FACET: RETHINKING WHOLE-SLIDE FOUNDATION MODELS THROUGH TASK-CONDITIONED VIEWS

Anonymous authors
Paper under double-blind review

ABSTRACT

Foundation models (FMs) have demonstrated great success in computational pathology, first at the patch level and more recently at the whole-slide level. Despite being trained on increasingly large WSI corpora and, in some cases, incorporating additional molecular or textual modalities, existing slide-level FMs generally produce a single representation per slide, agnostic to the downstream task. This design is inherently limiting: a whole-slide image contains diverse morphological signals, only a subset of which may be relevant for any given clinical or biological question. We propose FACET, a whole-slide foundation model that conditions slide representations on task descriptions, enabling dynamic inference-time aggregation tailored to each prediction objective. FACET is pretrained through supervised multitask learning across 29 carefully curated prediction tasks spanning six cancer types, aligning task-conditioned slide representations with text-derived class prototypes. We evaluate FACET on an external benchmark comprising 40 distinct tasks, including tasks and cancer types not observed during pretraining. FACET achieves state-of-the-art performance against leading slide-level foundation models and maintains high performance on both unseen tasks and unseen cancer types, demonstrating robust generalization. Together, these results demonstrate task conditioning as a powerful approach for learning generalizable whole-slide representations. We release code, label files, task manifests, and evaluation resources at `https://github.com/anonymous-95/FACET`.

1 INTRODUCTION

Histopathology is central to clinical decision-making, and its digitization into gigapixel whole-slide images (WSIs) has enabled large-scale computational analysis (Kumar et al., 2020). Pathology foundation models (FMs) have emerged as a powerful paradigm for learning reusable representations, first at the patch level (Chen et al., 2024; Lu et al., 2024) and more recently at the whole-slide level, where slide representations are aggregated from sets of patch features (Li et al., 2025).

Recent slide-level FMs have demonstrated the value of complementary supervision for shaping WSI representations, including molecular profiles (Zhang et al., 2026) and multi-stain views (Jaume et al., 2024). In particular, language has emerged as a useful source of semantic guidance, through alignment with pathology reports (Ding et al., 2025; Shaikovski et al., 2024) and multiscale visual–text representations (Wong et al., 2026; Liu et al., 2026).

Despite this progress, existing slide-level FMs retain a structural limitation: each WSI is compressed into a single fixed representation. Even when richer supervision is used during pretraining, such as textual or molecular alignment, these signals are typically used to learn a general-purpose slide embedding, rather than to adapt the representation to the downstream task at inference time. This is inherently restrictive for histopathology: each slide contains diverse morphological signals, only a subset of which may be relevant to any given clinical or biological task, making a universal task-agnostic representation an inefficient summary of the slide (Peled et al., 2026a; Noorbakhsh et al., 2020). We empirically examine this premise in Figure 1, demonstrating that task identity shapes which tissue regions are prioritized and that transferring attention patterns across tasks degrades performance, suggesting that slide pooling contains an intrinsically task-dependent component.

To this end, we propose FACET, a slide-level foundation model that uses task descriptions to guide WSI representation at inference time. Rather than assigning each slide a fixed embedding, FACET

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

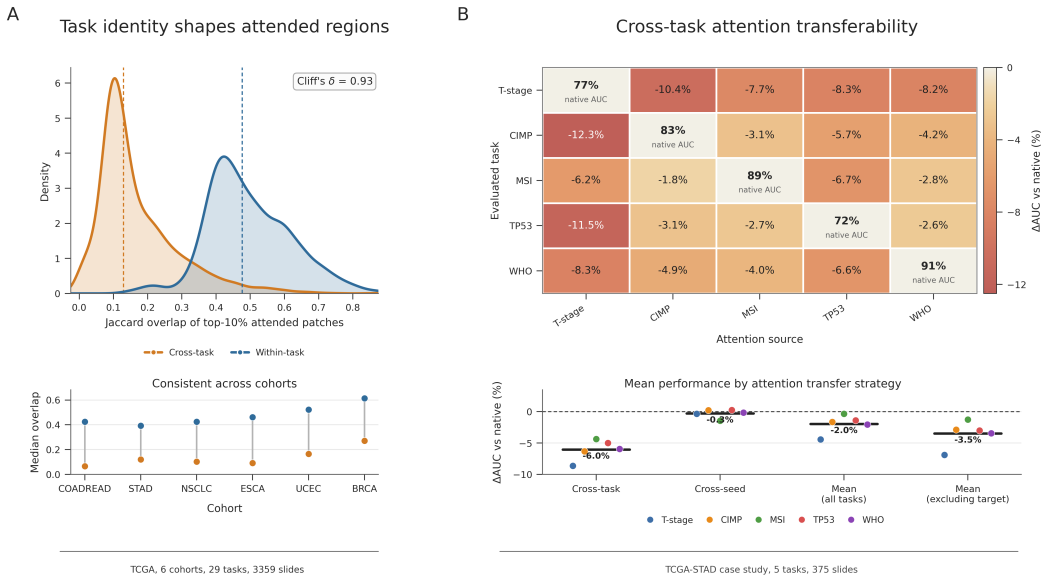

Figure 1: **Characterizing task attention motivates task-specific pooling.** (A) Independently trained ABMIL models generate attention maps per task and seed. Jaccard overlap of the top 10% most-attended patches was substantially higher across seeds of the same task than across different tasks (median 0.476 vs. 0.129; Cliff's $\delta = 0.93$), suggesting that task identity shapes attention beyond training stochasticity. This separation was consistent across all six cohorts. (B) In a TCGA-STAD case study, we evaluated cross-task attention transfer by substituting only the attention weights used for pooling, while keeping the rest of the model fixed. This transfer reduced AUC by 6.0% on average, with the largest losses involving T-stage and smaller losses between MSI and CIMP, potentially reflecting their known clinical association (Zong & Seto, 2014). Cross-seed transfer caused negligible degradation (-0.3%), indicating that cross-task losses were not driven by training stochasticity. Averaging attention across tasks reduced the loss to 2.0%, whereas excluding the target task increased it to 3.5%, indicating that part of the recovery was driven by the target task's own attention. Together, these results suggest that attention contains both shared and task-specific structure, motivating a foundation model that accounts for task variations as part of its core design. Further experimental details are provided in Appendix E.1.

aggregates patch features according to the prediction objective, capturing different task-aligned facets of the same slide. Figure 2 (bottom) illustrates this idea: the same slide yields different task-conditioned representations, each emphasizing the morphological evidence relevant to the queried task.

FACET is pretrained through supervised multitask learning over 29 carefully curated prediction tasks spanning six cancer types, aligning task-conditioned slide representations with text-derived class prototypes. We evaluate FACET on a broad external benchmark comprising 40 distinct tasks, where it outperforms leading slide-level foundation models. Importantly, these gains transfer to both unseen tasks and unseen cancer types, demonstrating robust generalization beyond the pretraining setup. Notably, FACET is pretrained on only 3,359 publicly available TCGA slides, compared with tens to hundreds of thousands of WSIs used by existing slide-level foundation models, as detailed in Sec. 4.5. Together, these results establish task conditioning as a powerful paradigm for learning generalizable whole-slide representations.

2 Related Work

Computational Pathology. Due to their gigapixel scale, WSIs are typically decomposed into patches and processed using multiple-instance learning (MIL) for slide-level prediction (Ilse et al., 2018). Recent MIL methods explore a range of aggregation strategies, including spatial context

modeling, as in PSA-MIL (Peled et al., 2026b); multi-scale learning, as in ViLa-MIL (Shi et al., 2024); and feature distillation, as in MHIM-MIL (Tang et al., 2026).

Patch-level Foundation Models in Pathology. Patch-level foundation models learn reusable histopathology representations through large-scale self-supervised pretraining (Chen et al., 2024; Zimmermann et al., 2024) or vision-language alignment (Huang et al., 2023; Lu et al., 2024). These models provide morphologically rich patch features, but they remain patch-centric and require a separate aggregation mechanism to form slide-level representations. Subsequent work has explored task-specific adaptation for slide-level prediction using semantic guidance, including alignment of visual tokens with textual class descriptions (Zhang et al., 2023), concept-guided feature adaptation (Huang et al., 2024), and prompt tuning for few-shot WSI classification (Chikontwe et al., 2024; Qu et al., 2024). While these methods demonstrate the value of incorporating task semantics into slide-level representation, they require separate retraining for each downstream objective. In contrast, FACET learns general-purpose slide representations that are dynamically adapted to downstream tasks at inference time through task-conditioned aggregation.

Slide-level Foundation Models in Pathology. Following the success of patch-level encoders, recent work has shifted toward slide-level FMs that learn WSI representations directly from sets of patch features. Existing approaches span self-supervised, weakly supervised, and fully supervised pretraining, as in GigaPath (Xu et al., 2024), CHIEF (Wang et al., 2024), and FEATHER (Shao et al., 2025), respectively. Vision-language models such as TITAN, PRISM, and PRISM2 align slide representations with pathology reports (Ding et al., 2025; Shaikovski et al., 2024; Vorontsov et al., 2026), with related approaches extending visual–text alignment across multiple spatial scales (Wong et al., 2026; Alawode et al., 2026; Liu et al., 2026). Other methods incorporate complementary supervision from molecular profiles, as in CARE (Zhang et al., 2026), or multi-stain views, as in Madeleine (Jaume et al., 2024). Beyond representation learning, recent VQA systems also use text during downstream reasoning: PRISM2 conditions diagnostic reasoning on the dialogue context after patch features have already been aggregated, while other approaches use the input question to retrieve relevant regions for LLM context (Vorontsov et al., 2026; Huang et al., 2026). Despite these advances, existing slide-level FMs generally retain a task-agnostic slide representation, with text providing useful semantic guidance during pretraining or after slide aggregation, rather than adapting the representation itself to the downstream prediction task. FACET instead uses task descriptions to dynamically adapt slide aggregation, producing task-specific representations at inference time.

3 METHODOLOGY

3.1 PROBLEM FORMULATION

Due to the scale of WSIs, tissue regions are first cropped into patches and encoded by a frozen patch encoder ϕ, yielding patch embeddings $x_i = \phi(\cdot) \in \mathbb{R}^d$. A slide is therefore represented as a set of feature vectors $X = \{x_i\}_{i=1}^N$. A slide-level encoder f_θ aggregates X into a compact representation $z \in \mathbb{R}^m$, which is subsequently used by downstream tasks $t \in \mathcal{T}$.

Existing slide-level foundation models typically instantiate f_θ as a transformer that maps the patch embeddings X to contextualized patch representations $\{h_i\}_{i=1}^N \in \mathbb{R}^m$, which are then pooled to produce a slide representation z. Many existing slide-level models use attention-based pooling (Ding et al., 2025; Shao et al., 2025; Shaikovski et al., 2024), in which the slide representation is given by

$$z = \sum_{i=1}^{N} a_i \, h_i = \mathbb{E}_{i \sim p_\theta(\cdot \mid X)}[h_i], \qquad a_i = p_\theta(i \mid X), \tag{1}$$

where $p_\theta(\cdot \mid X)$ is the learned attention distribution over patches.
Crucially, this formulation is *task-invariant*: a single representation z is assumed sufficient for all downstream objectives. Conceptually, this can be viewed from a task-marginal perspective: if patch relevance under task t is described by $p_\theta(i \mid X, t)$, then a task-agnostic weighting can be interpreted as averaging these task-specific relevance patterns under an assumed task prior $\pi(t)$,

$$p(i \mid X) = \sum_{t \in \mathcal{T}} \pi(t) \, p_\theta(i \mid X, t).$$

In this sense, a fixed slide representation is shaped by evidence that is useful on average across tasks, rather than by the requirements of the specific task being queried. This assumption is restrictive in histopathology, where downstream tasks are inherently heterogeneous, ranging from tumor subtyping and grading to biomarker prediction and outcome modeling, and often rely on distinct morphological cues expressed at different spatial and semantic scales. Figure 1 further provides empirical evidence for this variation by quantifying differences in attention patterns across tasks.

3.2 Task-Guided Slide Representation Learning

We address this limitation by replacing task-invariant slide representation learning with a task-conditioned formulation. Instead of requiring a single embedding z to support all downstream objectives, we learn a representation conditioned on the task, so that the evidence extracted from a slide depends on the prediction objective.

Formally, given a slide X and a task t, the slide encoder produces a task-conditioned representation $z_t \in \mathbb{R}^m$. Under attention-based aggregation, this can be written as

$$z_t = \sum_{i=1}^{N} a_i^{(t)} h_i = \mathbb{E}_{i \sim p_\theta(\cdot \mid X,t)}[h_i], \qquad a_i^{(t)} = p_\theta(i \mid X, t), \tag{2}$$

where $\{h_i\}_{i=1}^N$ denote patch representations, and $p_\theta(\cdot \mid X, t)$ is the induced task-dependent distribution over patches. Thus, each slide X is mapped to a family of representations $\{z_t(X)\}_{t \in \mathcal{T}}$, allowing the aggregated evidence to vary across tasks.

Task Representation via Text. Each task $t \in \mathcal{T}$ is represented by a natural-language description s_t that summarizes the prediction objective together with potentially relevant morphological cues observable in the slide. For example, a description for MSI prediction in colorectal cancer would encode the semantic context of the task, disease, and target classes, and may additionally suggest potentially relevant morphological cues such as immune infiltration and tumor architecture (Alexander et al., 2001). Task descriptions are generated using an LLM with a fixed prompting procedure conditioned only on the task objective and its class definitions, and remain fixed throughout training and inference. We further evaluate sensitivity to the description generator in Section 4.4. Example descriptions and the generation prompt are provided in Appendix B.3, with the full set released with our code. Each description is encoded by a pretrained text encoder g and mapped through a learned projection to obtain the task embedding

$$q_t = \mathrm{Proj}(g(s_t)).$$

The resulting embedding serves as the query for task-conditioned slide aggregation, providing the model with semantic information about which visual evidence may be relevant to the target prediction objective.

