0.6956±0.0149	0.0713±0.0018	**0.0553**±0.0012
Private-CRC	**0.3812**±0.0102	0.7227±0.0231	0.0752±0.0017	**0.0555**±0.0008
PANDA	**0.3565**±0.0070	0.7000±0.0179	0.0694±0.0019	**0.0508**±0.0017
BRACS	**0.3450**±0.0070	0.6702±0.0130	0.0718±0.0020	**0.0556**±0.0015
KIRC	**0.3460**±0.0083	0.6745±0.0065	0.0747±0.0020	**0.0575**±0.0014
KIRP	**0.3360**±0.0083	0.6798±0.0175	0.0737±0.0023	**0.0569**±0.0015
LUAD	**0.3432**±0.0105	0.7001±0.0201	0.0737±0.0021	**0.0571**±0.0017
STAD	**0.3312**±0.0067	0.6879±0.0151	0.0723±0.0018	**0.0562**±0.0014
UCEC	**0.3288**±0.0106	0.6676±0.0110	0.0729±0.0021	**0.0566**±0.0015

Table 12: Task performance with three teacher summaries at a common input width of 8,448. Every input retains the native base and appends four vectors from the same observed 20× field. The repeated regional mean is the reference. Higher is better.

Dataset	Metric	Repeated mean	Spatial summary	Permuted summary
CAM16	AUC	**0.8820**±0.0287	0.8716±0.0170	0.8730±0.0563
CAM17	F_1	**0.5141**±0.0307	0.5038±0.0433	0.5126±0.0409
Private-CRC	F_1	**0.8575**±0.0194	0.8491±0.0396	0.8455±0.0264
PANDA	κ	0.8569±0.0082	**0.8609**±0.0118	0.8607±0.0110
BRACS	F_1	**0.4319**±0.0873	0.4138±0.0936	0.4094±0.1125
KIRC	C-index	**0.7163**±0.0448	0.6978±0.0569	0.7013±0.0607
KIRP	C-index	0.7748±0.1123	**0.8447**±0.1206	0.8212±0.0797
LUAD	C-index	0.5636±0.0234	**0.5822**±0.0307	0.5638±0.0433
STAD	C-index	0.5394±0.0307	0.5095±0.0417	**0.5448**±0.0413
UCEC	C-index	0.6089±0.1076	**0.6313**±0.1085	0.5952±0.1106

without identifying a consistently superior recipient for teacher information. Five task initializations on one split describe training variability and do not establish a general ranking of architectures.

Table 13: Native $5\times$ and regional $20\times$ mean inputs across downstream models. Each model uses its own native reference on the same tissue support. Higher is better. Five task initializations share one data split. DSMIL is evaluated only for classification and grading.

	TransMIL		Gated ABMIL		DSMIL		RRT-MIL	
Dataset	Native $5\times$	Mean $20\times$	Native $5\times$	Mean $20\times$	Native $5\times$	Mean $20\times$	Native $5\times$	Mean $20\times$
CAM16	0.9605 $\pm$0.0240	**0.9642** $\pm$0.0205	0.9537 $\pm$0.0173	**0.9705** $\pm$0.0128	0.9580 $\pm$0.0036	**0.9759** $\pm$0.0097	0.9494 $\pm$0.0229	**0.9682** $\pm$0.0073
CAM17	**0.6743** $\pm$0.0552	0.6325 $\pm$0.0212	**0.6352** $\pm$0.0259	0.5690 $\pm$0.0300	**0.5522** $\pm$0.0537	0.5373 $\pm$0.0712	**0.6294** $\pm$0.0313	0.5992 $\pm$0.0500
Private-CRC	0.8010 $\pm$0.0317	**0.8497** $\pm$0.0134	**0.8288** $\pm$0.0250	0.8093 $\pm$0.0262	0.8397 $\pm$0.0114	**0.8444** $\pm$0.0126	**0.8443** $\pm$0.0123	0.8394 $\pm$0.0347
PANDA	0.8509 $\pm$0.0040	**0.8630** $\pm$0.0089	0.8363 $\pm$0.0038	**0.8518** $\pm$0.0045	0.8612 $\pm$0.0041	**0.8833** $\pm$0.0085	0.8488 $\pm$0.0050	**0.8832** $\pm$0.0044
BRACS	0.3709 $\pm$0.0319	**0.4035** $\pm$0.0246	**0.4141** $\pm$0.0530	0.3753 $\pm$0.0155	**0.4278** $\pm$0.0360	0.3490 $\pm$0.0322	**0.4603** $\pm$0.0303	0.4116 $\pm$0.0282
KIRC	**0.6800** $\pm$0.0147	0.6768 $\pm$0.0261	**0.6787** $\pm$0.0221	0.6355 $\pm$0.0083	Not evaluated		**0.6988** $\pm$0.0290	0.6846 $\pm$0.0322
KIRP	**0.8437** $\pm$0.0710	0.7603 $\pm$0.0578	**0.8563** $\pm$0.0165	0.8000 $\pm$0.0471	Not evaluated		**0.8429** $\pm$0.0821	0.7889 $\pm$0.0490
LUAD	0.5559 $\pm$0.0340	**0.5768** $\pm$0.0541	**0.5677** $\pm$0.0190	0.5565 $\pm$0.0144	Not evaluated		0.5741 $\pm$0.0729	**0.5763** $\pm$0.0369
STAD	0.5241 $\pm$0.0185	**0.5589** $\pm$0.0355	**0.5379** $\pm$0.0130	0.5318 $\pm$0.0152	Not evaluated		0.5986 $\pm$0.0381	**0.6008** $\pm$0.0218
UCEC	0.5976 $\pm$0.0476	**0.6000** $\pm$0.0466	0.4955 $\pm$0.0251	**0.4972** $\pm$0.0053	Not evaluated		**0.5773** $\pm$0.0521	0.5263 $\pm$0.0557

D NEURAL PREDICTION OF TEACHER TARGETS

This appendix examines prediction of high magnification targets from low magnification inputs. We first evaluate raw regional means on independent test fields, then separate the errors of projected mean and spatial components. Additional studies examine error correction, checkpoint selection and local image features. Each target retains its own coordinates and evaluation population.

D.1 REGIONAL MEAN PREDICTION ON INDEPENDENT TEST FIELDS

To test generalization of regional mean prediction, we reuse the selected direct and residual heads from Appendix E.4, with their original transformations and constant reference, the training target mean. Each realization samples eight groups from the excluded dataset and eight from each of the nine distillation datasets, taking at most 32 complete fields per group. The resulting task-test samples contain 231 to 256 fields from the excluded dataset and 2,278 to 2,304 from distillation datasets. None of these groups contributed to the corresponding distillation training or validation.

Group identities follow the available records. CAM17 uses challenge groups, CAM16 and Private-CRC use slides, PANDA uses images, and the other datasets use patient or case identifiers where available. The analysis preserves separation and weighting at these recorded units. Slide or image separation alone does not establish patient independence.

We reverse the distillation transformations before computing MSE in the original 1,536-coordinate UNI2-h space defined by Equation 31. Evaluation within the distillation datasets weights datasets equally, groups equally within a dataset and fields equally within a group. Evaluation on the excluded dataset uses the latter two levels. Two references make the errors interpretable. The native reference returns the region's original $5\times$ vector, while the constant returns the training target mean estimated from the original distillation training sample. Table 14 reports their errors alongside all four trained predictors.

Table 14: Raw regional mean MSE on independent test fields from the excluded dataset. Native $5\times$ is the reference. Zero and random identify final-layer initialization. The constant $\bar{m}_{\text{train}}$ is estimated from distillation training targets.

	Native $5\times$ x_i	Constant $\bar{m}_{\text{train}}$	Direct $f_\theta(B)_i$		Residual $x_i + r_\phi(B)_i$	
Dataset			Zero	Random	Zero	Random
CAM16	0.1759 ±0.0103	0.0875 ±0.0079	**0.0667** ±0.0067	0.0686 ±0.0066	0.0935 ±0.0061	0.0972 ±0.0076
CAM17	0.1939 ±0.0044	0.0912 ±0.0038	**0.0776** ±0.0049	0.0801 ±0.0053	0.1064 ±0.0048	0.1100 ±0.0045
Private-CRC	0.1876 ±0.0162	**0.1051** ±0.0034	0.1331 ±0.0040	0.1551 ±0.0085	0.1675 ±0.0171	0.2017 ±0.0235
PANDA	0.1554 ±0.0067	**0.0837** ±0.0071	0.0988 ±0.0120	0.1088 ±0.0112	0.1231 ±0.0107	0.1364 ±0.0106
BRACS	0.1882 ±0.0094	0.1099 ±0.0080	**0.0936** ±0.0082	0.0941 ±0.0089	0.1102 ±0.0065	0.1175 ±0.0074
KIRC	0.1697 ±0.0020	0.1218 ±0.0045	**0.1084** ±0.0089	0.1132 ±0.0096	0.1182 ±0.0099	0.1263 ±0.0104
KIRP	0.1549 ±0.0074	0.1099 ±0.0074	**0.0968** ±0.0041	0.1023 ±0.0068	0.1040 ±0.0054	0.1148 ±0.0099
LUAD	0.2074 ±0.0142	0.0954 ±0.0090	**0.0953** ±0.0104	0.1041 ±0.0136	0.1333 ±0.0131	0.1504 ±0.0182
STAD	0.1828 ±0.0083	0.0927 ±0.0107	**0.0866** ±0.0123	0.0902 ±0.0131	0.1080 ±0.0117	0.1144 ±0.0129
UCEC	0.1857 ±0.0093	**0.1118** ±0.0119	0.1183 ±0.0139	0.1294 ±0.0155	0.1261 ±0.0093	0.1386 ±0.0080

All four predictors improve on the constant across the independent test samples from the distillation datasets, but generalization to excluded datasets varies. Direct prediction with zero initialization improves on the constant in seven dataset averages. Residual prediction can approximate the teacher target more accurately than the native feature while remaining less accurate than the constant. These references distinguish progress toward the teacher from prediction of differences between regions.

To test whether prediction depends on correspondence between input and target, we permute complete input tuples within each dataset and evaluation population while keeping targets in place. The native feature, slide context and residual skip move together. All four predictors have lower MSE with aligned inputs in every excluded-dataset realization, as summarized in Table 15. The complete tuple changes, so the comparison does not separate local features from slide context. It evaluates trained heads rather than training on shuffled supervision.

Table 15: Raw regional mean MSE with aligned or permuted input tuples. The native vector, slide context and residual skip are permuted together. Permuted input is the reference within each prediction form.

Zero final-layer initialization				
	Direct		Residual	
Dataset	Permuted	Aligned	Permuted	Aligned
CAM16	0.1205±0.0136	**0.0667**±0.0067	0.1652±0.0131	**0.0935**±0.0061
CAM17	0.1216±0.0037	**0.0776**±0.0049	0.1733±0.0039	**0.1064**±0.0048
Private-CRC	0.1496±0.0041	**0.1331**±0.0040	0.2475±0.0130	**0.1675**±0.0171
PANDA	0.1145±0.0105	**0.0988**±0.0120	0.1769±0.0123	**0.1231**±0.0107
BRACS	0.1217±0.0084	**0.0936**±0.0082	0.1719±0.0088	**0.1102**±0.0065
KIRC	0.1584±0.0029	**0.1084**±0.0089	0.2234±0.0051	**0.1182**±0.0099
KIRP	0.1447±0.0080	**0.0968**±0.0041	0.1994±0.0064	**0.1040**±0.0054
LUAD	0.1296±0.0079	**0.0953**±0.0104	0.2090±0.0138	**0.1333**±0.0131
STAD	0.1226±0.0114	**0.0866**±0.0123	0.1883±0.0185	**0.1080**±0.0117
UCEC	0.1536±0.0133	**0.1183**±0.0139	0.2152±0.0155	**0.1261**±0.0093
Random final-layer initialization				
	Direct		Residual	
Dataset	Permuted	Aligned	Permuted	Aligned
CAM16	0.1229±0.0145	**0.0686**±0.0066	0.1682±0.0149	**0.0972**±0.0076
CAM17	0.1246±0.0051	**0.0801**±0.0053	0.1764±0.0045	**0.1100**±0.0045
Private-CRC	0.1732±0.0083	**0.1551**±0.0085	0.2799±0.0161	**0.2017**±0.0235
PANDA	0.1256±0.0093	**0.1088**±0.0112	0.1905±0.0115	**0.1364**±0.0106
BRACS	0.1229±0.0097	**0.0941**±0.0089	0.1783±0.0097	**0.1175**±0.0074
KIRC	0.1660±0.0044	**0.1132**±0.0096	0.2313±0.0055	**0.1263**±0.0104
KIRP	0.1509±0.0090	**0.1023**±0.0068	0.2095±0.0081	**0.1148**±0.0099
LUAD	0.1391±0.0106	**0.1041**±0.0136	0.2244±0.0172	**0.1504**±0.0182
STAD	0.1271±0.0124	**0.0902**±0.0131	0.1939±0.0187	**0.1144**±0.0129
UCEC	0.1664±0.0152	**0.1294**±0.0155	0.2263±0.0143	**0.1386**±0.0080

Direct prediction has lower MSE than residual prediction in every dataset average under both initializations. Proposition 1 gives the same optimum for unrestricted direct and residual functions. The empirical difference characterizes the finite networks and their optimization rather than a different target optimum.

D.2 Feature distributions before and after distillation

The task scores in Tables 1 and 2 compare observed teacher representations through separately trained slide models. Their inputs have no before and after distillation state. To examine learning separately, we compare native $5\times$ features, observed $20\times$ regional means and the saved initial and trained regional-mean predictors on the independent fields from Appendix D.1. The same fields, targets and training transformations are used before and after training. Each excluded dataset contributes five realizations with 231 to 256 fields from eight recorded groups per realization. Their total is 12,728 field appearances, with possible overlap across realizations. No task labels enter these measurements and no model is retrained.

For each evaluation field, m_i is the regional teacher mean from Lemma 1, and q_i is its prediction. Both use the original $d = 1536$ UNI2-h coordinates. Weights w_i give each recorded group equal mass and each field equal weight within its group. They sum to one, giving $\bar{m} = \sum_i w_i m_i$ and $\bar{q} = \sum_i w_i q_i$. We compare target MSE and the cosine similarity of corresponding vectors. For

positive target variance, the centered variance ratio is

$$v = \frac{\sum_i w_i \|q_i - \bar{q}\|_2^2}{\sum_i w_i \|m_i - \bar{m}\|_2^2}. \tag{25}$$

A ratio of one indicates equal total centered variation, without requiring the distributions or individual predictions to agree. We compute the Pearson correlation between $\|q_i - q_j\|_2$ and $\|m_i - m_j\|_2$ over $i < j$, weighted proportionally to $w_i w_j$. It describes relative distances between regions without requiring identical feature axes, following the representational similarity approach (Kriegeskorte et al., 2008). Constant outputs have no defined correlation of pairwise distances. A zero output also has no defined cosine similarity. These cases remain undefined rather than receiving an artificial score.

Raw MSE depends on vector magnitude as well as direction. For nonzero q_i and m_i, the Euclidean inner product gives

$$\|q_i - m_i\|_2^2 = \|q_i\|_2^2 + \|m_i\|_2^2 - 2\|q_i\|_2\|m_i\|_2 \cos(q_i, m_i). \tag{26}$$

Cosine similarity alone does not determine reconstruction error in these unnormalized coordinates. To separate agreement in the common feature average from agreement between regions, we also compute centered MSE. With the averages and weights defined above,

$$\frac{1}{d}\sum_i w_i \|q_i - m_i\|_2^2 = \frac{1}{d}\|\bar{q} - \bar{m}\|_2^2 + \frac{1}{d}\sum_i w_i \|(q_i - \bar{q}) - (m_i - \bar{m})\|_2^2. \tag{27}$$

The cross term vanishes because each centered representation has weighted mean zero. The first term measures average displacement and the second measures mismatch after each representation's average is removed. Centering uses evaluation vectors only to describe their errors and does not alter the predictor. These quantities distinguish common feature structure from reconstruction of regional differences, without attributing small raw errors to uniformly similar pathology patches.

Table 16 shows lower MSE and higher correlation of pairwise distances after distillation for both prediction forms in every dataset average. Their spread changes in opposite directions. Direct variance increases and residual variance decreases, while trained direct predictions remain less variable than teacher means and trained residual predictions remain more variable. Smaller direct spread relative to the teacher does not imply contraction from initialization. The two forms exhibit distinct learning behavior despite sharing a target and unrestricted optimum.

Zero final-layer initialization provides a distinct reference. The initial direct output is identically zero, whereas the initial residual prediction equals the native feature exactly. Training lowers direct MSE in eight dataset means, with higher errors on Private-CRC and PANDA. Residual MSE falls in every dataset mean. Table 17 shows that the lower direct and higher residual variance relative to the teacher persist under this initialization. The distinction prevents improvement over a random initial network from being mistaken for improvement over every reference.

The native feature provides an unchanged reference in the same coordinates. Trained direct prediction improves MSE and correlation of pairwise distances relative to native input in every dataset average under either initialization. Residual prediction lowers MSE throughout with zero initialization and in all datasets except Private-CRC with random initialization. These comparisons establish improvements relative to native input rather than superiority to every predictor. Table 14 retains the constant alongside all four trained predictors.

Under Proposition 1, reduced variation relative to teacher targets is compatible with conditional mean prediction but does not distinguish conditional target variation from estimation error. The direct outputs retain substantial variation and positive correlations of pairwise distances, while residual outputs remain more variable than the targets. Together with the permutation controls in Table 15, these measurements establish dependence on the input rather than a common constant prediction. Theorem 1 asks a different question about information relevant to the task. It does not identify a feature reconstruction error with representation loss or teacher access. Section 5.3 evaluates the downstream benefit separately.

This feature analysis uses CPU float32 inference. The raw reconstruction evaluation above uses GPU bfloat16 autocast, so its numerical values remain a separate evaluation. All measurements describe the saved predictors on sampled fields. The patch encoders remain frozen.

Table 16: Feature matching before and after distillation in the original 1,536 coordinates. Both predictors use random final-layer initialization. Native $5\times$ is the reference for MSE and correlation of pairwise distances. Variance ratios are descriptive, with teacher variation equal to one.

		MSE ↓			
	Native $5\times$	Direct		Residual	
Dataset	x_i	Initial	Trained	Initial	Trained
CAM16	0.1759 ± 0.0103	0.1206 ± 0.0108	$\mathbf{0.0686} \pm 0.0066$	0.1998 ± 0.0116	0.0972 ± 0.0076
CAM17	0.1939 ± 0.0044	0.1231 ± 0.0032	$\mathbf{0.0801} \pm 0.0053$	0.2183 ± 0.0045	0.1100 ± 0.0045
Private-CRC	0.1876 ± 0.0162	0.1574 ± 0.0042	$\mathbf{0.1551} \pm 0.0085$	0.2322 ± 0.0128	0.2017 ± 0.0235
PANDA	0.1554 ± 0.0067	0.1183 ± 0.0069	$\mathbf{0.1088} \pm 0.0112$	0.1855 ± 0.0056	0.1364 ± 0.0106
BRACS	0.1882 ± 0.0094	0.1412 ± 0.0096	$\mathbf{0.0941} \pm 0.0089$	0.2111 ± 0.0111	0.1174 ± 0.0074
KIRC	0.1697 ± 0.0020	0.1693 ± 0.0079	$\mathbf{0.1132} \pm 0.0096$	0.1992 ± 0.0033	0.1263 ± 0.0104
KIRP	0.1549 ± 0.0074	0.1486 ± 0.0082	$\mathbf{0.1023} \pm 0.0068$	0.1806 ± 0.0076	0.1148 ± 0.0099
LUAD	0.2074 ± 0.0142	0.1431 ± 0.0134	$\mathbf{0.1041} \pm 0.0136$	0.2398 ± 0.0164	0.1504 ± 0.0182
STAD	0.1828 ± 0.0083	0.1279 ± 0.0104	$\mathbf{0.0902} \pm 0.0131$	0.2047 ± 0.0095	0.1144 ± 0.0129
UCEC	0.1857 ± 0.0093	0.1527 ± 0.0136	$\mathbf{0.1294} \pm 0.0155$	0.2136 ± 0.0110	0.1386 ± 0.0080
		Centered variance ratio			
	Native $5\times$	Direct		Residual	
Dataset	x_i	Initial	Trained	Initial	Trained
CAM16	2.0689 ± 0.0950	0.2571 ± 0.0160	0.7269 ± 0.0635	2.3289 ± 0.1013	1.4388 ± 0.0469
CAM17	1.7259 ± 0.0926	0.2260 ± 0.0156	0.5245 ± 0.0261	1.9529 ± 0.1095	1.1984 ± 0.0582
Private-CRC	2.6330 ± 0.1512	0.4104 ± 0.0326	0.7964 ± 0.0392	3.0351 ± 0.1435	2.5611 ± 0.1402
PANDA	1.9803 ± 0.1472	0.2748 ± 0.0320	0.5430 ± 0.0589	2.2597 ± 0.1771	1.7437 ± 0.0949
BRACS	1.7239 ± 0.1369	0.2342 ± 0.0217	0.5302 ± 0.0356	1.9601 ± 0.1590	1.3829 ± 0.0799
KIRC	1.4746 ± 0.0529	0.2098 ± 0.0239	0.5387 ± 0.0494	1.6791 ± 0.0530	1.2758 ± 0.0789
KIRP	1.4454 ± 0.1226	0.1988 ± 0.0172	0.5414 ± 0.0352	1.6430 ± 0.1403	1.2545 ± 0.1020
LUAD	1.9351 ± 0.1296	0.2836 ± 0.0240	0.6153 ± 0.0695	2.2214 ± 0.1513	1.7643 ± 0.0939
STAD	1.6633 ± 0.1838	0.2057 ± 0.0280	0.4802 ± 0.0593	1.8749 ± 0.2114	1.3153 ± 0.1212
UCEC	1.5191 ± 0.0819	0.2119 ± 0.0167	0.5601 ± 0.0534	1.7307 ± 0.0967	1.2919 ± 0.0684
		Pearson correlation of pairwise distances ↑			
	Native $5\times$	Direct		Residual	
Dataset	x_i	Initial	Trained	Initial	Trained
CAM16	0.7720 ± 0.0334	0.7717 ± 0.0324	$\mathbf{0.8889} \pm 0.0248$	0.7786 ± 0.0325	0.8477 ± 0.0293
CAM17	0.5915 ± 0.0561	0.6104 ± 0.0336	$\mathbf{0.7572} \pm 0.0346$	0.5998 ± 0.0557	0.6435 ± 0.0532
Private-CRC	0.6855 ± 0.0434	0.6244 ± 0.0468	$\mathbf{0.7843} \pm 0.0175$	0.6870 ± 0.0472	0.6915 ± 0.0404
PANDA	0.5973 ± 0.0506	0.6179 ± 0.0388	$\mathbf{0.7768} \pm 0.0231$	0.6062 ± 0.0498	0.6569 ± 0.0266
BRACS	0.6206 ± 0.0424	0.6128 ± 0.0469	$\mathbf{0.7640} \pm 0.0375$	0.6283 ± 0.0419	0.6827 ± 0.0523
KIRC	0.7043 ± 0.0240	0.6879 ± 0.0407	$\mathbf{0.8140} \pm 0.0334$	0.7100 ± 0.0212	0.7685 ± 0.0378
KIRP	0.6957 ± 0.0312	0.6713 ± 0.0317	$\mathbf{0.7919} \pm 0.0294$	0.7009 ± 0.0314	0.7615 ± 0.0178
LUAD	0.6741 ± 0.0377	0.6489 ± 0.0351	$\mathbf{0.7889} \pm 0.0456$	0.6804 ± 0.0338	0.7068 ± 0.0172
STAD	0.6911 ± 0.0369	0.6142 ± 0.0590	0.7365 ± 0.0318	0.6918 ± 0.0371	$\mathbf{0.7446} \pm 0.0508$
UCEC	0.7244 ± 0.0301	0.6694 ± 0.0552	$\mathbf{0.7859} \pm 0.0676$	0.7259 ± 0.0386	0.7627 ± 0.0623

Table 17: **Variation and target geometry after training from a zero output layer.** Centered variance is divided by teacher variance, with one indicating equal total variation rather than an optimum. Pearson correlation compares pairwise distances between regions in the output and target spaces. Native $5\times$ is its reference. The five realizations use independent test fields from the excluded dataset, whereas the feature maps illustrate one slide and one realization.

	Centered variance ratio		Pearson correlation of pairwise distances ↑		
Dataset	Direct	Residual	Native $5\times$	Direct	Residual
CAM16	0.6997 ±0.0473	1.4150 ±0.0391	0.7720 ±0.0334	**0.8896** ±0.0140	0.8530 ±0.0245
CAM17	0.4978 ±0.0326	1.1662 ±0.0634	0.5915 ±0.0561	**0.7465** ±0.0289	0.6338 ±0.0564
Private-CRC	0.6253 ±0.0463	2.3830 ±0.1137	0.6855 ±0.0434	**0.8169** ±0.0282	0.6917 ±0.0423
PANDA	0.4742 ±0.0388	1.6700 ±0.1179	0.5973 ±0.0506	**0.7908** ±0.0294	0.6611 ±0.0397
BRACS	0.4876 ±0.0291	1.3241 ±0.0862	0.6206 ±0.0424	**0.7543** ±0.0411	0.6962 ±0.0524
KIRC	0.4748 ±0.0331	1.2216 ±0.0627	0.7043 ±0.0240	**0.8030** ±0.0316	0.7727 ±0.0355
KIRP	0.4956 ±0.0283	1.1852 ±0.0838	0.6957 ±0.0312	**0.7772** ±0.0448	0.7557 ±0.0219
LUAD	0.5454 ±0.0709	1.6263 ±0.1022	0.6741 ±0.0377	**0.7863** ±0.0389	0.7233 ±0.0286
STAD	0.4423 ±0.0677	1.2605 ±0.1082	0.6911 ±0.0369	0.7205 ±0.0439	**0.7589** ±0.0390
UCEC	0.4612 ±0.0404	1.2071 ±0.0569	0.7244 ±0.0301	0.7772 ±0.0645	**0.7799** ±0.0497

D.3 Teacher correspondence in a common feature map

The maps illustrate the regional mean prediction study from Section 4. The teacher target m_i is the arithmetic mean of sixteen aligned $20\times$ features, and the student input is the native $5\times$ feature together with its slide context. Direct diamonds show the target estimate $f_\theta(B)_i$. Residual diamonds show $x_i + r_\phi(B)_i$, where the correction $r_\phi(B)_i$ learns the teacher-minus-native difference $m_i - x_i$. Thus both plotted outputs estimate the teacher mean. Neither the correction alone nor a concatenation of native and output features is plotted. The networks use trained seed-42 parameters with zero final-layer initialization. This condition matches Table 3. It is not selected because zero initialization performs best on every displayed slide. The patch encoders remain frozen.