Task-Guided Slide Aggregation. Given patch-level features $X = \{x_i\}_{i=1}^N$, the slide encoder first produces contextualized patch representations $H = [h_1, \ldots, h_N] \in \mathbb{R}^{N \times m}$ using a transformer encoder. The task embedding q_t is then used as the query in a cross-attention pooling layer over H, yielding task-dependent patch weights and a corresponding task-specific slide representation. In matrix form, the pooling step is defined as

$$z_t = \mathrm{softmax}\left(\frac{(q_t W_q)(HW_k)^\top}{\sqrt{d}}\right)(HW_v) \tag{3}$$

where W_q, W_k, and W_v are learned projections. In practice, we use a multi-head attention layer followed by a residual connection. Thus, task conditioning acts directly at the aggregation stage, modulating which patch-level signals are emphasized in the resulting slide representation. Figure 2 outlines our overall framework (top) and illustrates how task semantics guide slide representations (bottom).

Supervised Text-Conditioned Objective. In addition to the task description s_t, each task $t \in \mathcal{T}$ is associated with a set of class descriptions $\{s_c^{(t)}\}_{c \in \mathcal{C}_t}$, where $\mathcal{C}_t$ denotes the label space of task t. Task and class descriptions are generated jointly using the procedure described in Section 3.2, with each class description characterizing a specific class within the corresponding task. Class descriptions are then encoded into textual class prototypes $r_c^{(t)} = \mathrm{Proj}(g(s_c^{(t)}))$. We then optimize a cross-entropy loss over cosine-similarity logits between z_t and the class prototypes. To further ground

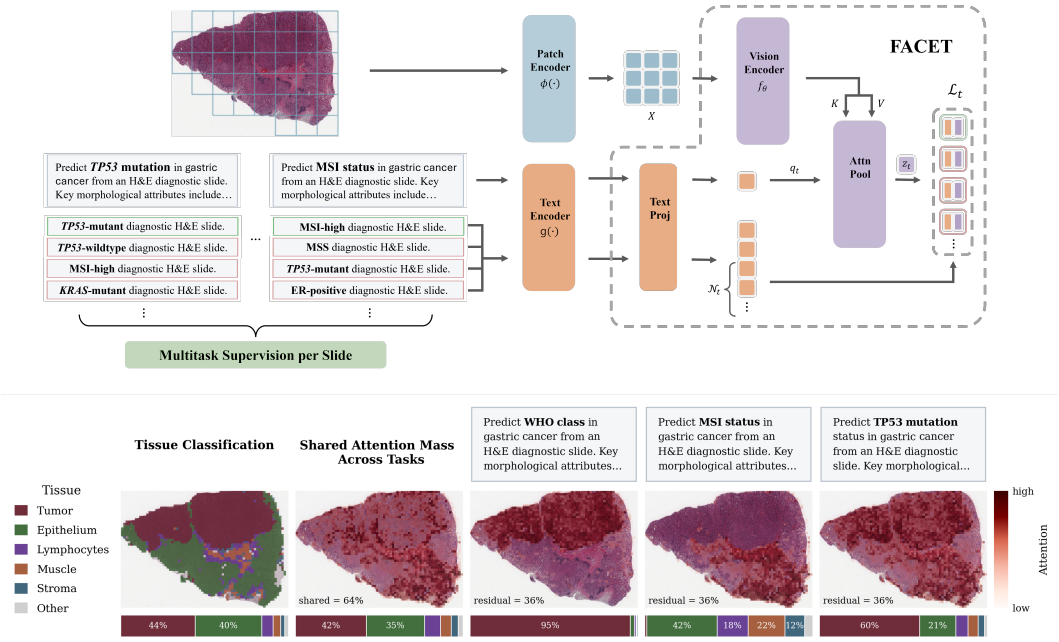

Figure 2: **Overview of the proposed framework.** Top: Multitask supervision and task-conditioned slide representation learning. A whole-slide image is partitioned into patches and encoded by a pretrained patch encoder $\phi(\cdot)$, while task descriptions and candidate class prompts are embedded by a pretrained text encoder $g(\cdot)$. The resulting visual tokens are further encoded by the vision encoder f_θ. Task-conditioned attention pooling uses the textual task embedding to query the visual tokens through cross-attention, producing a task-guided slide representation z_t that emphasizes visual evidence relevant to task t. Training aligns z_t with the true class prototype while contrasting it against sampled negative prototypes $\mathcal{N}_t$, grounding the slide representation in the semantics of the target task. Bottom: Qualitative visualization of task-conditioned attention. Approximately two-thirds of the attention mass is shared across tasks and is concentrated primarily along tumor boundaries and the surrounding microenvironment, suggesting that these regions provide common morphological evidence for multiple tasks. The residual attention maps reveal distinct task-dependent patterns: WHO classification primarily emphasizes tumor regions, whereas MSI shifts attention toward surrounding tissue and disproportionately emphasizes immune regions despite their limited spatial extent in the slide. Overall, this illustrates how FACET can emphasize different spatial evidence depending on the target task. Further details are given in the Appendix E.2.

slide aggregation in the queried task, we augment the loss with sampled cross-task prototypes that contrast the target class not only with in-task alternatives, but also with prototypes from other tasks. Thus, the overall objective is

$$s(z_t, r) = \frac{\cos(z_t, r)}{\tau}, \qquad \mathcal{L}_t = -\log \frac{\exp\left(s(z_t, r_{c^*}^{(t)})\right)}{\sum_{c \in \mathcal{C}_t} \exp\left(s(z_t, r_c^{(t)})\right) + \sum_{r \in \mathcal{N}_t} \exp(s(z_t, r))}, \tag{4}$$

where $c^* \in \mathcal{C}_t$ is the true label for task t, τ is a temperature parameter, and $\mathcal{N}_t \subset \{ r_c^{(t')} \mid t' \in \mathcal{T} \setminus \{t\},\ c \in \mathcal{C}_{t'} \}$ denotes a sampled set of k cross-task prototypes. By treating cross-task prototypes as negatives, even prototypes corresponding to true attributes of the slide can act as negatives under other tasks, encouraging the model to use the task description to disambiguate the correct class for the target task.

3.3 Multitask Pretraining

We pretrain FACET on 29 prediction tasks derived from six public TCGA cohorts — TCGA-BRCA, TCGA-NSCLC, TCGA-ESCA, TCGA-STAD, TCGA-COADREAD, and TCGA-UCEC — com-

prising 3,359 slides with a fixed train/validation split. The tasks were curated for clinical relevance and compatibility with multitask learning and span complementary prediction families, including tumor subtyping, grading, molecular status, and outcome-related objectives. Because they vary substantially in semantics, cohort size, class balance, and loss magnitude, effective pretraining requires balanced optimization across tasks.

Task Curation. Candidate tasks were constructed from cBioPortal[1] by screening for clinically and biologically meaningful attributes. We applied a label transformation step that grouped rare categories when appropriate, excluded ill-defined labels, and removed candidates with insufficient sample size or severe class imbalance. Since not all attributes are directly predictable from morphology, we further screened tasks for detectable slide-level signal using a simple ABMIL baseline. From the remaining pool, we manually selected a diverse set of tasks while ensuring label validity and biological relevance. Finally, following the procedure in Section 3.2, we used GPT-5.4 to generate the task and class descriptions consumed by the text encoder. We further evaluate sensitivity to the choice of text generator and text encoder in Section 4.4. Additional information, including full task definitions and prompting templates, is provided in Appendices B.2 and B.3, respectively.

Multitask Optimization. We train FACET with a hierarchically balanced sampling policy (Peled et al., 2026a), in which slide batches are sampled with equalized probability across cohorts, tasks, and classes, avoiding over-representation of dominant groups. Because loss magnitudes vary substantially across tasks, directly averaging task losses can destabilize training; we therefore normalize each task loss using an exponential moving average of its own loss trajectory. Survival tasks are optimized with a survival cross-entropy loss (Zadeh & Schmid, 2020), folded into the same multitask objective as in Equation 4. The final checkpoint is selected on a held-out validation split by maximizing $\mu - 0.5\sigma^2$, where μ and σ^2 denote the mean and variance of task-wise validation scores, favoring checkpoints with consistent, balanced performance across tasks to promote generalizable slide representations.

4 Experiments and Results

Experimental Setup. We evaluate FACET on nine *external* CPTAC cohorts (Edwards et al., 2015), comprising four cancer types represented during pretraining — CPTAC-BRCA, CPTAC-COAD, CPTAC-NSCLC, and CPTAC-UCEC — and five *unseen* cancer types — CPTAC-CCRCC, CPTAC-HNSC, CPTAC-GBM, CPTAC-OV, and CPTAC-PDA. Of the 40 evaluation tasks, our main analysis focuses on 34 *unseen* prediction tasks, providing a broad evaluation of generalization across both unseen prediction objectives and unseen cancer types. Results on additional tasks observed during pretraining are included in Appendix C.2. We adopt standardized task definitions and data splits from Patho-Bench (Zhang et al., 2025) to facilitate reproducible evaluation. Unless otherwise specified, classification and survival performance are reported as AUC and C-index, respectively, and all percentage improvements refer to relative changes in performance. Results are based on 50 cross-validation splits and are reported over 1,000 bootstrap resamples. Full evaluation task definitions are provided in Appendix C.2, and individual results are reported in Tables 5, 6, 7, and 8.

Implementation Details. FACET uses CONCH v1.5 as a frozen patch encoder and CONCH v1.0 text encoder to encode task and class descriptions (Lu et al., 2024), which are generated with GPT-5.4 following Section 3.2. The slide encoder comprises a two-layer, four-head transformer with a 256-dimensional hidden size and ALiBi positional encoding (Press et al., 2021), followed by task-conditioned cross-attention pooling. We use TRIDENT[2] for WSI preprocessing and patch extraction. Full architecture and training details are provided in Appendix A, with the complete configuration summarized in Table 3.

4.1 Benchmarking Against Slide-Level Foundation Models

We benchmark FACET on nine external cohorts from the CPTAC repository, and compare to leading slide-level foundation models, including CHIEF (Wang et al., 2024), GigaPath (Xu et al., 2024), PRISM (Shaikovski et al., 2024), Madeleine (Jaume et al., 2024), FEATHER (Shao et al., 2025), TITAN (Ding et al., 2025), and CARE (Zhang et al., 2026), using each model with its corresponding

[1] https://www.cbioportal.org

[2] https://github.com/mahmoodlab/trident/

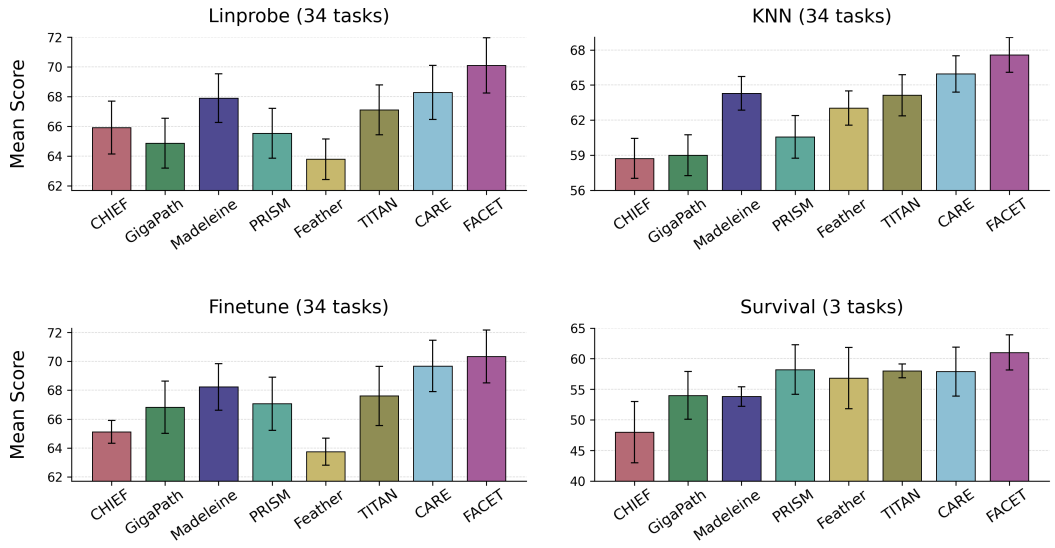

Figure 3: **Benchmarking against slide-level state-of-the-art foundation models.** Mean performance across 34 *unseen* classification tasks from external CPTAC cohorts, under linear probing, KNN ($k = 15$), and finetuning, together with CoxNet evaluation on three survival tasks. FACET achieves the highest average performance across all four settings, with the largest gains under frozen-representation evaluation, namely linear probing, KNN, and CoxNet. Under finetuning, FACET remains the strongest model, although the performance gap narrows as task-agnostic models gain task-specific adaptation to individual downstream tasks. Complete task-level results are provided in Tables 5, 6, 7, and 8 in the Appendix.

patch encoder. Figure 3 evaluates the resulting slide representations under four evaluation protocols: linear probing (linprobe), KNN ($k = 15$), and finetuning across 34 *unseen* classification tasks, together with CoxNet evaluation on three survival tasks. FACET achieves the highest average performance across all settings, with the largest gains observed when the slide encoder remains frozen, including linear probing, KNN, and CoxNet evaluations. Under linear probing and KNN, FACET improves over the strongest baseline, CARE, by 2.6% and 1.8%, respectively, while outperforming TITAN by 4.4% and 5.4%, and GigaPath by 8.0% and 14.5%. Across the three survival tasks, FACET also shows strong average performance, improving over the best baseline by 4.8%. Finetuning yields limited gains over linear probing, consistent with prior studies (Zhang et al., 2026; Ding et al., 2025). While FACET remains the strongest model, its advantage narrows in finetuning to 0.9% over CARE and 4.1% over TITAN, likely because finetuning introduces task-specific adaptation to otherwise task-agnostic representations. Notably, FACET achieves these results despite being pretrained on a substantially smaller WSI corpus than existing slide-level FMs, as discussed in Sec 4.5.