Reference construction and sampling. Within each dataset, we first sample 1,000 complete test regions by cycling through randomly ordered groups and drawing regions without replacement within each group. The seed combines 20260923, the dataset index and 42. The same identities supply native features, sixteen individual teacher features and the regional teacher means. Each of these three populations is clustered separately in the original 1,536 coordinates using Euclidean K-means. Five initializations, at most 200 iterations and tolerance 10^{-4} are used. Candidate cluster counts range from two to eight, and the largest silhouette score on a fixed subset of at most 768 vectors determines the count. These partitions organize the display and do not identify tissue classes.

Each population contributes 800 reference landmarks, distributed across its clusters in proportion to their size and including an actual observation nearest each cluster center. A common 50-component principal component analysis (PCA) transform reduces the 2,400 landmarks without whitening. A t-SNE map then uses Euclidean affinities, perplexity 40, seed 42, PCA initialization, 250 early iterations with exaggeration 12 and 500 subsequent iterations with exaggeration one (van der Maaten & Hinton, 2008). The remaining vectors are placed against these landmarks with openTSNE (Poličar et al., 2024), using median initialization from 25 neighbors, perplexity 40, learning rate 0.1, 250 iterations and maximum gradient norm 0.25. Prediction outputs do not fit the clusters, PCA or reference map. Reference and added coordinates share the same centered coordinate frame.

The display expands these reference populations to every cached region with sixteen aligned teacher features across all dataset partitions. This gives 1,007,109 native regions, 16,113,744 individual

teacher features and 1,007,109 regional means. The earlier cluster centers and reference map remain unchanged. Contours summarize every mapped reference with a square-bin histogram using 256 bins along the longest coordinate span and Gaussian smoothing of two bins. Each contour encloses approximately 75% of its cluster's smoothed density. Hollow markers identify actual observations nearest the cluster centers, with size reflecting their population share. Contours describe the reference distribution rather than uncertainty or anatomical boundaries.

For each dataset, one test slide is selected uniformly from sorted eligible slide identities using seed 20260923 combined with the dataset index. Every cached complete region of that slide is displayed for both prediction forms, with no further subsampling. Cached regions can exceed the 128 parents selected for the complete-field task study in Appendix B, because these illustrations use all eligible cached regions rather than that task selection. Trained predictions use the saved context from the whole native slide bag. Purple and red lines connect each native feature to its direct and residual outputs, respectively, while green lines connect it to its teacher mean. These connections identify the same physical region rather than successive training states. Cluster numbers are omitted because they do not identify tissue classes. All points and their coordinates remain unchanged. Figure 4 presents CAM16, CAM17, KIRC and KIRP in a compact grid. The examples contrast overlap with teacher neighborhoods, incomplete coverage of a teacher group and differences between direct and residual outputs. These dataset choices are editorial, while slide selection within each remains random. Each main-text panel overlays both prediction forms with one shared legend. The other six datasets appear below with separate direct and residual panels. The saved coordinates, points and connections remain unchanged. Display limits include every visible contour and observation, without reserving space for initial outputs that are not plotted. Equal axis scaling preserves the geometry within each map. Quantitative conclusions use the independent evaluations across all ten datasets.

Reconstruction accuracy and regional variation. Table 18 computes MSE on precisely the regions shown, averaging equally over regions and feature coordinates. Both trained predictors reduce error relative to native features on each displayed slide, with direct prediction giving the smaller error. On the KIRC slide in Figure 4, direct outputs have 0.5284 times the teacher's centered variance, compared with 1.5041 for residual outputs. Their correlations with teacher pairwise distances are 0.8816 and 0.9215, respectively. These calculations use the original 1,536 coordinates and equal region weights, with the definitions in Equation 25. Thus the wider residual distribution on this slide accompanies better agreement in relative distances but slightly higher average reconstruction error across the corresponding targets. These criteria describe different aspects of prediction.

The independent five-realization evaluation in Table 17 establishes the broader pattern. Every dataset average has direct variation below the teacher reference and residual variation above it. Both forms preserve regional differences, as their positive distance correlations and the permutation results in Table 15 show. Direct prediction's smaller spread is not a constant collapse. Nor is it a decrease from initialization, since training expands the initially zero direct output into an input-dependent representation. Distance correlations do not have a universal direct or residual ordering. For example, the KIRC dataset average favors direct prediction even though the displayed slide favors residual prediction. The illustrative slide and the independent evaluation fields answer different sampling questions.

What the remaining maps show. The remaining six maps extend the examples in Figure 4. PANDA shows a compact direct group while some residual outputs remain near native features. UCEC shows residual predictions occupying teacher neighborhoods that direct outputs cover less extensively. BRACS, Private-CRC, LUAD and STAD show different degrees of overlap between outputs and teacher references. Across all ten displayed slides, direct prediction has smaller variation and lower MSE than residual prediction. Residual variation is not always larger than teacher variation on these individual slides. The different visual patterns do not establish a common number of recovered clusters or a classification-versus-survival distinction.

Dataset contours describe a broader population than the teacher targets of the displayed slide. Their overlap with individual $20\times$ contours does not mean that the displayed means cover every teacher cluster or preserve every individual vector. Moreover, openTSNE positions each added output against reference landmarks without optimizing interactions among the added outputs. Apparent concentration is not a calibrated measure of variance or density in the original space. Connections share the native feature as their origin and identify its corresponding output and teacher mean.

Original-coordinate measurements determine reconstruction accuracy and variation. Native and teacher extraction histories also differ, so separation of their reference distributions cannot be attributed solely to magnification from this comparison.

Two kinds of averaging and the predictor bottleneck. Lemma 1 concerns averaging the sixteen observed teacher vectors within one region. Under uniform squared error, their variation around that regional mean contributes no predictor gradient for one shared output. Proposition 1 concerns a different average, the possible regional targets given the low magnification input. If that input does not determine the target, even the optimal deterministic predictor averages over the remaining possibilities. It need not reproduce the full distribution of teacher observations. These distinctions explain why overlap with individual teacher features is not evidence of recovering their within-region detail, and why less variable predictions can still minimize squared error.

The implemented heads also impose a concrete architectural restriction. Each network has a 256-unit GELU hidden layer followed by an affine output in 1,536 coordinates, with no final activation. For a given trained network, all direct outputs lie in an affine subspace of dimension at most 256. The residual network adds its correction to the native vector, allowing variation outside that correction subspace to pass through the skip connection. This distinction is consistent with their different spread despite the common unrestricted optimum in Proposition 1. It does not establish whether the main reconstruction difficulty comes from architecture, optimization or conditional uncertainty. In particular, the direct predictor's more limited output subspace is not evidence that the missing variation is unavailable in its input.

From reconstruction to task benefit. Theorem 1 distinguishes a teacher observation from a deterministic transformation of B. Its teacher access term is positive only when the teacher contributes information that reduces the best attainable task loss beyond what B permits. Feature uncertainty need not have that task consequence. A predictor can fail to match teacher features because of its finite approximation, and some unmatched feature variation may be irrelevant to Y. Neither a sparse set of mapped neighborhoods nor a nonzero MSE identifies the teacher access term.

The standalone task results in Table 23 show that lower reconstruction error does not consistently determine which prediction form performs better. KIRC with zero initialization favors direct prediction in mean C-index, despite its smaller spread, and the ordering reverses under random initialization. Wider residual coverage is therefore not a general explanation for better task performance. The native retention study in Appendix E.5 addresses a distinct consequence. Retaining the complete B removes representation loss, whereas replacing it by either output can change both representation loss and model excess. A residual sum alone does not guarantee recovery of B. Together, the maps and task controls motivate preserving useful regional distinctions while testing their task value separately from average reconstruction accuracy.

Table 18: **Reconstruction error for the exact regions in the feature maps.** Native features are the reference. Each entry is MSE over one displayed test slide and one trained model with zero final-layer initialization. Errors use the original 1,536 coordinates and equal weight per region. These are individual illustrative cases, without aggregation across training seeds. Lower is better.

Dataset	Regions	Native x_i	Direct $f_\theta(B)_i$	Residual $x_i + r_\phi(B)_i$
CAM16	128	0.1571	**0.0516**	0.0821
CAM17	135	0.2518	**0.1360**	0.1641
Private-CRC	149	0.2346	**0.1278**	0.1978
PANDA	52	0.1463	**0.1044**	0.1232
BRACS	134	0.1429	**0.0761**	0.0889
KIRC	128	0.1697	**0.0902**	0.0942
KIRP	131	0.1732	**0.0881**	0.1020
LUAD	128	0.2139	**0.0901**	0.1250
STAD	154	0.1652	**0.0822**	0.1100
UCEC	160	0.1747	**0.1040**	0.1144

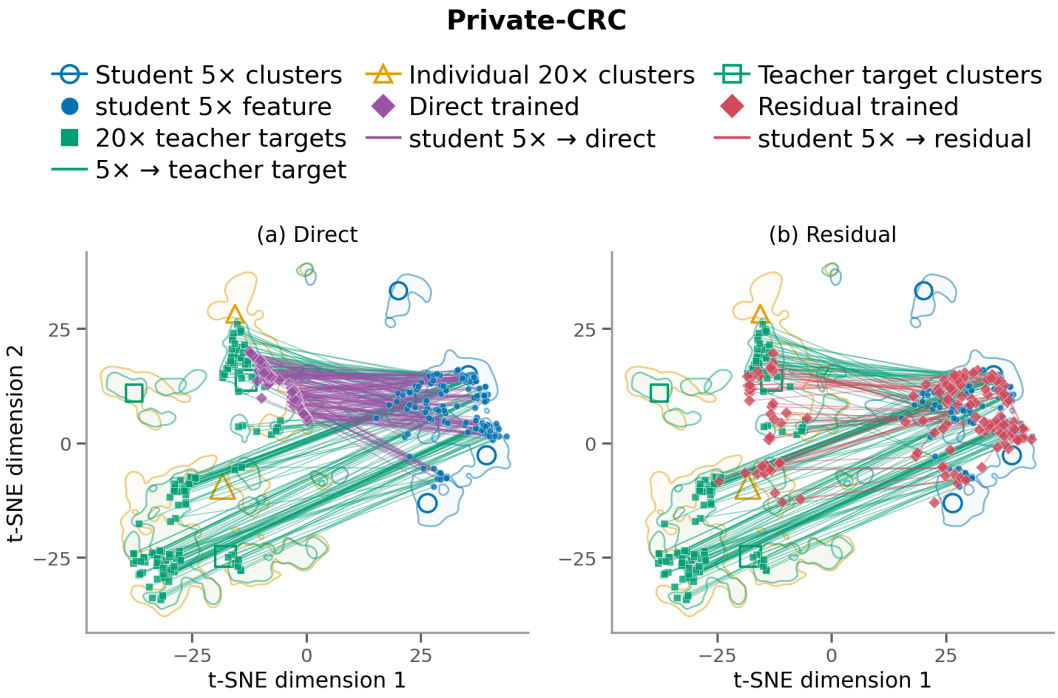


Figure 9: **Teacher correspondence in Private-CRC.** Direct and residual predictions for all 149 cached regions of one test slide. Some residual outputs remain near the native reference distribution while their corresponding teacher targets occupy different neighborhoods. Colors and connections follow Figure 4. Table 18 reports errors for these exact regions.

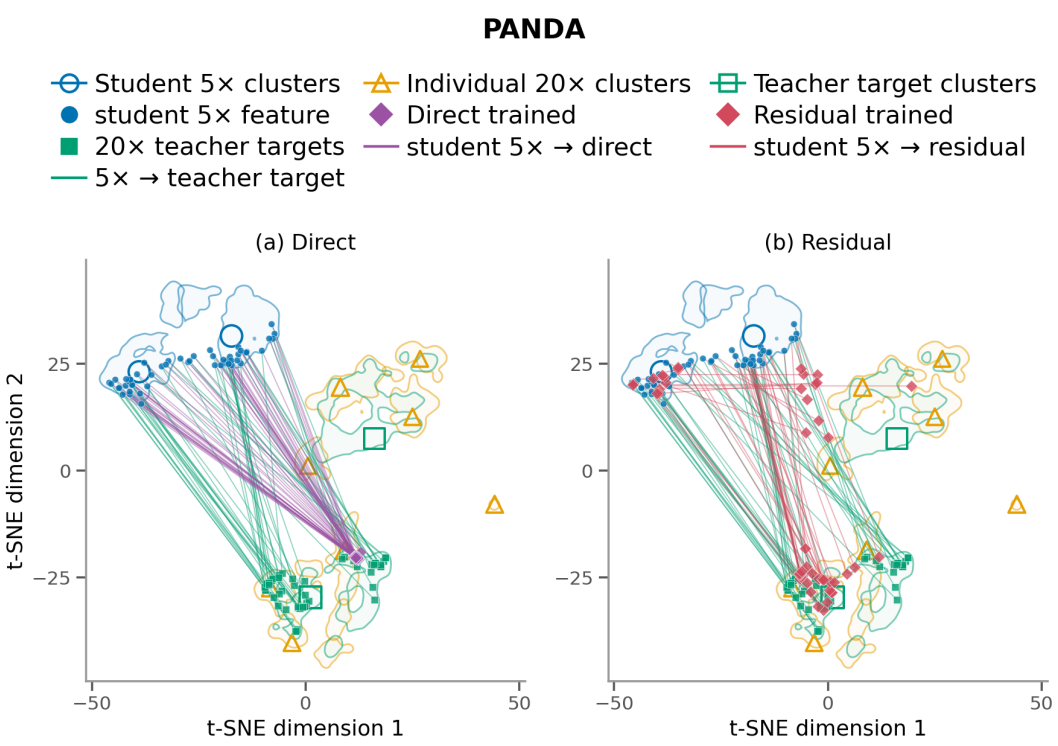


Figure 10: **Teacher correspondence in PANDA.** Direct and residual predictions for all 52 cached regions of one test slide. Direct outputs form a compact group near teacher means, while residual outputs also extend toward native neighborhoods. Colors and connections follow Figure 4. Table 18 reports errors for these exact regions.