4.2 Generalization Across Seen and Unseen Cancer Types

We next examine generalization across cancer types, separating cancer types represented during pretraining from entirely unseen cancer types. Importantly, all 34 classification tasks are drawn from external CPTAC cohorts and are unseen during pretraining; we partition them according to whether the corresponding cancer type was represented during FACET pretraining. The represented cancer types include breast, colorectal, non-small-cell lung, and endometrial cancers (CPTAC-BRCA, CPTAC-COAD, CPTAC-NSCLC, and CPTAC-UCEC), whereas renal clear cell, head and neck, glioblastoma, ovarian, and pancreatic cancers (CPTAC-CCRCC, CPTAC-HNSC, CPTAC-GBM, CPTAC-OV, and CPTAC-PDA) were not observed during pretraining. As shown in Figure 4, FACET achieves the strongest linprobe performance in both settings. Across the 22 tasks from cancer types represented during pretraining, FACET outperforms the strongest baseline by 3.1%, compared with 1.5% across the 12 tasks from unseen cancer types. These results show that FACET generalizes

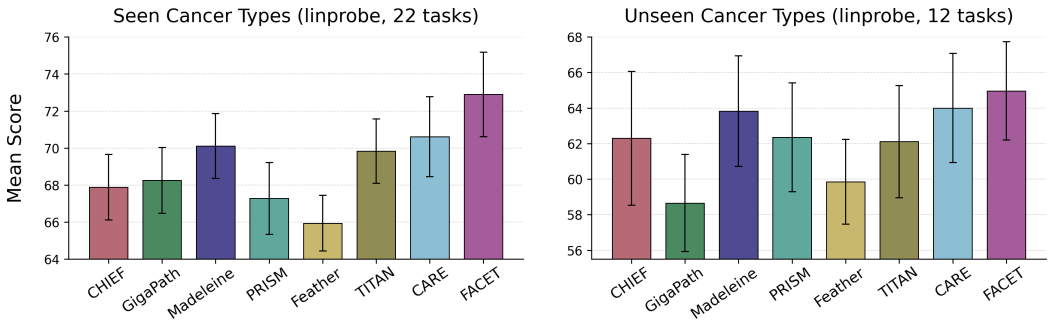

Figure 4: **Generalization across cancer types seen and unseen during pretraining.** Linear-probe performance across 34 *unseen* classification tasks, comparing cancer types seen versus unseen during pretraining. FACET achieves the highest average performance in both settings, with larger gains on cancer types represented during pretraining while retaining the strongest performance on entirely unseen cancer types. These results support the broad applicability of FACET beyond the pretraining setting.

across unseen tasks in both represented and entirely unseen cancer types, supporting its broader applicability beyond the pretraining setting.

4.3 Incorporating Clinical Context Through Task Descriptions

Pathologists rarely interpret morphology in isolation, but instead consider available clinical context alongside the tissue specimen. FACET provides a direct mechanism for incorporating such information: clinical attributes can be included in the task description, enabling clinical priors to be integrated into the slide representation without additional training. We examine this capability through a case study of smoking status in lung cancer, where smoking has been reported with fewer EGFR mutations and more frequent TP53 mutations (Devarakonda et al., 2021; Mounawar et al., 2007). For each patient, we retain the original task description and append a short statement indicating smoking status: *"The patient is a smoker."* or *"The patient is not a smoker."*. We use CPTAC-LUAD as the target lung cancer cohort and CPTAC-GBM as a control, where EGFR and TP53 are still relevant molecular targets, but smoking status is not expected to carry relevant biological information. Additionally, we randomly permute smoking status as an additional control for effects arising from task-description augmentation, averaging results over 10 independent permutations. As shown in Table 1, under linear probing, incorporating true smoking context improves LUAD performance by 2.8% for EGFR and 2.3% for TP53. In contrast, permuted smoking context decreases performance for both tasks, while neither true nor permuted smoking information improves performance in GBM. Together, these results provide a controlled example illustrating FACET's ability to incorporate clinically relevant context and improve downstream performance, without additional model training. Further experimental details are provided in Appendix E.3.

Table 1: Relative change in linear-probe AUC after augmenting task descriptions with true or randomly permuted smoking status. True smoking context improves both LUAD tasks, whereas neither true nor permuted smoking information improves performance in the GBM control cohort. Permuted smoking is averaged over 10 independent permutations; ± denotes 95% CI from 1000 paired test-set bootstraps.

Cohort	Mut.	FACET	+ Smoking	+ Perm. Smoking
LUAD	EGFR	81.1	↑2.8% ± 1.8%	↓1.5% ± 0.3%
	TP53	72.0	↑2.3% ± 1.3%	↓2.0% ± 0.4%
GBM	EGFR	66.3	↓2.9% ± 1.4%	↓1.9% ± 1.2%
	TP53	87.7	↓1.3% ± 1.5%	↓1.3% ± 0.4%

4.4 Ablation Studies

We evaluate ablations across 40 CPTAC tasks, using AUC for classification and C-index for survival. As shown in Table 2, replacing FACET's text-conditioned query with a learned query reduces performance by 4.4%.

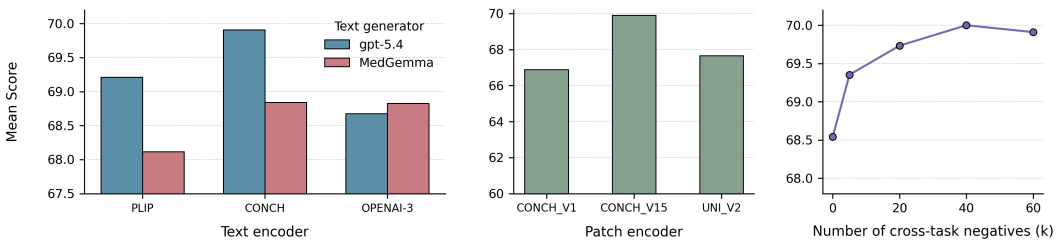

Figure 5: **Ablations.** Mean performance across 40 CPTAC tasks when varying the task-description generator and text encoder (left), patch encoder (middle), and number of cross-task negative prototypes (right).

Conditioning all encoder layers also reduces performance by 1.3%, suggesting that shared processing before conditioned aggregation is beneficial.

Table 2: **Ablation of task-conditioning design.** Mean performance across 40 CPTAC tasks. Task-conditioned aggregation outperforms both a task-agnostic learned query and conditioning throughout the encoder.

Conditioning strategy	Mean score
Learned query (no text)	66.86
Text conditioning at all layers	69.03
FACET (aggregation only)	**69.94**

Figure 5 further examines the model components. GPT-5.4-generated descriptions generally outperform those generated by MedGemma-27B (Sellergren et al., 2025), while both generators support effective task conditioning, indicating that FACET is not tied to a single description generator or exact phrasing. CONCH performs best as the text encoder, while CONCH v1.5 provides the strongest patch encoder. Increasing the number of cross-task negative prototypes improves performance up to approximately 40 negatives, after which gains plateau. Additional architectural and hyperparameter ablations are provided in Tables 9 and 10 in Appendix D.

4.5 Pretraining Dataset Scale

FACET is pretrained on 3,359 publicly available WSIs, an order-of-magnitude smaller corpus than those used by most recent slide-level foundation models. For comparison, CHIEF, GigaPath, TITAN, and PRISM are pretrained on approximately 61K, 171K, 336K, and 587K slides, respectively, while CARE uses 34K slides, FEATHER approximately 24K, and MADELEINE 16K multi-stain WSIs. Several of the largest pretraining corpora rely substantially on institutional or otherwise non-public data, making comparable scale difficult to reproduce using public repositories alone. Despite this smaller pretraining scale, FACET achieves strong downstream performance across our external benchmark.

5 Discussion

We introduce FACET, a whole-slide foundation model that learns task-adaptive slide representations, rather than assigning each slide one fixed representation. FACET uses natural-language task descriptions to guide slide aggregation toward morphological evidence relevant to the downstream task. Across an extensive external benchmark spanning 40 tasks, FACET consistently outperforms existing slide-level foundation models and generalizes to both unseen tasks and unseen cancer types, supporting its broad applicability. We further demonstrate FACET's ability to incorporate clinical metadata through task descriptions, providing clinical context beyond morphology alone. Notably, FACET achieves these results using only 3,359 pretraining slides, an order of magnitude fewer than many recent slide-level foundation models, which are often pretrained on tens to hundreds of thousands of slides. Together, these findings highlight the value of task-adaptive representation learning and suggest that representation design can provide an important axis of progress alongside dataset scaling.

REPRODUCIBILITY STATEMENT

We took several measures to support reproducibility throughout this work. We release our code together with pretrained FACET weights, which will be made available through HuggingFace. All experiments use publicly available TCGA and CPTAC whole-slide images. Importantly, downstream evaluation follows the standardized task definitions and predefined data splits from Patho-Bench, enabling consistent and reproducible comparison across models. The Appendix provides the complete pretraining task definitions (Section B), model and training configuration (Section A), and complete task-level results (Section C), together with the released code, label files, and task manifests to support transparent and reproducible use of FACET. We also release all task and class descriptions used in this work, together with the prompting templates and generation process used to construct them, including instructions for generating descriptions for new downstream tasks. These resources are intended to facilitate the use and evaluation of FACET and the reproduction of the reported results.

AI USE STATEMENT

In this work, we used generative AI tools to assist with improving the clarity and readability of the manuscript and with specific, well-defined coding tasks during implementation; all resulting changes were manually reviewed and verified for correctness. As part of our method, generative AI was also used to generate task and class descriptions from predefined prediction objectives and class definitions. We did not use generative AI for literature search or for reproducing results from other foundation models, which were obtained using their official codebases. Other required-disclosure categories were not applicable to this work. We have reviewed all AI-assisted work and take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

REFERENCES

Basit Alawode, Arif Mahmood, Muaz Khalifa Al Radi, Shahad Albastaki, Asim Khan, Muhammad Bilal, Moshira Ali Abdalla, Mohammed Bennamoun, and Sajid Javed. Mllm-hwsi: A multimodal large language model for hierarchical whole slide image understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13732–13743, 2026.

Julian Alexander, Toshiaki Watanabe, Tsung-Teh Wu, Asif Rashid, Shuan Li, and Stanley R Hamilton. Histopathological identification of colon cancer with microsatellite instability. *The American journal of pathology*, 158(2):527–535, 2001.

Chengfei Cai, Yangshu Zhou, Yiping Jiao, Liang Li, and Jun Xu. Prognostic analysis combining histopathological features and clinical information to predict colorectal cancer survival from whole-slide images. *Digestive Diseases and Sciences*, 69(8):2985–2995, 2024.

Richard J Chen, Tong Ding, Ming Y Lu, Drew FK Williamson, Guillaume Jaume, Bowen Chen, Andrew Zhang, Daniel Shao, Andrew H Song, Muhammad Shaban, et al. Towards a general-purpose foundation model for computational pathology. *Nature Medicine*, 2024.

Philip Chikontwe, Myeongkyun Kang, Miguel Luna, Siwoo Nam, and Sang Hyun Park. Low-shot prompt tuning for multiple instance learning based histology classification. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 285–295. Springer, 2024.

Siddhartha Devarakonda, Yize Li, Fernanda Martins Rodrigues, Sumithra Sankararaman, Humam Kadara, Chandra Goparaju, Irena Lanc, Kymberlie Pepin, Saiama N Waqar, Daniel Morgensztern, et al. Genomic profiling of lung adenocarcinoma in never-smokers. *Journal of Clinical Oncology*, 39(33):3747–3758, 2021.

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*, pp. 1–13, 2025.

Nathan J Edwards, Mauricio Oberti, Ratna R Thangudu, Shuang Cai, Peter B McGarvey, Shine Jacob, Subha Madhavan, and Karen A Ketchum. The cptac data portal: a resource for cancer proteomics research. *Journal of proteome research*, 14(6):2707–2713, 2015.

Maaike Galama, Nina Kozar-Gillan, Christina Embacher, Todd Dembo, Cornelius Böhm, Evelyn Ramberger, Julika Ribbat-Idel, Rosemarie Krupar, Verena Aumiller, Miriam Hägele, et al. Opentme: An open dataset of ai-powered h&e tumor microenvironment profiles from tcga. *arXiv preprint arXiv:2604.12075*, 2026.

Wentao Huang, Weimin Lyu, Peiliang Lou, Qingqiao Hu, Xiaoling Hu, Shahira Abousamra, Wenchao Han, Ruifeng Guo, Jiawei Zhou, Chao Chen, et al. Act like a pathologist: Tissue-aware whole slide image reasoning. *arXiv preprint arXiv:2603.00667*, 2026.

Yanyan Huang, Weiqin Zhao, Yihang Chen, Yu Fu, and Lequan Yu. Free lunch in pathology foundation model: Task-specific model adaptation with concept-guided feature enhancement. *Advances in Neural Information Processing Systems*, 37:79963–79995, 2024.

Zhi Huang, Federico Bianchi, Mert Yuksekgonul, Thomas J Montine, and James Zou. A visual–language foundation model for pathology image analysis using medical twitter. *Nature medicine*, 29(9):2307–2316, 2023.

Maximilian Ilse, Jakub Tomczak, and Max Welling. Attention-based deep multiple instance learning. In *International conference on machine learning*, pp. 2127–2136. PMLR, 2018.

Guillaume Jaume, Anurag Jayant Vaidya, Andrew Zhang, Andrew H Song, Richard J. Chen, Sharifa Sahai, Dandan Mo, Emilio Madrigal, Long Phi Le, and Mahmood Faisal. Multistain pretraining for slide representation learning in pathology. In *European Conference on Computer Vision*. Springer, 2024.

Jakob Nikolas Kather, Niels Halama, and Alexander Marx. 100,000 histological images of human colorectal cancer and healthy tissue. 2018.

Neeta Kumar, Ruchika Gupta, and Sanjay Gupta. Whole slide imaging (wsi) in pathology: current perspectives and future directions. *Journal of digital imaging*, 33(4):1034–1040, 2020.

Dong Li, Guihong Wan, Xintao Wu, Xinyu Wu, Xiaohui Chen, Yi He, Zhong Chen, Peter K Sorger, and Chen Zhao. Multi-modal foundation models for computational pathology: a survey. *Transactions on machine learning research*, 2025:5715, 2025.

Fengchun Liu, Songhan Jiang, Linghan Cai, Ziyue Wang, and Yongbing Zhang. Pathflip: Fine-grained language-image pretraining for versatile computational pathology. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 7132–7140, 2026.

Ming Y Lu, Bowen Chen, Drew FK Williamson, Richard J Chen, Ivy Liang, Tong Ding, Guillaume Jaume, Igor Odintsov, Long Phi Le, Georg Gerber, et al. A visual-language foundation model for computational pathology. *Nature medicine*, 30(3):863–874, 2024.

Mounia Mounawar, Anush Mukeria, Florence Le Calvez, Rayjean J Hung, Helene Renard, Alexis Cortot, Claire Bollart, David Zaridze, Paul Brennan, Paolo Boffetta, et al. Patterns of egfr, her2, tp53, and kras mutations of p14arf expression in non–small cell lung cancers in relation to smoking history. *Cancer research*, 67(12):5667–5672, 2007.

Javad Noorbakhsh, Saman Farahmand, Ali Foroughi Pour, Sandeep Namburi, Dennis Caruana, David Rimm, Mohammad Soltanieh-Ha, Kourosh Zarringhalam, and Jeffrey H Chuang. Deep learning-based cross-classifications reveal conserved spatial behaviors within tumor histological images. *Nature communications*, 11(1):6367, 2020.

Sharon Peled, Yosef E. Maruvka, and Moti Freiman. Multi-cancer framework with cancer-aware attention and adversarial mutual-information minimization for whole slide image classification. *Medical Image Analysis*, 109:103927, 2026a. ISSN 1361-8415. doi: https://doi.org/10.1016/j.media.2025.103927. URL https://www.sciencedirect.com/science/article/pii/S1361841525004736.

Sharon Peled, Yosef E Maruvka, and Moti Freiman. Psa-mil: a probabilistic spatial attention-based multiple instance learning for whole slide image classification. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1211–1220, 2026b.

Ofir Press, Noah A Smith, and Mike Lewis. Train short, test long: Attention with linear biases enables input length extrapolation. *arXiv preprint arXiv:2108.12409*, 2021.

Linhao Qu, Dingkang Yang, Dan Huang, Qinhao Guo, Rongkui Luo, Shaoting Zhang, and Xiaosong Wang. Pathology-knowledge enhanced multi-instance prompt learning for few-shot whole slide image classification. In *European conference on computer vision*, pp. 196–212. Springer, 2024.

Benoit Schmauch, Sarah S Elsoukkary, Amika Moro, Roma Raj, Chase J Wehrle, Kazunari Sasaki, Julien Calderaro, Patrick Sin-Chan, Federico Aucejo, and Daniel E Roberts. Combining a deep learning model with clinical data better predicts hepatocellular carcinoma behavior following surgery. *Journal of Pathology Informatics*, 15:100360, 2024.

Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, et al. Medgemma technical report. *arXiv preprint arXiv:2507.05201*, 2025.

George Shaikovski, Adam Casson, Kristen Severson, Eric Zimmermann, Yi Kan Wang, Jeremy D Kunz, Juan A Retamero, Gerard Oakley, David Klimstra, Christopher Kanan, et al. Prism: A multi-modal generative foundation model for slide-level histopathology. *arXiv preprint arXiv:2405.10254*, 2024.

Daniel Shao, Richard J Chen, Andrew H Song, Joel Runevic, Ming Y. Lu, Tong Ding, , and Faisal Mahmood. Do multiple instance learning models transfer? In *International conference on machine learning*, 2025.

Jiangbo Shi, Chen Li, Tieliang Gong, Yefeng Zheng, and Huazhu Fu. Vila-mil: Dual-scale vision-language multiple instance learning for whole slide image classification. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11248–11258. IEEE, 2024.

Satoshi Shirae, Shyam Sundar Debsarkar, Hiroharu Kawanaka, Bruce Aronow, and VB Surya Prasath. Multimodal ensemble fusion deep learning using histopathological images and clinical data for glioma subtype classification. *IEEE Access*, 13:57780–57797, 2025.

Wenhao Tang, Sheng Huang, Heng Fang, Fengtao Zhou, Bo Liu, and Qingshan Liu. Multiple instance learning framework with masked hard instance mining for gigapixel histopathology image analysis. *International Journal of Computer Vision*, 134(1):41, 2026.

Eugene Vorontsov, George Shaikovski, Adam Casson, Julian Viret, Eric Zimmermann, Neil Tenenholtz, Yi Kan Wang, Jan H Bernhard, Ran A Godrich, Juan A Retamero, et al. End-to-end multimodal pathology foundation model with clinical dialogue. *Nature Medicine*, pp. 1–11, 2026.

Xiyue Wang, Junhan Zhao, Eliana Marostica, Wei Yuan, Jietian Jin, Jiayu Zhang, Ruijiang Li, Hongping Tang, Kanran Wang, Yu Li, et al. A pathology foundation model for cancer diagnosis and prognosis prediction. *Nature*, 634(8035):970–978, 2024.

Bryan Wong, Jongwoo Kim, Huazhu Fu, and Mun Yi. Few-shot learning from gigapixel images via hierarchical vision-language alignment and modeling. *Advances in Neural Information Processing Systems*, 38:71313–71354, 2026.

Hanwen Xu, Naoto Usuyama, Jaspreet Bagga, Sheng Zhang, Rajesh Rao, Tristan Naumann, Cliff Wong, Zelalem Gero, Javier González, Yu Gu, et al. A whole-slide foundation model for digital pathology from real-world data. *Nature*, 630(8015):181–188, 2024.

Shekoufeh Gorgi Zadeh and Matthias Schmid. Bias in cross-entropy-based training of deep survival networks. *IEEE transactions on pattern analysis and machine intelligence*, 43(9):3126–3137, 2020.

Andrew Zhang, Guillaume Jaume, Anurag Vaidya, Tong Ding, and Faisal Mahmood. Accelerating data processing and benchmarking of ai models for pathology. *arXiv preprint arXiv:2502.06750*, 2025.

Di Zhang, Zhangpeng Gong, Xiaobo Pang, Jiashuai Liu, Junbo Lu, Hao Cui, Jiusong Ge, Zhi Zeng, Kai Yi, Yinghua Li, et al. Care: A molecular-guided foundation model with adaptive region modeling for whole slide image analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21078–21088, 2026.

Yunkun Zhang, Jin Gao, Mu Zhou, Xiaosong Wang, Yu Qiao, Shaoting Zhang, and Dequan Wang. Text-guided foundation model adaptation for pathological image classification. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 272–282. Springer, 2023.

Eric Zimmermann, Eugene Vorontsov, Julian Viret, Adam Casson, Michal Zelechowski, George Shaikovski, Neil Tenenholtz, James Hall, Thomas Fuchs, Nicolo Fusi, Siqi Liu, and Kristen Severson. Virchow2: Scaling self-supervised mixed magnification models in pathology. *arXiv preprint arXiv:2408.00738*, 2024.

Liang Zong and Yasuyuki Seto. Cpg island methylator phenotype, helicobacter pylori, epstein-barr virus, and microsatellite instability and prognosis in gastric cancer: a systematic review and meta-analysis. *PloS one*, 9(1):e86097, 2014.

APPENDIX

A ADDITIONAL IMPLEMENTATION DETAILS

A.1 MODEL ARCHITECTURE

Whole-slide images are processed using TRIDENT and partitioned into non-overlapping 512×512-pixel tiles at $20\times$ magnification. Each tile is encoded by the frozen CONCH v1.5 visual encoder, producing a 768-dimensional patch embedding. A residual MLP projects these embeddings to the 256-dimensional model space. The projected tokens are contextualized by a two-layer transformer with four attention heads and a dropout rate of 0.1. Spatial information is incorporated using two-dimensional ALiBi biases derived from the Manhattan distances between patch coordinates, without learned positional embeddings.

FACET represents each prediction objective using the task and class descriptions introduced in Section 3.2. These descriptions are encoded by the frozen CONCH v1.0 text encoder, which produces 512-dimensional text embeddings, and are mapped to the 256-dimensional model space through learned projections. The projected task embedding queries the contextualized patch tokens through multi-head cross-attention, yielding a task-conditioned slide representation.

For classification, the slide representation and projected class prototypes are ℓ_2-normalized, and logits are computed from their cosine similarities with a fixed temperature $\tau = 0.1$. The candidate set contains all classes belonging to the queried task together with 50 randomly sampled class prototypes from other pretraining tasks. The complete architecture configuration is summarized in Table 3.

A.2 PRETRAINING AND OPTIMIZATION DETAILS

FACET is pretrained on 3,359 slides from six TCGA cohorts and 29 curated prediction tasks, as described in Section 3.3. We reserve 15% of the cases for validation. Splitting is performed at the case level and stratified by cohort, task, and label, ensuring that slides from the same case cannot occur in both the training and validation sets.

Training uses a hierarchical sampling procedure that first balances cohorts, then tasks, and finally classes. Slides associated with multiple tasks are inversely weighted to prevent their repeated annotations from causing over-representation. Each task loss is normalized by an exponential moving average of its historical magnitude, with decay $\gamma = 0.99$, before losses are combined across tasks.

We optimize the model using AdamW with a learning rate of 10^{-5}, weight decay of 0.05, and cosine annealing over 50 epochs. Because a slide constitutes one micro-batch, gradients are accumulated

Table 3: Architecture and pretraining hyperparameters used for FACET.

Parameter	**Value**
Architecture	
Patch encoder	CONCH v1.5 (frozen)
Patch embedding dimension	768
Text encoder	CONCH v1.0 (frozen)
Text embedding dimension	512
Model hidden dimension	256
Transformer layers / attention heads	2 / 4
Spatial encoding	2D ALiBi (Manhattan distance)
Dropout	0.1
Patch size / magnification	512×512 pixels / $20\times$
Pretraining	
Optimizer	AdamW
Learning rate	1×10^{-5}
Weight decay	0.05
Learning-rate schedule	Cosine annealing, $T_{\max} = 50$
Micro-batch size / gradient accumulation	1 / 64
Maximum epochs	50
Temperature τ	0.1
Cross-task negative prototypes	50
EMA loss decay γ	0.99
Validation fraction	0.15 (case-level)
Checkpoint criterion	$\mu_{\text{val}} - 0.5\sigma^2_{\text{val}}$
Sampling policy	Balanced cohort–task–class

over 64 micro-batches before each optimization step. The final checkpoint maximizes

$$\mu_{\text{val}} - 0.5\sigma^2_{\text{val}},$$

where μ_{val} and σ^2_{val} are the mean and variance of the task-wise validation scores, respectively. This criterion favors checkpoints that perform strongly and consistently across the heterogeneous pretraining tasks. The pretraining hyperparameters are summarized in Table 3.

B Pretraining Task Curation and Definitions

B.1 Curation Pipeline

We construct the pretraining task collection from TCGA clinical metadata obtained through cBioPortal. For each cohort and clinical variable, we count the cases with non-missing values and the number of distinct categories. Variables with fewer than 80 available cases are excluded because they provide insufficient data for training and validation. We also exclude variables with fewer than two or more than ten categories, retaining only well-defined classification targets with a manageable number of classes.

Survival endpoints, including overall, disease-specific, and progression-free survival, are identified separately through their paired event-status and time-to-event variables. Closely related cohorts are combined where clinically appropriate, including COAD and READ as COADREAD and LUAD and LUSC as NSCLC. These combinations increase the number of available cases and the per-class sample counts.

Each candidate variable is reviewed using an LLM supplied with its column name, cohort, category counts, and structured curation instructions. The review flags variables unsuitable as morphology-based prediction targets, such as patient demographics, treatment information, and administrative or follow-up variables.

The review also proposes grouping biologically related classes and excluding ambiguous or insufficiently represented classes. These operations produce biologically coherent targets while reducing

label ambiguity and extreme class sparsity. For example, MSS and MSI-low are grouped as MSS in colorectal cancer because both are conventionally distinguished from MSI-high category. Indeterminate PR status in BRCA and "Not Evaluable" MSI labels in colorectal cancer are excluded because they do not represent definitive biological classes. Similarly, the GS and MSI subtypes in ESCA are excluded because each contains only one case, leaving the sufficiently represented ESCC and CIN classes. After class grouping and exclusion, we retain only candidates with at least 15 cases in every class. This minimum ensures that each class has sufficient representation for model training and reduces instability caused by extremely rare labels.

For each remaining classification candidate, we fit a simple ABMIL model to assess whether the attribute exhibits detectable slide-level morphological signal. Candidates with an AUC below 60% are excluded because they cannot be predicted reliably from the available histology and may therefore introduce predominantly noisy supervision during pretraining.

Finally, we manually verify the validity of the data sources, label definitions, class groupings, and exclusions. We then manually select 29 tasks across six cohorts, considering their clinical and biological relevance, sample size, class balance, ABMIL performance, and overall coverage of cohorts and prediction families. This selection promotes valid and biologically meaningful prediction targets while providing a diverse pretraining collection and limiting over-representation of any individual cohort or task family. Complete task definitions, label sources, class distributions, and applied groupings or exclusions are reported in Table 4.

B.2 Pretraining Tasks

The final pretraining collection comprises 29 tasks spanning six TCGA cohorts and a diverse set of clinical and biological prediction objectives. Table 4 reports the source cBioPortal attribute for each task, the original label distribution, all class grouping or exclusion operations applied during curation, and the resulting final label space used for pretraining.

B.3 Task and Class Description Generation

Task and class descriptions are generated from the curated task configurations using a fixed LLM prompting procedure. For each task, the generator receives the cancer type, prediction target, task type, and final class definitions, together with instructions to describe only morphological features potentially observable on routine H&E slides.

Importantly, these descriptions provide a natural-language representation of an independently curated prediction objective. Consequently, the textual representation contains semantic information from the disease context, target identity, and class definitions in addition to any generated morphological associations. For example, biomarker prediction in breast cancer and microsatellite-instability prediction in colorectal cancer remain semantically distinct even independently of the accompanying morphology.

The same generation procedure is used across all pretraining and evaluation tasks, and the resulting descriptions remain fixed throughout training and inference. We further evaluate the sensitivity of FACET to the choice of description generator and text encoder in Section 4.4.