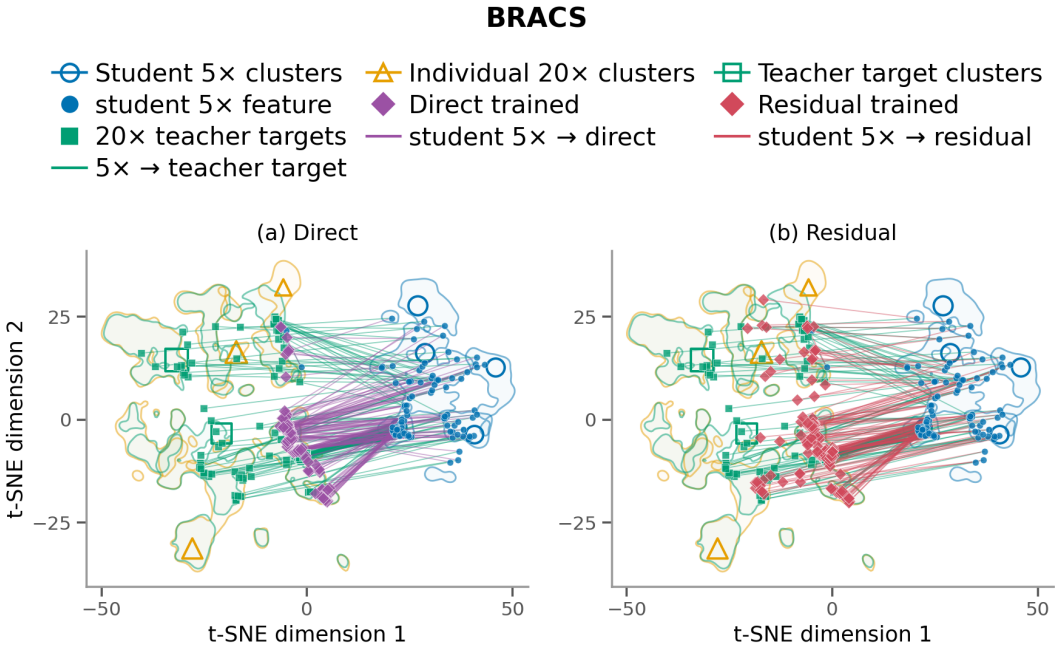


Figure 11: **Teacher correspondence in BRACS.** Direct and residual predictions for all 134 cached regions of one test slide. Direct outputs concentrate near central teacher neighborhoods, while residual outputs extend across more of the displayed target distribution. Colors and connections follow Figure 4. Table 18 reports errors for these exact regions.

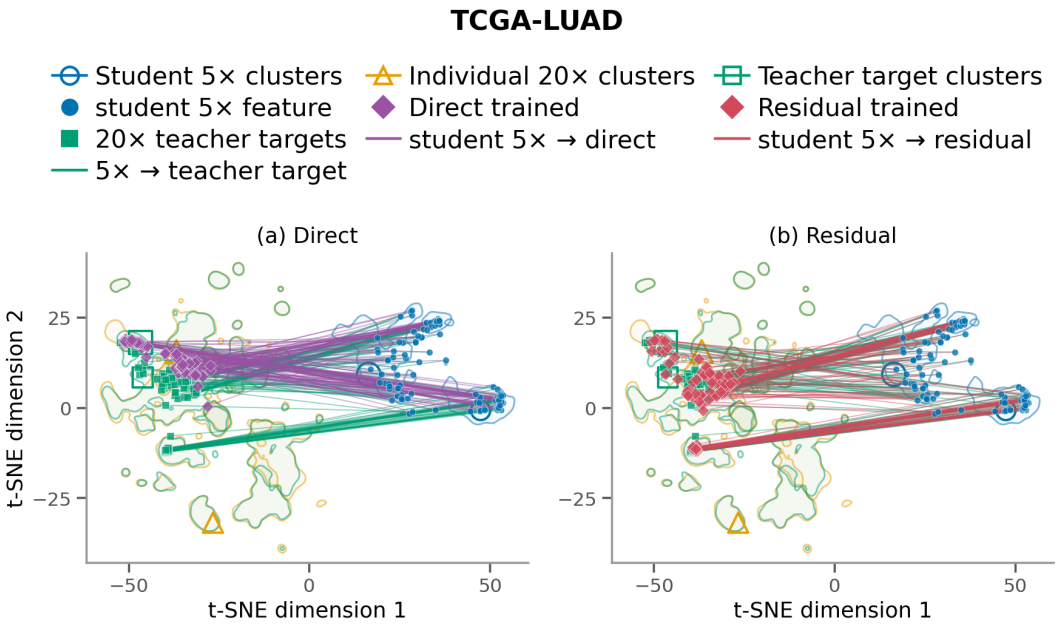


Figure 12: **Teacher correspondence in TCGA-LUAD.** Direct and residual predictions for all 128 cached regions of one test slide. Both prediction forms overlap teacher neighborhoods, but the residual outputs show a broader spread in this view. Colors and connections follow Figure 4. Table 18 reports errors for these exact regions.

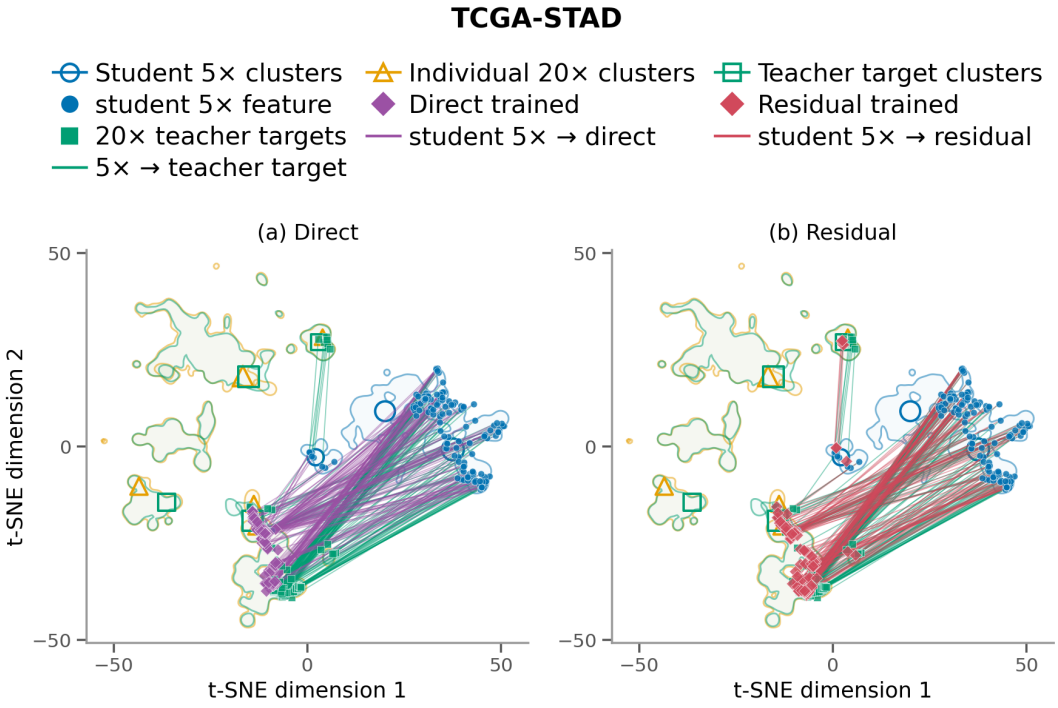


Figure 13: **Teacher correspondence in TCGA-STAD.** Direct and residual predictions for all 154 cached regions of one test slide. Both prediction forms approach the lower teacher neighborhoods, while direct outputs retain less variation in the original coordinates. Colors and connections follow Figure 4. Table 18 reports errors for these exact regions.

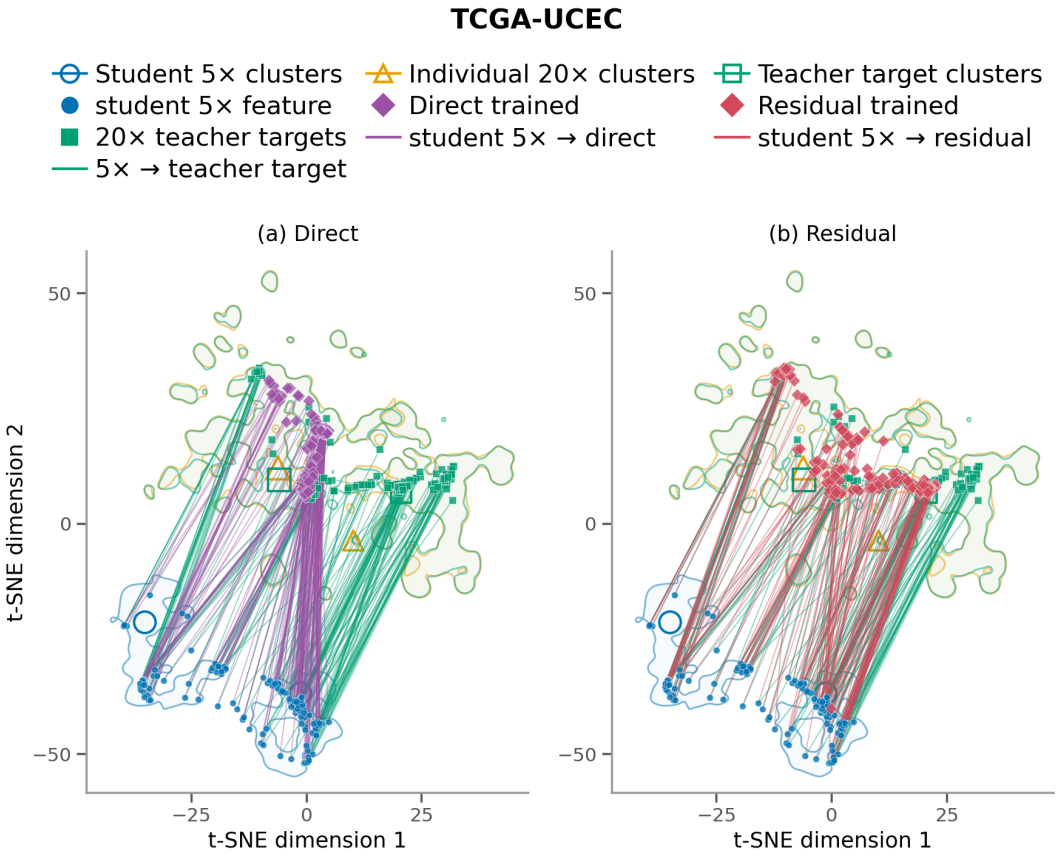


Figure 14: **Teacher correspondence in TCGA-UCEC.** Direct and residual predictions for all 160 cached regions of one test slide. Residual outputs extend across the right teacher neighborhoods more broadly than direct outputs, yet direct prediction has lower target MSE. Colors and connections follow Figure 4. Table 18 reports errors for these exact regions.

D.4 Predicting projected mean and spatial components

The next diagnostic asks which target components account for prediction error. It uses a different distillation partition and head from the raw analysis above. Its native input $b = [x; u] \in \mathbb{R}^{2304}$ concatenates the 1,536-coordinate UNI2-h feature and 768-coordinate DINOv3 feature from one $5\times$ region. The target has four blocks of 192 coordinates. A common MLP predicts all 768 outputs,

$$f_\theta(b) = W_2\,\mathrm{GELU}(W_1 b + a_1) + a_2, \qquad W_1 \in \mathbb{R}^{512\times 2304}, \quad W_2 \in \mathbb{R}^{768\times 512}. \tag{28}$$

The parameters θ comprise both weight matrices and their biases. GELU applies $\mathrm{GELU}(z) = z\Phi(z)$ coordinatewise, where Φ is the standard normal cumulative distribution function (Hendrycks & Gimpel, 2016). Squared error across the complete target trains all four components jointly.

Only base training groups determine the transformations. Their moments scale the input, while moments of projected native $5\times$ UNI2-h features normalize the targets. We center the mean block, use zero offsets for spatial blocks and apply the same 192 scales to all four. This preserves their different variances rather than giving every block unit variance.

Each distillation dataset contributes eight base training groups, eight correction groups, four validation groups for checkpoint selection and four independent evaluation groups. Eight diagnostic groups from the excluded dataset evaluate generalization without influencing fitting or selection. We sample at most 32 aligned fields per group and weight datasets, groups within datasets and fields within groups equally. Mean and spatial errors use identical fields. These samples differ from the task test fields above and from full downstream bags. Table 19 evaluates the base predictor before correction.

Training uses at most 4,096 AdamW updates with learning rate .001, weight decay .0001, batch size 1,024 and gradient clipping at 1. Total selection MSE chooses update 32 in thirty panels and update 128 in twenty. The reference always returns the target average from base training groups and uses the same coordinates, fields and weights as the network. Since it predicts no regional differences, improving on this constant tests predictive value beyond a common target centre.