Generation Prompt. The complete prompt used to generate task and class descriptions is provided below.

```
You are a computational pathology expert. Given a list of task configuration YAMLs describing
     histopathology tasks, generate a corresponding list of task_desc.json objects used to
     produce text embeddings for a multi-task whole-slide-image model.

### Context

I am training a supervised multi-task model on H&E whole-slide images across many tasks, with
     the goal of learning general-purpose slide representations. Each task has:
- A **task description** text, embedded and used to guide the model's attention toward task-
     relevant morphological features.
- A **class description** per class, embedded and used as class anchors -- the model's slide
     representation is compared against each class embedding to produce the final prediction.

### Input
```

Table 4: FACET pretraining tasks and label operations. Original and final labels are reported as class counts.

#	Cohort	Task (cBioPortal attribute)	Original labels	Label operations	Final labels
1	BRCA	ER status (ER Status)	"Positive": 542, "Negative": 156, "Not Performed": 30, "Performed but Not Available": 4, "Indeterminate": 2	**Excluded:** "Not Performed", "Performed but Not Available", "Indeterminate" **Grouped:** None	"ER_Positive": 542, "ER_Negative": 156
2	BRCA	HER2 status (HER2 Status)	"Negative": 587, "Positive": 100, "Equivocal": 9	**Excluded:** "Equivocal" **Grouped:** None	"HER2_Negative": 587, "HER2_Positive": 100
3	BRCA	PR status (PR Status)	"Positive": 470, "Negative": 226, "Not Performed": 31, "Performed but Not Available": 4, "Indeterminate": 3	**Excluded:** "Not Performed", "Performed but Not Available", "Indeterminate" **Grouped:** None	"Positive": 470, "Negative": 226
4	BRCA	Molecular subtype (Subtype)	"LumA": 436, "LumB": 176, "Basal": 152, "Her2": 65, "Normal": 29	**Excluded:** "Normal" **Grouped:** None	"LumA": 436, "LumB": 176, "Basal": 152, "Her2": 65
5	BRCA	Histologic type (Tumor Type)	"Infiltrating Ductal Carcinoma": 674, "Infiltrating Lobular Carcinoma": 180, "Other": 40, "Mixed Histology (NOS)": 26, "Mucinous Carcinoma": 17, "Metaplastic Carcinoma": 7, "Medullary Carcinoma": 6, "Infiltrating Carcinoma (NOS)": 1, "Breast Invasive Carcinoma": 1	**Excluded:** "Other", "Mixed Histology (NOS)", "Metaplastic Carcinoma", "Medullary Carcinoma", "Infiltrating Carcinoma (NOS)", "Breast Invasive Carcinoma" **Grouped:** None	"Ductal": 674, "Lobular": 180, "Mucinous": 17
6	COADREAD	DSS (Disease-specific Survival status)	"0:ALIVE OR DEAD TUMOR FREE": 387, "1:DEAD WITH TUMOR": 60	**Excluded:** None **Grouped:** None	"0:ALIVE OR DEAD TUMOR FREE": 387, "1:DEAD WITH TUMOR": 60
7	COADREAD	Histologic type (Cancer Type Detailed)	"Colon Adenocarcinoma": 314, "Rectal Adenocarcinoma": 118, "Mucinous Adenocarcinoma of the Colon and Rectum": 54, "Colorectal Adenocarcinoma": 5	**Excluded:** "Colorectal Adenocarcinoma" **Grouped:** ["Colon Adenocarcinoma", "Rectal Adenocarcinoma"]	"Conventional_Adenocarcinoma": 432, "Mucinous_Adenocarcinoma": 54
8	COADREAD	Hypermutation (Hyper-mutated)	"0.0": 110, "1.0": 22	**Excluded:** None **Grouped:** None	"Non-hypermutated": 110, "Hypermutated": 22
9	COADREAD	MLH1 silencing (MLH1 Silencing)	"0": 121, "1": 19	**Excluded:** None **Grouped:** None	"No_Silencing": 121, "Silenced": 19
10	COADREAD	MSI status (MSI Status)	"MSS": 118, "MSI-H": 26, "MSI-L": 25, "Not Evaluable": 1	**Excluded:** "Not Evaluable" **Grouped:** ["MSS", "MSI-L"]	"MSS": 143, "MSI": 26
11	ESCA	DSS (Disease-specific Survival status)	"0:ALIVE OR DEAD TUMOR FREE": 108, "1:DEAD WITH TUMOR": 45	**Excluded:** None **Grouped:** None	"0:ALIVE OR DEAD TUMOR FREE": 108, "1:DEAD WITH TUMOR": 45
12	ESCA	Tumor stage (American Joint Committee on Cancer Tumor Stage Code)	"T3": 73, "T2": 37, "T1": 24, "T4": 4, "T4a": 1, "T0": 1	**Excluded:** "T0", "T4a" **Grouped:** ["T1", "T2"] and ["T3", "T4"]	"Early_T": 61, "Late_T": 77
13	ESCA	Barrett's esophagus (Barretts esophagus)	"No": 93, "Yes-USA": 15, "Yes-UK": 10	**Excluded:** None **Grouped:** ["Yes-USA", "Yes-UK"]	"No": 93, "Yes": 25
14	ESCA	Histologic type (Cancer Type Detailed)	"Esophageal Squamous Cell Carcinoma": 90, "Esophageal Adenocarcinoma": 66	**Excluded:** None **Grouped:** None	"Esophageal Squamous Cell Carcinoma": 90, "Esophageal Adenocarcinoma": 66

15	ESCA	Lymph-node status (Primary Lymph Node Presentation Assessment)	“Yes”: 107, “No”: 32	**Excluded:** None **Grouped:** None	“Node_Positive”: 107, “Node_Negative”: 32
16	NSCLC	DSS (Disease-specific Survival status)	“0:ALIVE OR DEAD TUMOR FREE”: 321, “1:DEAD WITH TUMOR”: 102	**Excluded:** None **Grouped:** None	“0:ALIVE OR DEAD TUMOR FREE”: 321, “1:DEAD WITH TUMOR”: 102
17	NSCLC	Histologic type (Cancer Type Detailed)	“Lung Squamous Cell Carcinoma”: 458, “Lung Adenocarcinoma”: 446	**Excluded:** None **Grouped:** None	“Lung Squamous Cell Carcinoma”: 458, “Lung Adenocarcinoma”: 446
18	NSCLC	Mutation status (Mutation Status)	“NO”: 265, “YES”: 17	**Excluded:** None **Grouped:** None	“WT”: 265, “MUT”: 17
19	NSCLC	Lymph-node status (N Stage)	“N0”: 575, “N1”: 210, “N2”: 94, “Nx”: 14, “N3”: 7	**Excluded:** “Nx” **Grouped:** [“N1”, “N2”, “N3”]	“Node_Negative”: 575, “Node_Positive”: 311
20	NSCLC	Tumor stage (T Stage)	“T2”: 292, “T2a”: 155, “T3”: 101, “T1”: 101, “T1b”: 87, “T1a”: 68, “T2b”: 55, “T4”: 40, “Tx”: 2	**Excluded:** “Tx” **Grouped:** [“T1”, “T1a”, “T1b”, “T2”, “T2a”, “T2b”] and [“T3”, “T4”]	“Low_T”: 758, “High_T”: 141
21	STAD	MLH1 silencing (MLH1 Silencing)	“FALSE”: 168, “TRUE”: 33	**Excluded:** None **Grouped:** None	“FALSE”: 168, “TRUE”: 33
22	STAD	Molecular subtype (Molecular Subtype)	“CIN”: 128, “GS”: 50, “MSI”: 45, “EBV”: 23	**Excluded:** None **Grouped:** None	“CIN”: 128, “GS”: 50, “MSI”: 45, “EBV”: 23
23	STAD	MSI status (MSI Status)	“MSS”: 166, “MSI-H”: 45, “MSI-L”: 35	**Excluded:** None **Grouped:** [“MSS”, “MSI-L”]	“MSS”: 201, “MSI”: 45
24	STAD	Lymph-node status (Primary Lymph Node Presentation Assessment)	“Yes”: 341, “No”: 23	**Excluded:** None **Grouped:** None	“Yes”: 341, “No”: 23
25	STAD	WHO class (WHO Class)	“Tubular”: 117, “Poorly_Cohesive”: 58, “Mixed”: 17, “Papillary”: 16, “Mucinous”: 16	**Excluded:** None **Grouped:** None	“Tubular”: 117, “Poorly_Cohesive”: 58, “Mixed”: 17, “Papillary”: 16, “Mucinous”: 16
26	UCEC	DSS (Disease-specific Survival status)	“0:ALIVE OR DEAD TUMOR FREE”: 438, “1:DEAD WITH TUMOR”: 52	**Excluded:** None **Grouped:** None	“0:ALIVE OR DEAD TUMOR FREE”: 438, “1:DEAD WITH TUMOR”: 52
27	UCEC	Histologic grade (Neoplasm Histologic Grade)	“G3”: 267, “G2”: 117, “G1”: 96, “High Grade”: 10	**Excluded:** None **Grouped:** [“G1”, “G2”] and [“G3”, “High Grade”]	“Low”: 213, “High”: 277
28	UCEC	Molecular subtype (Subtype)	“CN_HIGH”: 146, “CN_LOW”: 138, “MSI”: 137, “POLE”: 47	**Excluded:** None **Grouped:** None	“CN_HIGH”: 146, “CN_LOW”: 138, “MSI”: 137, “POLE”: 47
29	UCEC	Histologic type (Tumor Type)	“Endometrioid Endometrial Adenocarcinoma”: 373, “Serous Endometrial Adenocarcinoma”: 96, “Mixed Serous and Endometrioid Carcinoma”: 21	**Excluded:** None **Grouped:** None	“Endometrioid”: 373, “Serous”: 96, “Mixed”: 21

```
You will receive a list of task configurations. Each configuration has these fields:
- `datasets`: list of dataset codes (e.g. "TCGA-ACC", "cptac_brca"). Use these to infer the
    source organ and cancer type.
- `task_col`: the prediction target column name (e.g. "TP53_mutation", "hyper_mutated", "
    dss__SURVIVAL").
- `original_column` (optional): the human-readable name of the original clinical variable (e.g
    . "Hyper-mutated", "Disease-specific Survival status").
- `original_time_column` (optional): for survival tasks, the human-readable name of the time
    variable.
- `task_type`: either "classification" or "survival".
- `metrics`: evaluation metrics (e.g. "macro-ovr-auc", "cindex").
- `label_dict`: mapping from integer class indices to class names.
- `sample_col`: the column used to identify samples.
- `num_samples`: total number of samples.
- `extra_cols` (optional): additional metadata columns.

### Important Assumption
```

```
All slides in this dataset are **standard diagnostic H&E slides**. Your descriptions must
     assume **no additional information beyond what is visible in a routine diagnostic H&E
     slide** -- no immunohistochemistry, no molecular assays, no special stains. Every
     morphological cue you mention must be identifiable on H&E alone.

### Task Text Rules

- Concise -- one to three short sentences.
- Describe the task generically, without mentioning dataset codes or dataset-specific metadata
      (no "TCGA", "CPTAC", no cohort IDs).
- Include the source organ and cancer type inferred from the dataset code (e.g. TCGA-COAD ->
     colon adenocarcinoma, cptac_brca -> breast cancer).
- For classification tasks: list all final classes (from `label_dict`) and briefly explain
     what each means.
- For survival tasks: describe the survival endpoint being predicted and explain the time-bin
     structure (censored vs. event bins).
- Mention only the key morphological or histopathological indicators that help differentiate
     the classes from one another -- features a pathologist would look for on **H&E only**.
     Keep this brief.

### Class Text Rules

- One short sentence per class.
- Describe what distinguishes this specific class from the others based solely on H&E
     morphology.
- Avoid details shared across all classes (these embeddings are used as contrastive anchors,
     so they must be maximally distinguishing).
- For survival tasks: describe both the time horizon (early vs. late bin) and the event status
      (censored vs. event occurred).

### Output Format

Return a JSON array of objects, one per input task, in the same order. Each object has:

{
    "task": "<task description>",
    "classes": {
        "0": "<class 0 description>",
        "1": "<class 1 description>",
        ...
    }
}

Keys under "classes" must be the string integer indices matching the keys in `label_dict`.
     Return only the JSON array, no extra commentary.
```

Example Task and Class Descriptions. Representative generated descriptions are shown below for microsatellite instability prediction in colorectal adenocarcinoma and estrogen receptor status prediction in breast carcinoma.

```
{
  "task": "Classify microsatellite instability status in colorectal adenocarcinoma. Classes:
       MSI vs. MSS. Key H&E indicators of MSI include abundant tumor-infiltrating lymphocytes,
        mucinous or medullary features, poor differentiation, and pushing growth pattern,
       whereas MSS tumors more often appear as conventional gland-forming adenocarcinomas.",
  "classes": {
    "0": "MSI tumor with dense lymphocytic infiltration, mucinous or medullary features, and
         reduced gland formation.",
    "1": "MSS tumor with more conventional gland-forming adenocarcinoma morphology and less
         prominent immune infiltration."
  }
}


{
  "task": "Predict estrogen receptor status in breast carcinoma. Classes: ER_Negative vs.
       ER_Positive. On H&E, ER-positive tumors more often show lower to intermediate grade,
       tubule formation, and less pleomorphism, whereas ER-negative tumors more often show
       high grade, solid growth, marked atypia, and brisk mitotic activity.",
  "classes": {
    "0": "ER-negative breast carcinoma with high-grade morphology, reduced tubule formation,
         marked nuclear pleomorphism, and frequent mitoses.",
    "1": "ER-positive breast carcinoma with more luminal-type morphology, including better
         gland formation, lower nuclear grade, and less proliferative appearance."
  }
}
```

The complete set of task and class descriptions used in FACET is released together with the code and task manifests, including descriptions for all 29 TCGA pretraining tasks and 40 CPTAC evaluation tasks.

C Evaluation Details and Results

C.1 Evaluation Protocols

We evaluate all models using standardized downstream task definitions and data splits from Patho-Bench (Zhang et al., 2025). Whole-slide representations are extracted using TRIDENT with each foundation model's corresponding pretrained patch and slide encoders. For FACET, slide representations are generated by conditioning the slide encoder on the downstream task description.

For classification tasks, we evaluate models using linear probing, k-nearest neighbors (KNN), and end-to-end finetuning, and performance is measured using area under the ROC curve (AUC). Linear probing trains a linear classifier on frozen slide representations, while KNN performs non-parametric classification using $k = 15$ nearest training examples.

Given the scale of the evaluation, comprising 40 downstream tasks assessed across 50 cross-validation splits, exhaustive model- and task-specific hyperparameter optimization would substantially increase the computational scope of the benchmark. Moreover, for some downstream tasks the available sample size is fewer than 150 samples in total, making extensive retraining both impractical and particularly prone to overfitting. We therefore adopted a budget-aware finetuning strategy aimed at adapting each pretrained model to the downstream task, rather than exhaustively optimizing or retraining it for each model–task combination. For each model, we evaluated a limited set of hyperparameter configurations on a representative subset of downstream tasks, varying the number of training epochs from 5 to 15 and the learning rate from 1×10^{-4} to 1×10^{-6}. Based on these experiments, we selected a model-specific configuration that provided stable performance across the evaluated tasks and applied it consistently across the remaining downstream tasks for that model. Across models, a generally stable configuration used AdamW with a learning rate of 1×10^{-5}, cosine learning-rate scheduling, a weight decay of 0.05, a batch size of 8, 15 training epochs, and weighted class balancing. Importantly, fine-tuning did not consistently improve over linear probing. This is consistent with previously reported behavior in histopathology, where end-to-end fine-tuning can underperform linear probing in limited-data regimes, likely because of overfitting. We observed a similar pattern in our experiments, including several cases of severe overfitting or optimization instability, particularly when the per-fold training set contained fewer than approximately 100 samples. Where applicable, we retain model-specific training components and follow the corresponding official training procedures; for example, CARE additionally uses its auxiliary training loss.

For survival tasks, FACET retains the same cosine-similarity logits used for classification, but replaces the classification cross-entropy with a survival cross-entropy loss (Zadeh & Schmid, 2020). Survival times are discretized into four equal-frequency bins during pretraining. For downstream evaluation, frozen slide representations are evaluated using CoxNet following the Patho-Bench implementation, with performance measured by the concordance index (C-index).

All reported results are aggregated over 50 predefined cross-validation splits and 1,000 bootstrap resamples. Confidence intervals are computed using a normal approximation based on the bootstrap standard error.

C.2 Evaluation Tasks and Per-Task Results

Our evaluation comprises 40 prediction tasks across nine external CPTAC cohorts. The main-text benchmark focuses on 34 classification tasks that were not observed during FACET pretraining, spanning both cancer types represented during pretraining and entirely unseen cancer types. The Appendix reports the complete task-level evaluation, including tasks whose prediction objectives were represented during pretraining.

Tables 5, 6, and 7 report per-task results for all 37 classification tasks under linear probing, KNN, and finetuning, and Table 8 reports the three overall-survival tasks under CoxNet.

Under linear probing, FACET achieves the best result on 12 of 37 tasks, roughly twice as many as CARE (6) and TITAN (5). FACET ranks among the top two on 18 tasks, compared with 14 for CARE and 7 for TITAN. This suggests that FACET's aggregate improvement reflects consistently strong performance across tasks rather than a small number of large gains.

D Additional Ablations

All ablations are evaluated on the 15% case-level validation split used during pretraining. We report the mean validation AUC across the 25 classification tasks; the remaining four pretraining tasks are disease-specific survival tasks and are excluded from this analysis.

D.1 Architectural Ablations

We evaluate the model hidden dimension, number of transformer layers, and number of attention heads while holding the remaining training configuration fixed. Results are reported in Table 9.

D.2 Hyperparameter Ablations

We additionally evaluate the number of training epochs, effective batch size, and learning rate while holding the remaining configuration fixed. Results are reported in Table 10.

E Supplementary Analyses

E.1 Task Attention Analysis (Figure 1)

For the analysis in Figure 1, we train independent ABMIL models for each prediction task. All models start from the same fixed random initialization, rather than from pretrained weights, so that differences in learned attention are not attributable to different initializations.

Attention similarity (Figure 1, left) is evaluated using the highest-attended 10% of patches for each slide. For the within-task comparison, we measure overlap across independently trained models with different random seeds, such that training differs in batch composition, dropout, and other sources of stochasticity. This provides a reference for the variability in attention induced by the training process itself. For the cross-task comparison, each pair of tasks is evaluated using models trained with the same random seed and on the same dataset, such that the primary difference between the models is the prediction target. This enables a more controlled comparison of task-dependent attention patterns.

For the attention-transfer experiment (Figure 1, right), only the attention weights used for slide pooling are replaced. The feature extractor and classifier of the target-task model remain unchanged. This isolates the effect of aggregation as part of the representation learning.

E.2 Task-Conditioned Attention Visualization (Figure 2)

The attention visualization in Figure 2 is computed from the task-conditioned cross-attention layer of the pretrained FACET model. To clearly visualize differences in attention across tasks, the displayed maps correspond to head #1.

To compare attention across tasks, attention values are aligned at the patch level for the same slide. The shared attention component is defined as the elementwise minimum attention across the evaluated tasks. For each task, the residual attention map is obtained by subtracting this shared component from the original task-specific attention map. This decomposition is used only for visualization and does not modify the model or its predictions. Each attention map is independently log-normalized to improve visualization clarity.

Patch-level tissue categories are assigned using a tissue classifier trained on the NCT-CRC-HE-100K dataset (Kather et al., 2018), and the resulting annotations were additionally compared manually against Atlas H&E-TME (Galama et al., 2026), a pathology foundation model for spatial profiling

Table 5: Linear-probing results on 37 CPTAC classification tasks. We report AUC with 95% confidence intervals. The best result for each task is bolded. Tasks marked with † correspond to prediction targets seen during FACET pretraining; all other tasks are unseen.

Cohort	Task	CHIEF	GigaPath	PRISM	Madeleine	FEATHER	TITAN	CARE	**FACET**
CPTAC-BRCA	Immune class	$69.8_{\pm 1.8}$	$71.6_{\pm 2.0}$	$64.6_{\pm 2.3}$	$73.1_{\pm 1.8}$	$68.1_{\pm 1.9}$	$69.2_{\pm 2.4}$	$\mathbf{76.1}_{\pm 1.7}$	$73.5_{\pm 1.9}$
	PIK3CA mut.	$\mathbf{63.5}_{\pm 2.5}$	$55.1_{\pm 3.6}$	$58.6_{\pm 3.9}$	$62.7_{\pm 3.2}$	$52.3_{\pm 3.2}$	$61.7_{\pm 3.6}$	$56.3_{\pm 3.6}$	$58.1_{\pm 3.0}$
	TP53 mut.	$82.4_{\pm 2.2}$	$76.4_{\pm 2.8}$	$\mathbf{84.2}_{\pm 1.7}$	$83.6_{\pm 2.5}$	$73.1_{\pm 3.0}$	$82.4_{\pm 2.0}$	$83.5_{\pm 2.2}$	$83.7_{\pm 2.3}$
CPTAC-COAD	ACVR2A mut.	$78.4_{\pm 3.6}$	$65.8_{\pm 3.9}$	$78.4_{\pm 3.0}$	$75.6_{\pm 3.8}$	$67.8_{\pm 3.7}$	$76.1_{\pm 3.1}$	$76.7_{\pm 3.7}$	$\mathbf{82.2}_{\pm 2.8}$
	APC mut.	$64.0_{\pm 4.6}$	$71.4_{\pm 3.8}$	$66.3_{\pm 3.5}$	$69.9_{\pm 4.2}$	$57.9_{\pm 3.9}$	$77.8_{\pm 3.4}$	$80.1_{\pm 3.3}$	$\mathbf{81.9}_{\pm 2.6}$
	ARID1A mut.	$72.9_{\pm 3.9}$	$79.3_{\pm 3.8}$	$71.0_{\pm 4.5}$	$77.2_{\pm 3.2}$	$65.5_{\pm 4.9}$	$72.6_{\pm 4.5}$	$78.2_{\pm 3.8}$	$\mathbf{85.2}_{\pm 3.0}$
	Immune class	$60.6_{\pm 1.9}$	$62.4_{\pm 2.1}$	$61.1_{\pm 2.1}$	$63.2_{\pm 2.1}$	$65.0_{\pm 2.8}$	$57.3_{\pm 1.9}$	$61.4_{\pm 2.2}$	$\mathbf{68.7}_{\pm 2.6}$
	KRAS mut.	$64.5_{\pm 3.2}$	$67.9_{\pm 3.1}$	$58.2_{\pm 3.1}$	$65.3_{\pm 3.5}$	$62.9_{\pm 4.1}$	$59.4_{\pm 3.1}$	$65.8_{\pm 3.7}$	$\mathbf{71.4}_{\pm 3.2}$
	MSI-H†	$85.8_{\pm 3.2}$	$84.0_{\pm 2.7}$	$89.3_{\pm 2.7}$	$85.6_{\pm 3.3}$	$73.5_{\pm 3.4}$	$87.1_{\pm 2.7}$	$89.7_{\pm 2.2}$	$\mathbf{94.6}_{\pm 1.3}$
	PIK3CA mut.	$63.3_{\pm 3.1}$	$\mathbf{71.6}_{\pm 3.5}$	$49.9_{\pm 3.3}$	$69.9_{\pm 2.8}$	$64.3_{\pm 3.5}$	$64.7_{\pm 4.1}$	$60.9_{\pm 3.4}$	$64.6_{\pm 3.3}$
	SETD1B mut.	$75.6_{\pm 3.5}$	$72.2_{\pm 4.3}$	$74.1_{\pm 4.2}$	$78.4_{\pm 4.1}$	$70.3_{\pm 5.1}$	$78.5_{\pm 3.7}$	$78.6_{\pm 3.7}$	$\mathbf{88.6}_{\pm 2.8}$
	TP53 mut.	$66.2_{\pm 3.6}$	$64.4_{\pm 3.3}$	$61.7_{\pm 3.6}$	$62.1_{\pm 4.0}$	$66.2_{\pm 3.4}$	$\mathbf{72.0}_{\pm 2.6}$	$69.9_{\pm 3.2}$	$70.8_{\pm 3.9}$
CPTAC-LUAD	EGFR mut.	$72.8_{\pm 3.5}$	$78.8_{\pm 2.4}$	$81.0_{\pm 2.6}$	$79.0_{\pm 3.1}$	$74.6_{\pm 2.5}$	$77.3_{\pm 2.8}$	$80.7_{\pm 3.0}$	$\mathbf{81.1}_{\pm 2.3}$
	Immune class	$71.7_{\pm 2.9}$	$63.2_{\pm 2.0}$	$66.4_{\pm 2.2}$	$72.9_{\pm 2.8}$	$64.0_{\pm 1.8}$	$67.7_{\pm 2.3}$	$68.2_{\pm 2.5}$	$\mathbf{74.9}_{\pm 2.0}$
	KRAS mut.	$62.9_{\pm 2.8}$	$69.8_{\pm 2.9}$	$67.9_{\pm 3.5}$	$64.8_{\pm 2.7}$	$66.6_{\pm 3.5}$	$69.6_{\pm 3.4}$	$\mathbf{75.0}_{\pm 2.6}$	$67.3_{\pm 2.8}$
	STK11 mut.	$83.4_{\pm 2.7}$	$82.7_{\pm 3.1}$	$78.5_{\pm 3.3}$	$85.0_{\pm 3.4}$	$81.5_{\pm 3.4}$	$86.4_{\pm 2.4}$	$\mathbf{88.5}_{\pm 2.4}$	$88.1_{\pm 3.1}$
	TP53 mut.	$70.7_{\pm 3.0}$	$71.1_{\pm 3.0}$	$\mathbf{74.6}_{\pm 2.3}$	$72.5_{\pm 2.9}$	$69.0_{\pm 3.4}$	$70.8_{\pm 2.6}$	$74.0_{\pm 2.9}$	$72.0_{\pm 2.9}$
CPTAC-LSCC	ARID1A mut.	$53.4_{\pm 5.1}$	$47.5_{\pm 3.0}$	$57.8_{\pm 4.3}$	$51.1_{\pm 3.7}$	$52.9_{\pm 3.9}$	$\mathbf{61.3}_{\pm 4.5}$	$44.9_{\pm 4.9}$	$45.2_{\pm 3.8}$
	Histologic grade†	$62.2_{\pm 2.5}$	$\mathbf{73.4}_{\pm 2.7}$	$60.2_{\pm 3.5}$	$68.0_{\pm 2.7}$	$65.4_{\pm 3.2}$	$59.5_{\pm 3.8}$	$65.0_{\pm 2.5}$	$64.6_{\pm 2.7}$
	Immune class	$70.4_{\pm 1.4}$	$73.4_{\pm 2.1}$	$65.3_{\pm 2.2}$	$\mathbf{76.4}_{\pm 1.7}$	$76.2_{\pm 2.1}$	$75.8_{\pm 2.5}$	$74.2_{\pm 1.5}$	$73.8_{\pm 2.4}$
	KEAP1 mut.	$64.2_{\pm 3.4}$	$65.1_{\pm 4.0}$	$51.7_{\pm 4.0}$	$\mathbf{68.2}_{\pm 3.5}$	$62.0_{\pm 3.8}$	$55.3_{\pm 4.3}$	$61.9_{\pm 4.2}$	$60.1_{\pm 4.0}$
CPTAC-UCEC	CTNNB1 mut.	$72.7_{\pm 3.3}$	$73.4_{\pm 3.1}$	$72.3_{\pm 2.7}$	$67.0_{\pm 3.5}$	$69.0_{\pm 3.0}$	$69.0_{\pm 3.7}$	$68.3_{\pm 3.4}$	$\mathbf{78.3}_{\pm 2.5}$
	Immune class	$55.3_{\pm 3.0}$	$56.9_{\pm 2.0}$	$69.6_{\pm 2.4}$	$64.3_{\pm 3.1}$	$60.8_{\pm 2.3}$	$67.0_{\pm 2.9}$	$\mathbf{70.2}_{\pm 2.5}$	$67.8_{\pm 2.1}$
	PTEN mut.	$54.6_{\pm 4.4}$	$61.5_{\pm 5.0}$	$\mathbf{66.8}_{\pm 4.7}$	$60.6_{\pm 4.3}$	$60.5_{\pm 3.5}$	$64.2_{\pm 4.9}$	$60.0_{\pm 4.2}$	$64.6_{\pm 4.6}$
CPTAC-CCRCC	BAP1 mut.	$72.6_{\pm 3.9}$	$62.6_{\pm 5.6}$	$59.9_{\pm 4.0}$	$\mathbf{74.7}_{\pm 3.3}$	$61.6_{\pm 5.3}$	$64.6_{\pm 3.9}$	$68.0_{\pm 4.8}$	$69.4_{\pm 3.1}$
	Immune class	$52.3_{\pm 2.1}$	$56.6_{\pm 2.7}$	$54.1_{\pm 2.5}$	$54.2_{\pm 2.3}$	$54.3_{\pm 2.8}$	$\mathbf{57.2}_{\pm 2.1}$	$53.1_{\pm 2.3}$	$54.7_{\pm 2.8}$
	PBRM1 mut.	$51.4_{\pm 3.3}$	$47.4_{\pm 3.0}$	$59.7_{\pm 3.8}$	$54.6_{\pm 3.1}$	$54.6_{\pm 3.0}$	$58.1_{\pm 3.0}$	$\mathbf{60.4}_{\pm 3.5}$	$57.1_{\pm 3.1}$
	VHL mut.	$62.3_{\pm 3.2}$	$\mathbf{65.3}_{\pm 3.1}$	$56.6_{\pm 4.6}$	$59.5_{\pm 3.0}$	$49.7_{\pm 3.4}$	$50.1_{\pm 4.1}$	$47.0_{\pm 3.1}$	$60.0_{\pm 3.4}$
CPTAC-GBM	EGFR mut.	$\mathbf{73.2}_{\pm 3.0}$	$62.7_{\pm 3.6}$	$64.4_{\pm 3.8}$	$68.7_{\pm 2.3}$	$70.0_{\pm 2.7}$	$72.4_{\pm 3.5}$	$67.9_{\pm 2.6}$	$66.3_{\pm 3.0}$
	Immune class	$73.9_{\pm 2.1}$	$67.5_{\pm 2.7}$	$72.6_{\pm 1.9}$	$72.9_{\pm 2.0}$	$73.4_{\pm 2.7}$	$\mathbf{76.2}_{\pm 2.0}$	$74.2_{\pm 1.7}$	$69.8_{\pm 2.3}$
	TP53 mut.	$86.2_{\pm 2.4}$	$76.6_{\pm 2.6}$	$85.0_{\pm 3.0}$	$86.7_{\pm 2.7}$	$69.3_{\pm 2.9}$	$85.0_{\pm 2.1}$	$81.9_{\pm 2.7}$	$\mathbf{87.7}_{\pm 2.5}$
CPTAC-HNSC	CASP8 mut.	$47.1_{\pm 6.1}$	$52.9_{\pm 5.0}$	$59.2_{\pm 5.7}$	$59.9_{\pm 5.3}$	$50.1_{\pm 7.0}$	$61.1_{\pm 5.1}$	$\mathbf{73.7}_{\pm 6.2}$	$72.7_{\pm 4.5}$
	Histologic grade†	$74.4_{\pm 2.5}$	$71.3_{\pm 3.3}$	$72.7_{\pm 2.9}$	$73.7_{\pm 2.4}$	$72.4_{\pm 3.3}$	$\mathbf{77.2}_{\pm 3.1}$	$76.1_{\pm 2.8}$	$73.5_{\pm 3.3}$
	Immune class	$65.4_{\pm 1.8}$	$63.0_{\pm 2.2}$	$\mathbf{74.1}_{\pm 1.6}$	$70.2_{\pm 1.8}$	$67.9_{\pm 2.2}$	$61.3_{\pm 1.5}$	$70.3_{\pm 2.2}$	$66.9_{\pm 2.1}$
CPTAC-OV	Immune class	$\mathbf{66.6}_{\pm 3.6}$	$53.3_{\pm 3.4}$	$59.3_{\pm 2.9}$	$58.4_{\pm 3.8}$	$56.3_{\pm 4.1}$	$54.0_{\pm 3.3}$	$65.3_{\pm 3.9}$	$63.9_{\pm 3.9}$
CPTAC-PDA	Immune class	$54.3_{\pm 2.5}$	$55.0_{\pm 2.0}$	$\mathbf{59.6}_{\pm 2.3}$	$52.7_{\pm 2.9}$	$55.5_{\pm 2.0}$	$57.6_{\pm 2.5}$	$55.4_{\pm 2.7}$	$52.9_{\pm 2.1}$
	SMAD4 mut.	$42.0_{\pm 3.3}$	$41.3_{\pm 4.2}$	$44.2_{\pm 4.3}$	$53.2_{\pm 3.7}$	$55.0_{\pm 4.1}$	$47.8_{\pm 4.1}$	$51.0_{\pm 3.8}$	$\mathbf{58.5}_{\pm 4.1}$