D.5 Reconstruction errors by component

Prediction improves the regional mean much more consistently than the spatial components. In Table 19, the mean head beats the constant on every independent evaluation within the distillation datasets and in eight excluded-dataset averages. Every nonconstant component instead has higher error than the constant in every realization on both populations. This separates useful mean prediction from a persistent spatial deficit in the fitted network. Under Proposition 1, these observed errors still combine conditional target variation with estimation error, so the deficit does not identify an irreducible limit.

The four target components have equal width, so total MSE averages their four errors. Table 19 shows that the constant makes much more error per coordinate on the mean than on the spatial components. A shared objective can consequently improve through mean prediction while spatial error remains high relative to its own reference. This identifies a reason to report the components separately, without attributing the spatial deficit to loss weighting.

D.6 Independent neural error correction

We test whether an additional network can correct errors left by the selected base predictor. Let f_A be the frozen head fitted to distillation training groups A. A second MLP r_B learns from disjoint groups B using the same input, 512 hidden units, GELU and 768 outputs. A zero final layer initializes the corrected prediction $q = f_A + r_B$ at the original function f_A. This makes any subsequent change attributable to training the correction.

The same selection groups choose both checkpoints by total MSE. Independent evaluation within the distillation datasets and on excluded-dataset groups selects neither checkpoint. Correction introduces both extra capacity and new training groups, so its effect combines those changes. Table 20 separates the resulting mean and spatial errors. This diagnostic stops at feature prediction and fits no downstream task model.

For each excluded dataset, correction lowers average total MSE on independent groups from the distillation datasets, mainly by improving the regional mean (Table 20). The mean accounts for 95.7%

Table 19: MSE for regional mean and spatial components. The constant is the reference within each component. Panel A evaluates independent groups from the distillation datasets, and Panel B evaluates the excluded dataset. Neither population selects predictors.

A. Independent evaluation within distillation datasets

	Regional mean		Spatial components	
Excluded dataset	Constant	Network	Constant	Network
CAM16	0.6361±0.0305	**0.3648**±0.0173	**0.0524**±0.0022	0.0697±0.0031
CAM17	0.6665±0.0335	**0.3702**±0.0219	**0.0527**±0.0025	0.0676±0.0021
Private-CRC	0.6851±0.0567	**0.4007**±0.0191	**0.0527**±0.0010	0.0698±0.0019
PANDA	0.6592±0.0310	**0.3775**±0.0188	**0.0487**±0.0011	0.0662±0.0023
BRACS	0.6295±0.0249	**0.3653**±0.0098	**0.0533**±0.0012	0.0691±0.0014
KIRC	0.6419±0.0226	**0.3671**±0.0165	**0.0552**±0.0016	0.0718±0.0026
KIRP	0.6320±0.0533	**0.3584**±0.0220	**0.0542**±0.0020	0.0705±0.0015
LUAD	0.6612±0.0309	**0.3552**±0.0162	**0.0564**±0.0017	0.0730±0.0013
STAD	0.6681±0.0490	**0.3671**±0.0204	**0.0542**±0.0017	0.0708±0.0014
UCEC	0.6269±0.0235	**0.3499**±0.0109	**0.0543**±0.0008	0.0706±0.0019

B. Excluded dataset evaluation

	Regional mean		Spatial components	
Excluded dataset	Constant	Network	Constant	Network
CAM16	0.6106±0.0182	**0.4501**±0.0418	**0.0464**±0.0015	0.0684±0.0036
CAM17	0.5618±0.0415	**0.4599**±0.0539	**0.0550**±0.0039	0.0724±0.0038
Private-CRC	**0.7193**±0.0842	0.7721±0.0848	**0.0845**±0.0096	0.1162±0.0075
PANDA	**0.5319**±0.0264	0.5602±0.0312	**0.0845**±0.0049	0.1106±0.0069
BRACS	0.7437±0.1001	**0.6339**±0.0924	**0.0454**±0.0056	0.0646±0.0081
KIRC	0.8488±0.0866	**0.6799**±0.1207	**0.0421**±0.0022	0.0657±0.0053
KIRP	0.7377±0.0775	**0.6091**±0.0801	**0.0393**±0.0016	0.0621±0.0045
LUAD	0.6055±0.0427	**0.5535**±0.0466	**0.0456**±0.0023	0.0738±0.0065
STAD	0.5791±0.0191	**0.5166**±0.0253	**0.0430**±0.0021	0.0652±0.0040
UCEC	0.9264±0.1741	**0.8966**±0.1914	**0.0490**±0.0017	0.0767±0.0096

Table 20: Absolute component MSE before and after neural error correction. The base predictor is the reference within each component. Spatial MSE averages three equally sized components. Neither evaluation population determines the selected predictors.

Independent evaluation within distillation datasets

	Regional mean		Spatial components	
Excluded dataset	Base	Corrected	Base	Corrected
CAM16	0.3648±0.0173	**0.3401**±0.0199	0.0697±0.0031	**0.0696**±0.0034
CAM17	0.3702±0.0219	**0.3361**±0.0156	0.0676±0.0021	**0.0665**±0.0019
Private-CRC	0.4007±0.0191	**0.3748**±0.0152	0.0698±0.0019	**0.0694**±0.0023
PANDA	0.3775±0.0188	**0.3532**±0.0097	0.0662±0.0023	**0.0659**±0.0031
BRACS	0.3653±0.0098	**0.3455**±0.0167	0.0691±0.0014	**0.0686**±0.0016
KIRC	0.3671±0.0165	**0.3526**±0.0244	0.0718±0.0026	**0.0716**±0.0028
KIRP	0.3584±0.0220	**0.3328**±0.0158	0.0705±0.0015	**0.0704**±0.0016
LUAD	0.3552±0.0162	**0.3342**±0.0241	0.0730±0.0013	**0.0727**±0.0021
STAD	0.3671±0.0204	**0.3463**±0.0124	**0.0708**±0.0014	0.0709±0.0017
UCEC	0.3499±0.0109	**0.3255**±0.0136	0.0706±0.0019	**0.0699**±0.0025

Excluded dataset evaluation

	Regional mean		Spatial components	
Excluded dataset	Base	Corrected	Base	Corrected
CAM16	**0.4501**±0.0418	0.4592±0.0420	**0.0684**±0.0036	0.0690±0.0036
CAM17	0.4599±0.0539	**0.4528**±0.0478	**0.0724**±0.0038	0.0726±0.0038
Private-CRC	**0.7721**±0.0848	0.7773±0.0735	**0.1162**±0.0075	0.1168±0.0069
PANDA	0.5602±0.0312	**0.5584**±0.0301	**0.1106**±0.0069	0.1112±0.0069
BRACS	0.6339±0.0924	**0.5908**±0.0751	**0.0646**±0.0081	0.0652±0.0079
KIRC	0.6799±0.1207	**0.6558**±0.1257	**0.0657**±0.0053	0.0669±0.0053
KIRP	0.6091±0.0801	**0.5872**±0.0958	**0.0621**±0.0045	0.0629±0.0041
LUAD	0.5535±0.0466	**0.5405**±0.0434	**0.0738**±0.0065	0.0746±0.0064
STAD	0.5166±0.0253	**0.4971**±0.0185	**0.0652**±0.0040	0.0655±0.0045
UCEC	0.8966±0.1914	**0.8963**±0.2074	**0.0767**±0.0096	0.0776±0.0090

Table 21: Task performance when a common native $5\times$ base receives zeros, the teacher mean or a prediction from global or spatial input. Zero augmentation is the reference. The teacher mean requires $20\times$ features during evaluation.

Dataset	Metric	$[B;0]$	$[B;M]$	Global input	Spatial input
CAM16	AUC	0.8455 ± 0.0485	$\mathbf{0.8935}\pm0.0635$	0.8729 ± 0.0384	0.8516 ± 0.0515
CAM17	F_1	0.5197 ± 0.0336	0.5158 ± 0.0700	0.5144 ± 0.0381	$\mathbf{0.5275}\pm0.0310$
Private-CRC	F_1	0.8305 ± 0.0276	$\mathbf{0.8378}\pm0.0267$	0.8277 ± 0.0228	0.8276 ± 0.0367
PANDA	κ	0.8386 ± 0.0067	$\mathbf{0.8601}\pm0.0137$	0.8387 ± 0.0124	0.8330 ± 0.0062
BRACS	F_1	$\mathbf{0.4211}\pm0.0710$	0.4168 ± 0.0542	0.4178 ± 0.0575	0.4125 ± 0.0856
KIRC	C-index	0.6823 ± 0.0548	0.6879 ± 0.0586	0.6876 ± 0.0718	$\mathbf{0.7186}\pm0.0538$
KIRP	C-index	0.7352 ± 0.1486	0.7850 ± 0.1042	$\mathbf{0.8278}\pm0.0529$	0.8155 ± 0.0828
LUAD	C-index	0.5381 ± 0.0614	0.5551 ± 0.0459	$\mathbf{0.5678}\pm0.0224$	0.5343 ± 0.0527
STAD	C-index	0.5104 ± 0.0188	$\mathbf{0.5553}\pm0.0669$	0.5335 ± 0.0561	0.4940 ± 0.0619
UCEC	C-index	0.6499 ± 0.1086	0.7039 ± 0.1145	$\mathbf{0.7105}\pm0.0840$	0.7048 ± 0.0618

of the pooled error reduction. We compute that fraction by averaging each block's reduction over the fifty realizations, then dividing the mean reduction by the sum across all four blocks. It is not an average of realization-level fractions. Correction demonstrates remaining learnable error in the mean while producing only a small reduction in spatial error.

Generalization to excluded datasets is less consistent. Correction lowers total error in six dataset averages and improves the mean on average, with adverse mean effects on CAM16 and Private-CRC. Spatial error rises in every dataset average. Each nonconstant block remains worse than its constant reference in every realization on both evaluation populations, before and after correction. Correction improves regional mean prediction while leaving the spatial deficit unresolved.

D.7 Spatial access to the low magnification image

To test whether local image features improve prediction, we give a predictor either repeated global features or a spatial grid. Both inputs contain sixteen 768-coordinate DINOv3 vectors with two position coordinates each. The global version repeats the normalized class token. The spatial version averages the same encoder's normalized 16×16 patch-token map into a 4×4 grid. Both come from one image field. Flattening produces 12,320 inputs for a $12320 \to 512 \to 768$ GELU network. The target is a 768-coordinate projection of the regional teacher mean. It contains neither individual teacher vectors nor nonconstant spatial components. The study comprises 100 distillation fits and 200 TransMIL fits across ten datasets and five realizations.

The predictors train separately with shared architecture, initialization, distillation training and validation groups, and normalization. Each uses 4,096 updates, batch size 1,024, learning rate .001, weight decay .0001 and gradient clipping at 1. Distillation validation selects checkpoints, including early evaluations at updates 0, 32, 128 and 256. For task evaluation, the 2,304-coordinate native base receives zeros, the teacher mean or a prediction, producing 3,072 coordinates. Every model uses the same complete fields, task initialization, splits and training procedure within a realization, with at most 128 regions per slide.

Local input changes reconstruction and task performance in different ways, as Figure 15 shows. It lowers average feature error on CAM16, CAM17, Private-CRC, PANDA and LUAD, whereas only CAM17 and KIRC gain on the average task score. KIRC gains despite worse reconstruction. These descriptive orderings motivate separate feature and task evaluations without establishing a causal relationship between them.

The task model retains global UNI2-h and DINOv3 features, but these do not necessarily recover the local grid seen by the spatial predictor. Complete-input recovery, required to cancel representation loss in Theorem 1, is therefore not established for this comparison. Appendix E.5 tests delivery separately by holding the prediction and its native input unchanged.