Table 6: KNN ($k = 15$) results on 37 CPTAC classification tasks. We report AUC with 95% confidence intervals. The best result for each task is bolded. Tasks marked with † correspond to prediction targets seen during FACET pretraining; all other tasks are unseen.

Cohort	Task	CHIEF	GigaPath	PRISM	Madeleine	FEATHER	TITAN	CARE	**FACET**
CPTAC-BRCA	Immune class	$70.8_{\pm 2.4}$	$70.5_{\pm 2.2}$	$64.4_{\pm 2.5}$	$\mathbf{74.7}_{\pm 2.3}$	$67.0_{\pm 2.4}$	$68.5_{\pm 2.2}$	$72.0_{\pm 2.4}$	$65.5_{\pm 2.3}$
	PIK3CA mut.	$64.4_{\pm 3.2}$	$56.4_{\pm 3.6}$	$53.8_{\pm 3.8}$	$61.4_{\pm 3.6}$	$55.8_{\pm 3.3}$	$58.2_{\pm 3.5}$	$62.2_{\pm 4.1}$	$\mathbf{65.5}_{\pm 3.2}$
	TP53 mut.	$79.8_{\pm 2.2}$	$72.9_{\pm 2.7}$	$81.0_{\pm 2.2}$	$81.3_{\pm 2.7}$	$75.9_{\pm 3.0}$	$81.1_{\pm 2.4}$	$\mathbf{84.1}_{\pm 2.6}$	$82.2_{\pm 2.2}$
CPTAC-COAD	ACVR2A mut.	$71.8_{\pm 3.6}$	$65.1_{\pm 4.4}$	$66.2_{\pm 3.3}$	$63.3_{\pm 3.9}$	$71.7_{\pm 4.5}$	$\mathbf{79.3}_{\pm 3.5}$	$70.1_{\pm 3.8}$	$76.3_{\pm 3.6}$
	APC mut.	$54.2_{\pm 4.4}$	$72.1_{\pm 4.3}$	$70.3_{\pm 4.0}$	$66.8_{\pm 4.2}$	$62.6_{\pm 4.4}$	$\mathbf{72.3}_{\pm 3.7}$	$68.7_{\pm 3.4}$	$69.6_{\pm 3.1}$
	ARID1A mut.	$59.1_{\pm 4.6}$	$56.5_{\pm 4.0}$	$46.6_{\pm 4.6}$	$74.8_{\pm 3.8}$	$64.9_{\pm 5.4}$	$60.6_{\pm 4.4}$	$68.5_{\pm 4.5}$	$\mathbf{79.9}_{\pm 3.3}$
	Immune class	$53.6_{\pm 2.4}$	$55.2_{\pm 2.2}$	$53.7_{\pm 2.7}$	$59.3_{\pm 2.3}$	$\mathbf{61.0}_{\pm 2.8}$	$55.1_{\pm 2.5}$	$57.4_{\pm 2.7}$	$59.2_{\pm 2.5}$
	KRAS mut.	$\mathbf{60.2}_{\pm 3.5}$	$53.5_{\pm 3.1}$	$57.1_{\pm 3.5}$	$52.1_{\pm 3.7}$	$59.6_{\pm 3.8}$	$59.4_{\pm 3.7}$	$57.5_{\pm 3.8}$	$59.9_{\pm 3.4}$
	MSI-H†	$77.3_{\pm 3.0}$	$77.8_{\pm 3.3}$	$62.3_{\pm 3.0}$	$75.0_{\pm 3.1}$	$72.0_{\pm 3.8}$	$78.2_{\pm 3.1}$	$74.1_{\pm 2.2}$	$\mathbf{89.0}_{\pm 1.8}$
	PIK3CA mut.	$53.8_{\pm 3.8}$	$64.2_{\pm 3.9}$	$46.9_{\pm 3.5}$	$59.9_{\pm 3.7}$	$64.8_{\pm 3.7}$	$60.7_{\pm 4.1}$	$61.1_{\pm 4.0}$	$\mathbf{67.3}_{\pm 3.7}$
	SETD1B mut.	$62.9_{\pm 4.0}$	$65.0_{\pm 4.3}$	$48.5_{\pm 4.3}$	$67.9_{\pm 4.4}$	$69.7_{\pm 5.4}$	$67.5_{\pm 3.8}$	$70.2_{\pm 4.2}$	$\mathbf{84.1}_{\pm 3.2}$
	TP53 mut.	$44.0_{\pm 4.0}$	$51.6_{\pm 3.3}$	$44.3_{\pm 4.0}$	$60.7_{\pm 4.4}$	$62.7_{\pm 3.6}$	$68.1_{\pm 3.2}$	$70.5_{\pm 3.5}$	$\mathbf{72.6}_{\pm 3.7}$
CPTAC-LUAD	EGFR mut.	$67.8_{\pm 3.6}$	$73.4_{\pm 3.3}$	$73.2_{\pm 3.0}$	$73.4_{\pm 3.2}$	$74.4_{\pm 3.3}$	$79.6_{\pm 3.1}$	$\mathbf{79.8}_{\pm 3.2}$	$74.4_{\pm 2.9}$
	Immune class	$\mathbf{64.7}_{\pm 2.6}$	$58.0_{\pm 2.4}$	$57.4_{\pm 2.3}$	$62.9_{\pm 2.5}$	$54.2_{\pm 2.4}$	$62.2_{\pm 2.5}$	$64.0_{\pm 2.4}$	$60.2_{\pm 2.5}$
	KRAS mut.	$62.7_{\pm 3.2}$	$\mathbf{70.7}_{\pm 3.4}$	$64.3_{\pm 3.4}$	$64.0_{\pm 3.2}$	$68.7_{\pm 3.8}$	$67.3_{\pm 3.4}$	$67.0_{\pm 3.3}$	$62.2_{\pm 3.1}$
	STK11 mut.	$74.7_{\pm 3.3}$	$83.8_{\pm 3.2}$	$84.0_{\pm 3.8}$	$83.3_{\pm 3.7}$	$81.7_{\pm 3.5}$	$83.7_{\pm 3.3}$	$\mathbf{87.6}_{\pm 3.0}$	$84.5_{\pm 3.2}$
	TP53 mut.	$62.4_{\pm 3.0}$	$66.4_{\pm 3.1}$	$69.4_{\pm 2.9}$	$66.6_{\pm 2.9}$	$61.5_{\pm 3.3}$	$66.5_{\pm 2.9}$	$\mathbf{70.8}_{\pm 3.0}$	$67.9_{\pm 3.0}$
CPTAC-LSCC	ARID1A mut.	$50.7_{\pm 4.8}$	$55.7_{\pm 3.9}$	$57.4_{\pm 4.6}$	$58.0_{\pm 3.7}$	$\mathbf{62.7}_{\pm 4.7}$	$53.6_{\pm 4.7}$	$46.8_{\pm 4.8}$	$55.6_{\pm 4.4}$
	Histologic grade†	$63.0_{\pm 3.1}$	$64.2_{\pm 3.1}$	$63.1_{\pm 3.7}$	$59.9_{\pm 3.4}$	$59.0_{\pm 3.6}$	$54.2_{\pm 3.9}$	$\mathbf{64.7}_{\pm 3.3}$	$57.2_{\pm 3.3}$
	Immune class	$55.3_{\pm 1.9}$	$60.7_{\pm 2.0}$	$55.9_{\pm 2.2}$	$73.0_{\pm 2.0}$	$\mathbf{73.5}_{\pm 2.3}$	$69.8_{\pm 2.3}$	$70.5_{\pm 2.1}$	$68.2_{\pm 2.3}$
	KEAP1 mut.	$55.9_{\pm 3.6}$	$66.0_{\pm 3.8}$	$58.8_{\pm 4.6}$	$\mathbf{69.1}_{\pm 3.8}$	$54.9_{\pm 3.9}$	$45.3_{\pm 5.1}$	$64.8_{\pm 4.3}$	$68.2_{\pm 4.0}$
CPTAC-UCEC	CTNNB1 mut.	$63.8_{\pm 3.4}$	$64.8_{\pm 3.8}$	$68.5_{\pm 3.0}$	$67.2_{\pm 3.7}$	$68.7_{\pm 3.8}$	$65.9_{\pm 4.0}$	$68.3_{\pm 3.5}$	$\mathbf{74.3}_{\pm 3.3}$
	Immune class	$48.5_{\pm 2.8}$	$41.7_{\pm 2.6}$	$58.5_{\pm 2.4}$	$57.8_{\pm 2.9}$	$46.9_{\pm 2.8}$	$59.6_{\pm 2.9}$	$61.5_{\pm 2.4}$	$\mathbf{62.8}_{\pm 2.5}$
	PTEN mut.	$52.2_{\pm 5.0}$	$54.4_{\pm 5.3}$	$\mathbf{78.3}_{\pm 4.8}$	$63.7_{\pm 5.0}$	$58.9_{\pm 4.1}$	$66.2_{\pm 4.8}$	$64.3_{\pm 4.7}$	$75.1_{\pm 4.9}$
CPTAC-CCRCC	BAP1 mut.	$62.9_{\pm 4.2}$	$57.8_{\pm 5.6}$	$65.1_{\pm 4.8}$	$\mathbf{74.3}_{\pm 4.0}$	$68.5_{\pm 5.4}$	$67.5_{\pm 4.5}$	$72.4_{\pm 5.0}$	$69.1_{\pm 3.3}$
	Immune class	$47.8_{\pm 2.3}$	$51.7_{\pm 2.6}$	$\mathbf{56.2}_{\pm 2.9}$	$52.7_{\pm 2.4}$	$53.9_{\pm 2.6}$	$49.0_{\pm 2.5}$	$55.6_{\pm 2.5}$	$49.7_{\pm 2.6}$
	PBRM1 mut.	$55.3_{\pm 3.3}$	$47.9_{\pm 3.7}$	$\mathbf{60.0}_{\pm 3.8}$	$56.4_{\pm 3.2}$	$59.0_{\pm 3.6}$	$57.4_{\pm 3.4}$	$54.5_{\pm 3.6}$	$59.9_{\pm 3.6}$
	VHL mut.	$55.8_{\pm 3.8}$	$58.8_{\pm 3.5}$	$55.9_{\pm 4.7}$	$58.7_{\pm 3.8}$	$47.8_{\pm 3.9}$	$46.8_{\pm 4.0}$	$55.9_{\pm 3.9}$	$\mathbf{64.3}_{\pm 3.8}$
CPTAC-GBM	EGFR mut.	$67.2_{\pm 3.5}$	$52.0_{\pm 3.7}$	$57.0_{\pm 3.5}$	$64.8_{\pm 2.9}$	$62.0_{\pm 3.0}$	$\mathbf{75.6}_{\pm 3.2}$	$69.6_{\pm 3.0}$	$66.3_{\pm 3.1}$
	Immune class	$66.2_{\pm 2.4}$	$61.2_{\pm 2.5}$	$74.1_{\pm 2.3}$	$66.4_{\pm 2.4}$	$71.9_{\pm 2.5}$	$74.5_{\pm 2.3}$	$\mathbf{75.0}_{\pm 2.0}$	$68.4_{\pm 2.2}$
	TP53 mut.	$74.8_{\pm 2.3}$	$60.9_{\pm 3.2}$	$74.8_{\pm 3.2}$	$74.9_{\pm 2.6}$	$61.1_{\pm 3.5}$	$78.4_{\pm 2.5}$	$\mathbf{80.4}_{\pm 2.6}$	$75.9_{\pm 3.1}$
CPTAC-HNSC	CASP8 mut.	$45.4_{\pm 6.1}$	$52.8_{\pm 5.5}$	$44.1_{\pm 5.9}$	$56.5_{\pm 5.2}$	$\mathbf{74.2}_{\pm 8.0}$	$56.7_{\pm 5.6}$	$70.6_{\pm 6.1}$	$72.2_{\pm 5.0}$
	Histologic grade†	$76.5_{\pm 2.9}$	$74.9_{\pm 3.7}$	$61.9_{\pm 3.1}$	$75.7_{\pm 3.0}$	$73.8_{\pm 3.4}$	$71.3_{\pm 3.0}$	$\mathbf{78.9}_{\pm 3.2}$	$68.7_{\pm 3.5}$
	Immune class	$50.4_{\pm 2.2}$	$46.6_{\pm 2.2}$	$58.1_{\pm 2.0}$	$59.1_{\pm 2.2}$	$66.3_{\pm 2.2}$	$62.2_{\pm 2.1}$	$\mathbf{67.9}_{\pm 2.5}$	$59.4_{\pm 2.2}$
CPTAC-OV	Immune class	$52.2_{\pm 3.8}$	$55.3_{\pm 3.4}$	$55.9_{\pm 3.3}$	$57.9_{\pm 3.9}$	$55.0_{\pm 4.0}$	$60.7_{\pm 3.8}$	$58.4_{\pm 4.0}$	$\mathbf{66.0}_{\pm 3.7}$
CPTAC-PDA	Immune class	$52.6_{\pm 2.7}$	$51.4_{\pm 2.3}$	$49.6_{\pm 2.5}$	$50.1_{\pm 2.8}$	$48.2_{\pm 2.5}$	$51.7_{\pm 2.7}$	$\mathbf{57.8}_{\pm 2.7}$	$48.7_{\pm 2.5}$
	SMAD4 mut.	$32.9_{\pm 3.8}$	$31.5_{\pm 4.3}$	$49.5_{\pm 4.4}$	$52.7_{\pm 4.5}$	$53.2_{\pm 4.4}$	$49.2_{\pm 4.2}$	$51.9_{\pm 4.3}$	$\mathbf{62.8}_{\pm 4.6}$