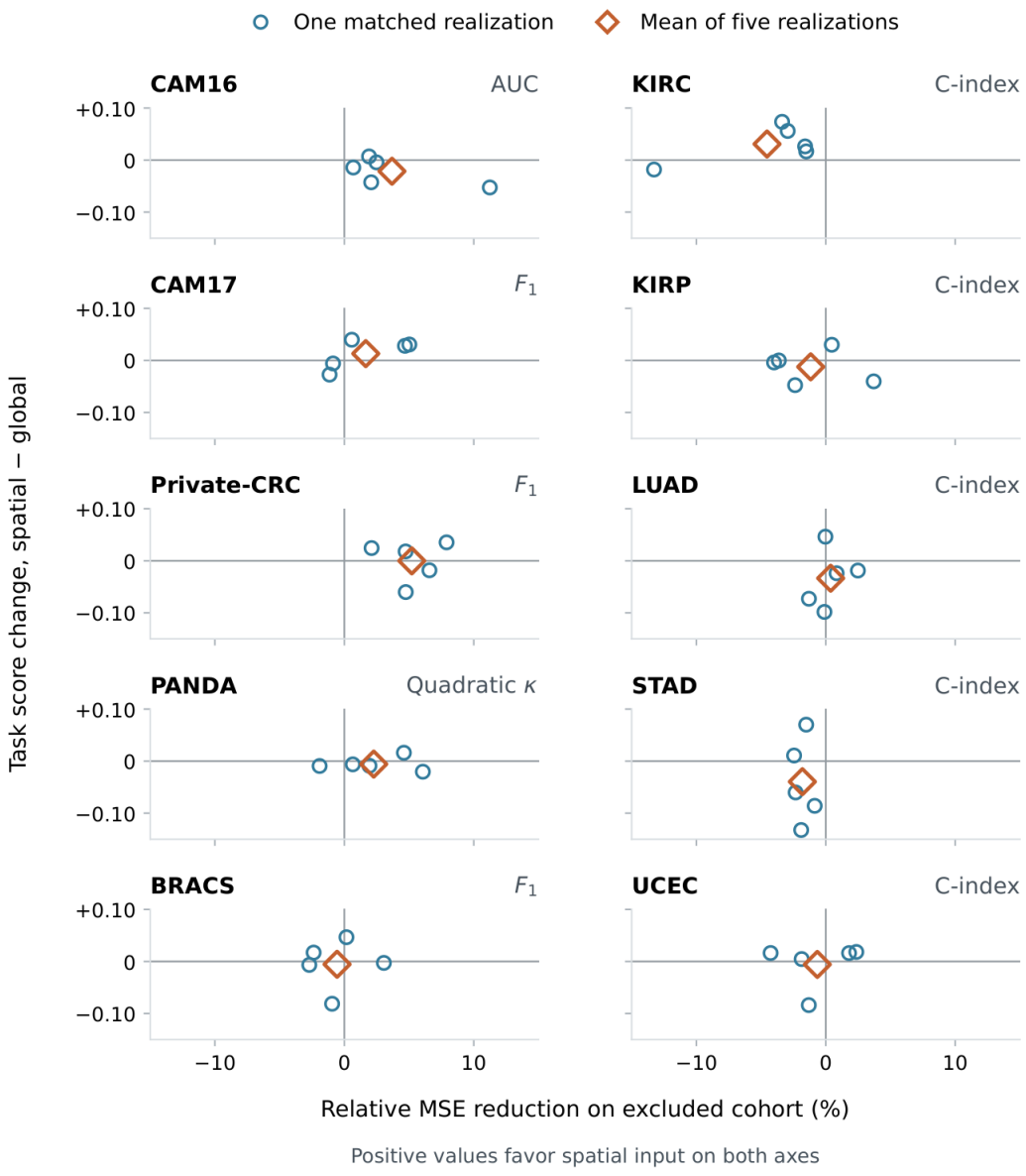


Figure 15: Spatial and global input for a common regional mean target. Horizontal position is $100(\mathrm{MSE}_{\mathrm{global}} - \mathrm{MSE}_{\mathrm{spatial}})/\mathrm{MSE}_{\mathrm{global}}$. Vertical position is the spatial minus global task score. Circles show individual realizations and diamonds show their mean. Positive values favor spatial input on each axis.

E Downstream Use and Attribution to Distillation

A useful teacher target does not ensure a useful prediction, and a useful prediction does not by itself establish a benefit from cross-resolution distillation. This appendix separates those questions with initial predictors, output removal and native retention. Each study compares task models on common tissue support and input width. The final study tests a different route to transfer by adapting the image encoder and deploying its features.

E.1 Learning the nonconstant spatial component

We ask whether Stage 1 training improves the spatial head after the regional mean is already learned. The study retaining all feature channels in Appendix C.6 compares the trained nonconstant head with its exact initialization while retaining the same trained mean. It also omits the spatial output while retaining that mean. Table 10 reports the absolute NLL for these three choices, using the same trained mean with zero spatial output as its reference. Table 10 adds their primary scores, native input with zero augmentation and observed teacher summaries. Zero spatial output retains the trained mean, whereas zero augmentation removes the entire summary. Every comparison uses the same complete fields and 8,448-coordinate input.

Learning the spatial head lowers average NLL on PANDA, BRACS, KIRC, KIRP, LUAD and UCEC, while CAM16, CAM17, Private-CRC and STAD have higher averages than the initial head. These effects vary across realizations. BRACS has lower NLL after learning in three realizations, whereas PANDA, KIRC and LUAD have lower NLL in all five. Classification NLL uses class probabilities, while survival NLL retains slide weights and training-derived bins. Each loss is interpreted on its own scale.

The comparison with the trained mean alone answers a different delivery question. CAM17 has lower mean NLL when spatial predictions are supplied despite higher NLL than the initial spatial head. STAD has higher mean NLL than both references. On PANDA, the mean NLL improvement over the trained mean coexists with improvement in only two realizations. KIRC and LUAD still have higher mean NLL than native input with zero augmentation. Learning a head, benefiting from its output beyond a predicted mean and improving beyond native input are therefore distinct results. The primary scores likewise show no common ordering across datasets.

E.2 General augmentation controls

To examine model excess, we test whether distillation makes an unchanged native input easier for the task model to use. The input B from Equation 1 is the bag $B = (b_i)_i$ of native vectors defined in Section 4. Every task input retains b_i. We append zeros, trained predictions q_i, teacher summaries t_i or the network's exact initial predictions q_i^I,

$$A_0 = ([b_i; 0])_i, \quad A_P = ([b_i; q_i])_i, \quad A_M = ([b_i; t_i])_i, \quad A_I = ([b_i; q_i^I])_i. \tag{29}$$

Zero padding keeps the task input width constant. The initial predictor controls for adding the network transformation before Stage 1 training. Both initial and trained predictions are functions of the retained input, so Equation 3 expresses their population loss difference as a difference in model excess. The observed teacher supplies a separate reference with access to (B, T).

E.3 Removing spatial predictions while retaining the same mean

To test delivery on unrestricted native bags, we reuse one saved predictor of the projected regional mean and three spatial components. Its full output has four blocks of 192 coordinates. The removal version keeps the mean unchanged and zeros the spatial blocks. Both append 768 coordinates to the native UNI2-h and DINOv3 base, giving a 3,072-coordinate input. A new task model trains for each version using common splits, initialization and settings. Every original $5\times$ row is retained, distinguishing this support from complete teacher fields and from the separate full-channel study.

Table 22 compares the complete trained output with the same trained mean alone, exact initial prediction and native input with zeros. Fifty additional task fits supply the mean-only condition across ten datasets and five realizations. Zeroing spatial components also lowers the added feature energy, so the intervention combines omission with that change.

Omitting the spatial blocks has mixed effects. PANDA improves over the full prediction in all five realizations, and its mean score also exceeds the zero and initial references. In contrast, CAM16 and LUAD remain below the native reference after omission. Removing a harmful component can improve an augmented model without establishing a general benefit from distillation.

Table 22: Task scores after omitting projected spatial predictions on unrestricted native bags. Every input retains the native UNI2-h and DINOv3 base and has 3,072 coordinates. Native input with zero augmentation is the reference. Higher is better.

Dataset / metric	Native 5× Zero augmentation	Native 5× Initial prediction	Native 5× Trained mean + spatial	Native 5× Same trained mean
CAM16 / AUC	**0.9770**±0.0129	0.9625±0.0308	0.9589±0.0293	0.9487±0.0315
CAM17 / F_1	0.6300±0.0294	0.6397±0.0262	**0.6462**±0.0445	0.6079±0.0719
Private-CRC / F_1	0.8244±0.0482	**0.8462**±0.0154	0.8426±0.0312	0.8238±0.0383
PANDA / κ	0.8425±0.0121	0.8489±0.0149	0.8430±0.0103	**0.8508**±0.0132
BRACS / F_1	0.4141±0.0340	0.4336±0.0714	0.4321±0.0780	**0.4570**±0.0587
KIRC / C-index	0.7017±0.0456	0.7104±0.0127	**0.7151**±0.0277	0.7026±0.0395
KIRP / C-index	0.8262±0.0836	0.7984±0.0342	0.7203±0.1028	**0.8267**±0.0343
LUAD / C-index	**0.5457**±0.0297	0.5099±0.0663	0.4866±0.0679	0.5004±0.0636
STAD / C-index	**0.5198**±0.0395	0.5004±0.0703	0.4975±0.0275	0.5030±0.0342
UCEC / C-index	**0.7027**±0.0652	0.6575±0.0832	0.6772±0.0691	0.6988±0.1115

E.4 Direct and residual prediction across ten datasets

To test whether reconstruction error ranks task representations, we compare direct and residual predictions of the same regional mean. This experiment trains task models separately for the two representations. A $3072 \to 256 \to 1536$ GELU predictor receives the standardized native $5\times$ vector x_i and the standardized mean c of native vectors across its slide. Direct output $q_i = f_i$ and residual output $q_i = x_i + f_i$ use common target scales $s_{m,d}$ to predict m_i,

$$\mathcal{L} = \mathbb{E}_{\text{train}} \frac{1}{1536} \sum_{d=1}^{1536} \left(\frac{q_{id} - m_{id}}{s_{m,d}} \right)^2 . \tag{30}$$

The expectation $\mathbb{E}_{\text{train}}$ averages over the distillation training population. Scales estimated from that population balance the coordinates during training. We reverse those transformations for reconstruction error in the original UNI2-h space. Figure 16 uses the historical evaluation on downstream training fields, while Table 3 uses independent test fields from Appendix D.1. For population e with its recorded group and field weights, error is

$$\text{MSE}_e(q, m) = \mathbb{E}_e \left[\frac{1}{1536} \sum_{d=1}^{1536} (q_{id} - m_{id})^2 \right] . \tag{31}$$

The historical diagnostic samples eight groups and 234 to 256 downstream training fields from each excluded dataset. Task performance is evaluated on separate test cases. The native reference sets $q_i = x_i$ on the same fields, measuring the discrepancy between native features and the teacher target. It equals the residual predictor at zero initialization before training, as confirmed by direct calculation of $\|m_i - x_i\|^2$. We count this reference once per dataset and realization.

Appendix D.1 uses the training target mean, estimated from the original distillation training sample, as a constant reference for the same selected heads on independent test fields. The native vector tests reduction of the discrepancy between magnifications, whereas the constant tests improvement over a common target centre. Because the populations differ, the two analyses retain separate errors and realization counts.

Each distillation dataset contributes at most sixteen training groups and eight validation groups, with 32 fields per group. The objective contains no cosine or native-retention term. AdamW uses batch size 1,024, learning rate .001, weight decay .0001, gradient clipping at 1 and 4,096 updates. Distillation validation chooses among checkpoints saved every 512 updates. All 200 predictors select the first candidate at update 512, so these results describe that selection procedure rather than an experiment on training duration.

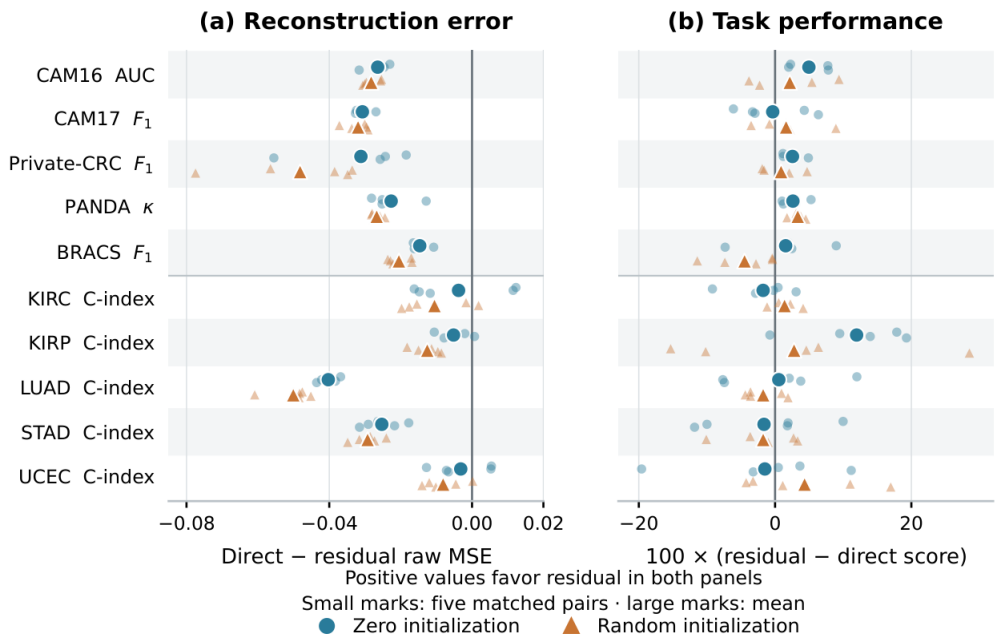


Figure 16: **Reconstruction error need not rank task performance.** Small marks show five realizations and large marks their means. Positive differences favor residual prediction in both panels. Reconstruction uses direct minus residual MSE on historical downstream training fields from datasets excluded from distillation. Task performance uses residual minus direct test score multiplied by 100 in each named metric. Table 23 reports absolute task scores.

We cross direct or residual output with random or zero initialization of the final layer. A zero layer starts direct output at zero and residual output at x_i. The native skip also changes the finite function class around the 256-coordinate bottleneck. Proposition 1 equates unrestricted optima but permits these differences in finite outputs and optimization. Across ten datasets and five realizations, each task model receives only the 1,536 predicted coordinates. Native features are not supplied as a separate block.

Direct prediction has lower mean reconstruction error in every dataset under both initializations, yet residual prediction often performs better downstream (Figure 16). Their preferences disagree in 35 of fifty realizations with zero initialization and 28 with random initialization. These results demonstrate that reconstruction error alone cannot rank the resulting task representations. The original-coordinate analysis in Appendix D.2 shows that the same prediction forms also differ in variation across regions. That difference is a candidate explanation to investigate, not an isolated cause of the task outcomes. Under Theorem 1, the standalone outputs can differ in both representation loss and model excess. Their reconstruction errors cannot separate these contributions.

Table 23 places both trained prediction forms beside native input. The best form varies across datasets and initialization. A higher mean than the other predictor does not ensure a higher mean than native input. This comparison uses standalone 1,536-coordinate predictions, while the following experiment changes the separate native block at a common width of 3,072.

E.5 Native retention with an unchanged teacher prediction

To isolate delivery from reconstruction, we reuse the direct and residual predictions above and change only whether the native vector is supplied alongside them. For each saved q_i, the retained version gives the task model $[x_i; q_i]$, while the zero native block version gives $[0_d; q_i]$. Both contain 3,072 coordinates in the same region order. The two task fits share splits, initial weights, training and selection. Their predictions and reconstruction errors are identical. Write $q_i^D = f_\theta(B)_i$ for direct prediction and $q_i^R = x_i + r_\phi(B)_i$ for residual prediction.

Table 23: Task scores for native $5\times$ features and standalone predictions of the regional $20\times$ mean. Each input has 1,536 coordinates. Zero and random specify final-layer initialization before distillation. Native $5\times$ input is the reference. Higher is better.

Dataset / metric	Native $5\times$	Direct Zero init.	Residual Zero init.	Direct Random init.	Residual Random init.
CAM16 / AUC	0.9494 ± 0.0198	0.9142 ± 0.0422	$\mathbf{0.9638}\pm0.0244$	0.9224 ± 0.0464	0.9440 ± 0.0417
CAM17 / F_1	0.6093 ± 0.0556	0.6116 ± 0.0857	0.6081 ± 0.0385	0.6069 ± 0.0683	$\mathbf{0.6231}\pm0.0494$
Private-CRC / F_1	0.8432 ± 0.0198	0.8249 ± 0.0179	$\mathbf{0.8504}\pm0.0142$	0.8412 ± 0.0258	0.8499 ± 0.0239
PANDA / κ	0.8496 ± 0.0089	0.8281 ± 0.0203	$\mathbf{0.8542}\pm0.0116$	0.8154 ± 0.0175	0.8486 ± 0.0104
BRACS / F_1	0.4170 ± 0.0527	0.4337 ± 0.0907	$\mathbf{0.4493}\pm0.0443$	0.4068 ± 0.0790	0.3620 ± 0.0708
KIRC / C-index	0.7007 ± 0.0279	$\mathbf{0.7082}\pm0.0250$	0.6906 ± 0.0407	0.6909 ± 0.0458	0.7047 ± 0.0423
KIRP / C-index	0.7707 ± 0.1060	0.6881 ± 0.0308	$\mathbf{0.8078}\pm0.1011$	0.7430 ± 0.1636	0.7709 ± 0.0964
LUAD / C-index	0.5327 ± 0.0682	0.5296 ± 0.0692	0.5351 ± 0.0300	$\mathbf{0.5636}\pm0.0464$	0.5461 ± 0.0373
STAD / C-index	0.5164 ± 0.0790	0.5192 ± 0.0732	0.5029 ± 0.0812	$\mathbf{0.5321}\pm0.0551$	0.5146 ± 0.0770
UCEC / C-index	$\mathbf{0.6756}\pm0.0830$	0.6576 ± 0.0973	0.6425 ± 0.1134	0.6193 ± 0.1241	0.6626 ± 0.1135

The study crosses prediction form, predictor final-layer initialization and native retention, producing eight conditions and 400 task fits across ten datasets and five realizations. Every predictor has undergone Stage 1 training. Zero and random describe how that training began. Fifty additional native-plus-zero references are reused across initialization panels at the same task width. They compare augmentation with the native base, while an exact initial predictor would be needed to attribute its effect to distillation.

Complete native input and the risk comparison. For this experiment, the deployment input B in Equation 1 is the complete bag of native UNI2-h vectors. The predictor uses each vector and the bag's arithmetic mean, without DINOv3. Taking the first block of $[x_i; q_i]$ recovers every native vector and hence the slide context. All 14,623 recorded bags have at most 6,488 rows, below the cap of 49,152, so no native row is lost through subsampling. Projection and attention occur later inside task model g.

Complete recovery makes representation loss zero for native retention in Theorem 1. Replacing the separate native block with zeros can leave some native information in q_i. A residual prediction, for example, contains x_i in the sum $x_i + r_i(B)$ even when that sum is not invertible. The loss difference between retained and zero native block conditions therefore combines any representation loss with their difference in model excess. Keeping q_i unchanged lets the experiment measure the practical effect of delivery without attributing it to better teacher reconstruction.

Task effects of retaining native features. Figure 17 reports changes within each realization for both primary scores and NLL, extending the NLL comparison in Figure 6. Tables 24 and 25 report absolute results. Native retention improves average primary scores more often for direct predictions than for residual predictions. Residual outputs already include a native skip, which makes the additional block a different intervention for the two forms. NLL favors retention across most conditions of both forms, but neither retention nor teacher prediction consistently improves on the matched native reference. The consequences of delivery depend on the task criterion and vary across the five joint distillation and task realizations.

NLL is the saved arithmetic mean of slide classification cross entropy or discrete survival loss. Survival uses the shared training-derived bins before case aggregation. Classification retains the trainer's probability convention, which differs from the saved case-analysis convention in Appendix C.5. We keep these loss magnitudes separate. The reference $[x_i; 0_d]$ shares support, input width, initial task weights, splits and training with all eight prediction conditions. It is a different task fit from the 1,536-coordinate native reference in Table 23.

A coordinate change that preserves information. An invertible coordinate change provides a separate check of information preservation. Starting from x and residual correction r, we form $u = (x + r)/\sqrt{2}$ and $v = (x - r)/\sqrt{2}$, with inverse $x = (u + v)/\sqrt{2}$ and $r = (u - v)/\sqrt{2}$. Rewriting a saved task model's first affine layer expresses the same function in these coordinates

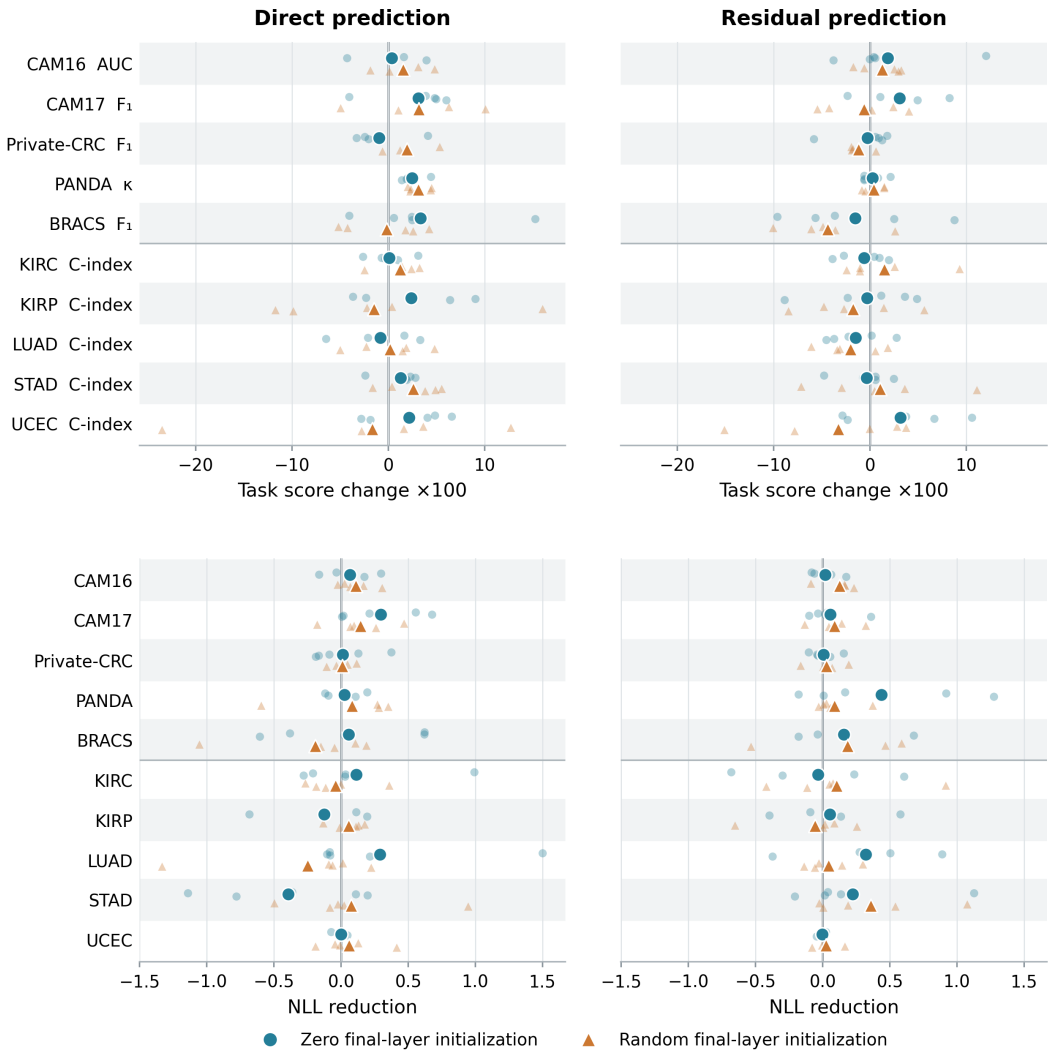


Figure 17: **Native retention changes task results for unchanged predictions.** Upper panels show 100 times the primary score for $[x_i; q_i]$ minus $[0; q_i]$. Lower panels show NLL for $[0; q_i]$ minus $[x_i; q_i]$. Positive values favor retention. Columns distinguish prediction forms, while circles and triangles distinguish zero and random final-layer initialization. All predictors are trained. Small marks show five corresponding task differences and large marks their means. Appendix E.5 reports absolute scores.

Table 24: Task scores with the same trained prediction and either a zero or retained native block. All inputs have 3,072 coordinates. The matched native $5\times$ reference $[x_i; 0_d]$ is reused across panels. Higher is better.

Dataset / metric	Native $5\times$ $[x_i; 0_d]$	Direct $[0_d; q_i^D]$	Direct + native $[x_i; q_i^D]$	Residual $[0_d; q_i^R]$	Residual + native $[x_i; q_i^R]$
Zero final-layer initialization before distillation					
CAM16 / AUC	0.9669±0.0183	0.9415±0.0306	0.9450±0.0290	0.9486±0.0556	**0.9672**±0.0146
CAM17 / F_1	0.6080±0.0327	0.5934±0.0464	**0.6248**±0.0339	0.5852±0.0608	0.6161±0.0505
Private-CRC / F_1	**0.8391**±0.0337	0.8159±0.0438	0.8063±0.0511	0.8370±0.0227	0.8346±0.0326
PANDA / κ	**0.8517**±0.0221	0.8251±0.0155	0.8496±0.0122	0.8445±0.0122	0.8471±0.0184
BRACS / F_1	0.3809±0.1116	0.3987±0.0517	**0.4323**±0.1115	0.4268±0.0741	0.4118±0.0515
KIRC / C-index	0.6984±0.0353	0.7095±0.0278	0.7106±0.0286	**0.7248**±0.0329	0.7187±0.0369
KIRP / C-index	**0.8363**±0.0911	0.8012±0.0756	0.8250±0.1022	0.8007±0.0782	0.7979±0.0849
LUAD / C-index	0.5453±0.0578	0.5343±0.0668	0.5262±0.0546	**0.5602**±0.0526	0.5453±0.0686
STAD / C-index	0.5495±0.0290	0.5487±0.0463	0.5617±0.0502	**0.5802**±0.0497	0.5768±0.0381
UCEC / C-index	0.6693±0.0576	0.6446±0.1337	0.6664±0.1133	0.6424±0.0568	**0.6743**±0.0797
Random final-layer initialization before distillation					
CAM16 / AUC	0.9669±0.0183	0.9455±0.0336	0.9611±0.0285	0.9563±0.0160	**0.9693**±0.0234
CAM17 / F_1	0.6080±0.0327	0.6068±0.0334	**0.6384**±0.0582	0.6164±0.0298	0.6105±0.0345
Private-CRC / F_1	0.8391±0.0337	0.8147±0.0467	0.8339±0.0330	**0.8494**±0.0298	0.8376±0.0404
PANDA / κ	0.8517±0.0221	0.8208±0.0118	**0.8521**±0.0109	0.8479±0.0169	0.8518±0.0100
BRACS / F_1	0.3809±0.1116	0.4313±0.0258	0.4299±0.0373	**0.4514**±0.0665	0.4077±0.0483
KIRC / C-index	0.6984±0.0353	0.6953±0.0591	0.7076±0.0471	0.7027±0.0484	**0.7177**±0.0470
KIRP / C-index	**0.8363**±0.0911	0.7936±0.0832	0.7790±0.0789	0.8044±0.0917	0.7869±0.0934
LUAD / C-index	0.5453±0.0578	0.5644±0.0759	**0.5663**±0.0585	0.5436±0.0525	0.5234±0.0400
STAD / C-index	0.5495±0.0290	0.5389±0.0356	**0.5651**±0.0396	0.5524±0.0721	0.5629±0.0669
UCEC / C-index	0.6693±0.0576	0.6764±0.1530	0.6600±0.0783	**0.7051**±0.1016	0.6726±0.0675

Table 25: Task NLL with the same trained prediction and either a zero or retained native block. All inputs have 3,072 coordinates. The matched native $5\times$ reference $[x_i; 0_d]$ is reused across panels. Lower is better.

Dataset	Native $5\times$ $[x_i; 0_d]$	Direct $[0_d; q_i^D]$	Direct + native $[x_i; q_i^D]$	Residual $[0_d; q_i^R]$	Residual + native $[x_i; q_i^R]$
Zero final-layer initialization before distillation					
CAM16	**0.2610**±0.0953	0.3580±0.1922	0.2924±0.1396	0.3253±0.1347	0.3048±0.1512
CAM17	0.6529±0.2769	0.9189±0.3316	**0.6241**±0.3428	0.7350±0.4200	0.6791±0.2495
Private-CRC	0.6192±0.0378	0.5823±0.1906	0.5690±0.0727	0.5371±0.0869	**0.5297**±0.1282
PANDA	1.2043±0.4426	1.0014±0.0692	**0.9764**±0.0898	1.5970±0.7405	1.1580±0.2178
BRACS	1.5287±0.6716	1.5428±0.3129	1.4855±0.5011	1.5966±0.4032	**1.4386**±0.3485
KIRC	1.8947±0.5592	1.7613±0.5343	**1.6468**±0.2609	1.7733±0.2047	1.8066±0.5783
KIRP	1.1656±0.4045	**1.1319**±0.1381	1.2582±0.2996	1.2343±0.2811	1.1800±0.2694
LUAD	**2.1217**±0.1648	2.6486±0.7593	2.3585±0.7398	2.7231±0.7288	2.4006±0.4799
STAD	2.6422±0.7560	**2.2103**±0.3926	2.6048±0.7348	2.7300±0.6291	2.5068±0.4820
UCEC	0.9924±0.1125	0.9382±0.0483	0.9375±0.0790	**0.9268**±0.0816	0.9309±0.0756
Random final-layer initialization before distillation					
CAM16	0.2610±0.0953	0.3762±0.1650	0.2669±0.1596	0.3478±0.1310	**0.2221**±0.1084
CAM17	**0.6529**±0.2769	0.8497±0.2992	0.7054±0.3072	0.7946±0.3287	0.7055±0.3602
Private-CRC	0.6192±0.0378	0.5807±0.0714	0.5703±0.0761	0.5354±0.0870	**0.5078**±0.0814
PANDA	1.2043±0.4426	1.1324±0.1417	1.0520±0.2602	1.1098±0.4763	**1.0221**±0.3193
BRACS	**1.5287**±0.6716	1.5451±0.4189	1.7362±0.5892	1.8601±0.5966	1.6750±0.4275
KIRC	1.8947±0.5592	1.8774±0.4209	1.9181±0.4312	1.9248±0.4497	**1.8219**±0.1382
KIRP	1.1656±0.4045	1.1808±0.2036	1.1251±0.1656	**1.0906**±0.1580	1.1478±0.2383
LUAD	2.1217±0.1648	**2.1058**±0.1759	2.3550±0.6127	2.4996±0.5208	2.4547±0.6061
STAD	2.6422±0.7560	2.4557±0.3826	**2.3816**±0.6888	2.8569±0.4871	2.4986±0.3642
UCEC	0.9924±0.1125	1.0223±0.2452	0.9615±0.0645	0.9622±0.0719	**0.9363**±0.0998

without retraining. All fifty dataset and realization replays agree within the recorded tolerance of .001, with maximum absolute logit difference 1.91×10^{-5}. They cover 14,685 slide appearances, including repeated slides across realizations. This validates the conversion for saved functions and illustrates complete recovery in the theorem. It does not equate fresh optimization outcomes. Both coordinates remain available in this replay.

E.6 Individual features or their mean as auxiliary supervision

The final study tests transfer through the image encoder itself. We adapt UNI2-h by distillation on the training images, freeze it and train downstream models on its features. The references are the published encoder, an encoder trained by mean KD and that model's mean-prediction output. Two further encoders combine mean KD with a flow objective whose target is either the regional mean or a sampled individual teacher feature. Their comparison changes the auxiliary target while retaining mean supervision and the same downstream readout.

Architecture and objectives. Let I_i be the raw $5\times$ patch and $c_i = E_\theta(I_i) \in \mathbb{R}^{1536}$ its adapted encoder feature. UNI2-h uses rank-8 LoRA updates on query and value projections in all 24 blocks, with pretrained weights and key projections frozen (Chen et al., 2024; **?**; Hu et al., 2022). A GELU head r_ψ with widths $1536 \to 1536 \to 1536$ predicts $m_i = K_i^{-1} \sum_{j=1}^{K_i} h_{ij}$ from c_i. The $1 \le K_i \le 16$ teacher vectors are the observed aligned $20\times$ patches assigned to the region. The distillation data do not require all sixteen positions. With coordinate scales $s_{m,d}$ estimated from the distillation training targets, the common mean loss is

$$\mathcal{L}_{\text{mean}} = \mathbb{E}_i \left[\frac{1}{1536} \sum_{d=1}^{1536} \text{SmoothL1}_1 \left(\frac{r_\psi(c_i)_d - m_{i,d}}{s_{m,d}} \right) + .1 \left(1 - \cos\big(r_\psi(c_i), m_i\big)\right) \right]. \tag{32}$$

SmoothL1 uses unit threshold, with value $e^2/2$ for $|e| < 1$ and $|e| - 1/2$ otherwise. Cosine similarity uses tolerance 10^{-8}. This objective accompanies encoder adaptation and differs from frozen-feature MSE prediction, so results across the two studies cannot isolate the loss function.

The auxiliary flow target is $a_i = m_i$ for mean supervision or $a_i = h_{iJ}$ for individual supervision, with J uniform over the K_i observed patches. Both use the centre μ and scale s_h estimated from distillation training data to form $A_i = (a_i - \mu) \oslash s_h$, where $\oslash$ is coordinatewise division. With Gaussian noise $\epsilon \sim \mathcal{N}(0, I)$ and time $s \sim \text{Unif}[0, 1)$, the field v_ω learns the velocity along a straight path (Lipman et al., 2023),

$$z_s = (1 - s)\epsilon + sA_i, \tag{33}$$

$$\mathcal{L}_{\text{flow}} = \mathbb{E}_{i,s,\epsilon,J} \left[\frac{1}{1536} \| v_\omega(z_s, s, c_i) - (A_i - \epsilon) \|_2^2 \right], \tag{34}$$

$$\mathcal{L} = \mathcal{L}_{\text{mean}} + .1\mathcal{L}_{\text{flow}}. \tag{35}$$

Only individual-feature supervision uses the draw J. Neither its index nor its spatial position enters the encoder or velocity field. Teacher features, path state and target velocity are detached, while the conditioning feature c_i carries gradients to the encoder. The velocity field has three modulated residual MLP blocks of width 512 and a 128-dimensional sinusoidal time encoding. Changing the sampled target also changes the path input z_s. The shared-output premise of Lemma 1 therefore does not hold, so replacing individual targets by their mean can change this auxiliary objective.

For both flow-trained encoders, downstream models receive only $E_\theta(I_i)$. We discard the mean head and velocity field, solve no differential equation and generate no teacher features at deployment. The separate head readout uses $r_\psi(E_\theta(I_i))$ from the mean-KD model. All task tokens have 1,536 coordinates. In Theorem 1, B now denotes raw images, which the adapted features need not recover. Both representation loss and model excess can consequently change, even though the auxiliary-target comparison holds task width constant.

Distillation populations and optimization. Three distillation partitions provide excluded-dataset evaluation across all ten datasets. The first trains on Private-CRC, PANDA, KIRC, KIRP, LUAD and STAD, then evaluates CAM16, CAM17, BRACS and UCEC. Partition A trains on CAM16, CAM17, BRACS, UCEC, KIRP and LUAD, then evaluates Private-CRC, PANDA and KIRC. Partition B trains

on CAM16, CAM17, BRACS, UCEC, Private-CRC and PANDA, then evaluates KIRP, LUAD and STAD. Each partition shares its distillation sample across the three objectives and five realizations. Evaluation datasets enter neither distillation training nor validation. These experiments use six distillation datasets, while the frozen-feature studies use nine.

Each distillation dataset contributes 1,024 fitting and 256 selection occurrences. Sampling first chooses a group, then a slide and an eligible region, retaining repeated occurrences in the objective. The six datasets contribute equal mass. Means and scales are estimated from fitting occurrences only. Individual-feature moments average within each region before averaging across occurrences, so regions with more observed teacher patches receive no extra weight. The mean loss uses coordinate SD of regional targets. The flow loss uses the common centre and coordinate SD of individual features, balanced by region. Both scales are floored at 10^{-4}.

Regions used for distillation can contain incomplete teacher grids. The first partition has 4,142 complete grids among 6,144 fitting occurrences, with two to fifteen observed patches in the remainder. Partitions A and B have 5,011 and 4,085 complete grids and include some regions with one patch. Uniformly sampling an observed teacher vector therefore weights the available patches within each region, rather than every physical position of a complete field.

Training uses 2,000 updates, effective batch size 96 and microbatches of 16. Each update includes sixteen fitting occurrences from each distillation dataset. AdamW uses learning rates 10^{-5} for encoder updates and 10^{-4} for the prediction head and field, weight decay .05 on decay-eligible parameters, ten percent warmup, cosine decay and gradient clipping at 1. BF16 forward passes retain FP32 parameters and losses, with TF32 disabled. Paired distillation realizations share their initial encoder, mean head, velocity field, occurrence schedule, noise and time streams across the models with flow supervision. Teacher indices are sampled independently of the noise and time draws.

Distillation validation compares checkpoints from updates 200 through 2,000 using each model's combined objective. Flow evaluation uses four fixed noise and time draws per occurrence, with one time from each quarter of the interval. The initial model is recorded but cannot be selected. All thirty flow fits choose update 2,000, holding selected training duration constant in the target comparison. Three objectives across three partitions and five realizations require 45 distillation fits.

Task comparison. All five task representations use the same selection of at most 128 raw $5\times$ patches per slide, with shared splits, task initialization and training within a realization. CAM16, CAM17, Private-CRC and BRACS train for fifteen epochs, and PANDA and survival for twenty. Coupled distillation and task seeds give 250 fits across five realizations. Test membership changes across realizations but matches across models within each. Because distillation partitions differ across evaluation datasets, variation in effects cannot be assigned solely to task category.

Table 26 compares all five readouts, including the mean-KD head. The frozen reference changes the interpretation of adapted models. Individual-feature flow has a higher mean than regional-mean flow on KIRP, yet both remain below the frozen encoder. Both flow encoders exceed the frozen reference on LUAD and STAD. The prediction head can also change the ordering relative to its own encoder, so auxiliary target choice and downstream readout require separate comparisons.

The preferred flow target varies across datasets. Individual-feature flow has higher test macro F_1 than regional-mean flow on CAM17, Private-CRC and BRACS, while CAM16 and PANDA favor regional-mean flow. Survival cohorts show both directions. Private-CRC favors individual-feature flow in all five test realizations, whereas KIRC favors regional-mean flow in all five. Their preferred targets reverse on average validation scores in Table 27. KIRP's positive test average for individual-feature flow combines two gains, two losses and one tie. These patterns support sensitivity to the auxiliary target without a rule for choosing it by task category or establishing recovery of the conditional distribution of teacher patches.

Table 26: Task scores from frozen UNI2-h, the mean-KD encoder and its prediction head, and encoders trained with mean KD plus flow supervision. Every readout has 1,536 coordinates on the same sampled $5\times$ patches. Published frozen UNI2-h is the reference. Higher is better.

Dataset / metric	Frozen UNI2-h	Mean KD Encoder	Mean KD Mean head	Mean KD + flow Regional mean	Mean KD + flow Individual feature
CAM16 / AUC	0.8808±0.0568	0.8734±0.0787	**0.8982**±0.0548	0.8827±0.0775	0.8749±0.0768
CAM17 / F_1	0.5025±0.0435	0.5115±0.0539	**0.5403**±0.0355	0.5077±0.0423	0.5211±0.0606
Private-CRC / F_1	0.8411±0.0262	0.8353±0.0509	0.8353±0.0408	0.8225±0.0434	**0.8492**±0.0282
PANDA / κ	0.8499±0.0155	**0.8519**±0.0131	0.8458±0.0117	0.8496±0.0122	0.8495±0.0067
BRACS / F_1	0.4589±0.0541	0.4331±0.0477	**0.4608**±0.0606	0.4204±0.0339	0.4347±0.0363
KIRC / C-index	0.7321±0.0407	0.7206±0.0176	0.7186±0.0451	**0.7378**±0.0176	0.7069±0.0332
KIRP / C-index	**0.7921**±0.0721	0.7519±0.0963	0.7504±0.1031	0.6861±0.1161	0.7744±0.1057
LUAD / C-index	0.5307±0.0481	0.5674±0.0424	0.5453±0.0380	0.5537±0.0268	**0.5783**±0.0396
STAD / C-index	0.5532±0.0235	0.5944±0.0312	0.5628±0.0607	0.5912±0.0316	**0.5953**±0.0331
UCEC / C-index	**0.6561**±0.0591	0.5980±0.0711	0.6348±0.1202	0.5924±0.0718	0.5976±0.0671

Table 27: Validation task scores at the selected downstream checkpoints. Both encoders retain mean KD. Regional-mean flow is the reference for individual-feature flow. Higher is better.

Dataset / metric	Flow toward regional mean	Flow toward individual feature
CAM16 / AUC	**0.8598**±0.0900	0.8538±0.0921
CAM17 / F_1	0.5468±0.0535	**0.5549**±0.1129
Private-CRC / F_1	**0.8889**±0.0303	0.8800±0.0367
PANDA / κ	**0.8709**±0.0114	0.8691±0.0136
BRACS / F_1	0.4201±0.0934	**0.4333**±0.0992
KIRC / C-index	0.6957±0.0499	**0.6965**±0.0566
KIRP / C-index	**0.8618**±0.0556	0.8587±0.0366
LUAD / C-index	**0.6386**±0.0513	0.6357±0.0734
STAD / C-index	**0.6428**±0.0863	0.6390±0.0849
UCEC / C-index	**0.8026**±0.0777	**0.8026**±0.0777