Table 7: Finetuning results on 37 CPTAC classification tasks. We report AUC with 95% confidence intervals. The best result for each task is bolded. Tasks marked with † correspond to prediction targets seen during FACET pretraining; all other tasks are unseen.

Cohort	Task	CHIEF	GigaPath	PRISM	Madeleine	FEATHER	TITAN	CARE	**FACET**
CPTAC-BRCA	Immune class	$64.7_{\pm 2.7}$	$\mathbf{76.1}_{\pm 1.9}$	$69.4_{\pm 1.7}$	$72.9_{\pm 2.0}$	$64.4_{\pm 3.0}$	$75.8_{\pm 2.1}$	$75.2_{\pm 1.8}$	$74.0_{\pm 2.3}$
	PIK3CA mut.	$60.5_{\pm 4.4}$	$55.8_{\pm 3.7}$	$58.1_{\pm 4.1}$	$\mathbf{61.3}_{\pm 3.2}$	$60.3_{\pm 4.3}$	$58.6_{\pm 3.4}$	$59.2_{\pm 3.1}$	$59.9_{\pm 3.1}$
	TP53 mut.	$80.2_{\pm 4.3}$	$79.3_{\pm 2.4}$	$84.9_{\pm 1.9}$	$84.0_{\pm 2.3}$	$77.6_{\pm 4.5}$	$84.7_{\pm 2.2}$	$\mathbf{86.2}_{\pm 2.1}$	$85.6_{\pm 2.2}$
CPTAC-COAD	ACVR2A mut.	$75.4_{\pm 4.8}$	$78.6_{\pm 3.3}$	$\mathbf{82.0}_{\pm 3.4}$	$71.4_{\pm 4.3}$	$69.1_{\pm 6.0}$	$81.3_{\pm 2.7}$	$77.3_{\pm 3.9}$	$80.4_{\pm 3.7}$
	APC mut.	$67.0_{\pm 4.7}$	$73.7_{\pm 3.4}$	$76.5_{\pm 3.3}$	$70.8_{\pm 3.7}$	$59.3_{\pm 4.0}$	$75.7_{\pm 2.4}$	$\mathbf{77.6}_{\pm 2.9}$	$74.1_{\pm 3.5}$
	ARID1A mut.	$69.4_{\pm 5.7}$	$76.3_{\pm 5.0}$	$65.9_{\pm 5.4}$	$80.4_{\pm 3.6}$	$66.7_{\pm 5.6}$	$67.5_{\pm 5.3}$	$75.1_{\pm 4.4}$	$\mathbf{83.5}_{\pm 3.6}$
	Immune class	$63.8_{\pm 2.4}$	$62.1_{\pm 2.0}$	$60.7_{\pm 3.1}$	$60.0_{\pm 2.4}$	$62.6_{\pm 3.2}$	$57.8_{\pm 2.6}$	$63.8_{\pm 2.6}$	$\mathbf{64.1}_{\pm 2.5}$
	KRAS mut.	$64.7_{\pm 4.9}$	$\mathbf{68.9}_{\pm 3.4}$	$58.8_{\pm 3.3}$	$61.8_{\pm 3.2}$	$62.7_{\pm 4.3}$	$64.5_{\pm 3.8}$	$67.2_{\pm 3.6}$	$68.4_{\pm 3.2}$
	MSI-H†	$77.8_{\pm 5.4}$	$89.2_{\pm 1.9}$	$88.2_{\pm 2.4}$	$79.4_{\pm 2.9}$	$72.5_{\pm 5.4}$	$86.3_{\pm 2.5}$	$88.7_{\pm 2.1}$	$\mathbf{91.3}_{\pm 2.3}$
	PIK3CA mut.	$59.3_{\pm 4.0}$	$\mathbf{69.8}_{\pm 3.9}$	$49.5_{\pm 3.6}$	$66.2_{\pm 3.7}$	$59.9_{\pm 4.6}$	$64.5_{\pm 3.5}$	$61.9_{\pm 4.1}$	$65.7_{\pm 3.8}$
	SETD1B mut.	$71.9_{\pm 4.8}$	$78.0_{\pm 3.3}$	$75.8_{\pm 3.8}$	$72.1_{\pm 4.4}$	$75.2_{\pm 4.6}$	$79.6_{\pm 4.1}$	$82.3_{\pm 3.4}$	$\mathbf{84.0}_{\pm 3.3}$
	TP53 mut.	$65.9_{\pm 4.8}$	$66.5_{\pm 3.9}$	$60.0_{\pm 3.5}$	$64.6_{\pm 3.7}$	$64.3_{\pm 4.8}$	$70.9_{\pm 2.9}$	$73.6_{\pm 3.2}$	$\mathbf{74.1}_{\pm 3.7}$
CPTAC-LUAD	EGFR mut.	$62.5_{\pm 3.3}$	$77.4_{\pm 3.0}$	$79.0_{\pm 2.8}$	$76.9_{\pm 3.0}$	$70.0_{\pm 5.2}$	$80.6_{\pm 2.5}$	$\mathbf{82.3}_{\pm 2.6}$	$80.6_{\pm 2.6}$
	Immune class	$65.9_{\pm 2.6}$	$69.8_{\pm 2.5}$	$66.2_{\pm 2.3}$	$71.1_{\pm 2.6}$	$62.8_{\pm 2.5}$	$65.6_{\pm 2.2}$	$71.0_{\pm 2.4}$	$\mathbf{74.4}_{\pm 2.0}$
	KRAS mut.	$60.6_{\pm 4.3}$	$71.3_{\pm 3.0}$	$67.9_{\pm 3.5}$	$63.1_{\pm 3.0}$	$66.5_{\pm 4.4}$	$\mathbf{74.8}_{\pm 3.1}$	$73.9_{\pm 2.4}$	$65.3_{\pm 2.7}$
	STK11 mut.	$61.7_{\pm 5.7}$	$84.9_{\pm 2.8}$	$88.8_{\pm 2.3}$	$83.9_{\pm 3.7}$	$76.3_{\pm 5.7}$	$89.9_{\pm 2.1}$	$\mathbf{90.5}_{\pm 2.4}$	$87.9_{\pm 3.0}$
	TP53 mut.	$65.4_{\pm 4.0}$	$70.3_{\pm 2.9}$	$73.0_{\pm 2.9}$	$72.6_{\pm 2.6}$	$63.3_{\pm 3.6}$	$69.6_{\pm 2.7}$	$\mathbf{73.8}_{\pm 2.5}$	$72.6_{\pm 2.8}$
CPTAC-LSCC	ARID1A mut.	$52.1_{\pm 5.0}$	$48.5_{\pm 4.0}$	$48.5_{\pm 4.8}$	$51.7_{\pm 3.9}$	$\mathbf{53.9}_{\pm 5.4}$	$32.6_{\pm 3.7}$	$42.8_{\pm 5.0}$	$46.2_{\pm 4.6}$
	Histologic grade†	$59.9_{\pm 3.8}$	$\mathbf{71.3}_{\pm 2.6}$	$64.9_{\pm 3.6}$	$68.0_{\pm 3.0}$	$66.7_{\pm 3.6}$	$68.1_{\pm 3.1}$	$70.3_{\pm 2.9}$	$68.9_{\pm 3.0}$
	Immune class	$66.5_{\pm 2.6}$	$\mathbf{77.3}_{\pm 2.2}$	$67.7_{\pm 1.8}$	$76.1_{\pm 1.9}$	$65.1_{\pm 3.2}$	$73.8_{\pm 2.2}$	$76.0_{\pm 1.8}$	$76.3_{\pm 2.1}$
	KEAP1 mut.	$63.9_{\pm 5.1}$	$67.7_{\pm 3.2}$	$53.8_{\pm 4.9}$	$\mathbf{70.5}_{\pm 3.1}$	$59.5_{\pm 5.1}$	$59.4_{\pm 3.9}$	$64.9_{\pm 3.9}$	$59.5_{\pm 3.2}$
CPTAC-UCEC	CTNNB1 mut.	$71.7_{\pm 4.6}$	$74.4_{\pm 3.7}$	$73.4_{\pm 3.3}$	$71.8_{\pm 3.5}$	$68.7_{\pm 4.5}$	$75.9_{\pm 3.9}$	$71.6_{\pm 4.0}$	$\mathbf{78.5}_{\pm 3.3}$
	Immune class	$63.2_{\pm 2.8}$	$56.8_{\pm 2.3}$	$\mathbf{69.1}_{\pm 2.5}$	$61.8_{\pm 2.2}$	$63.1_{\pm 2.9}$	$64.6_{\pm 2.4}$	$68.6_{\pm 2.1}$	$65.4_{\pm 2.6}$
	PTEN mut.	$73.6_{\pm 5.2}$	$62.7_{\pm 5.5}$	$\mathbf{77.7}_{\pm 4.2}$	$68.3_{\pm 5.3}$	$67.7_{\pm 5.1}$	$66.6_{\pm 5.2}$	$60.9_{\pm 5.4}$	$69.5_{\pm 4.8}$
CPTAC-CCRCC	BAP1 mut.	$76.5_{\pm 5.6}$	$75.0_{\pm 3.4}$	$71.2_{\pm 3.4}$	$79.2_{\pm 2.9}$	$63.4_{\pm 5.3}$	$72.1_{\pm 3.8}$	$73.6_{\pm 3.8}$	$\mathbf{80.2}_{\pm 2.3}$
	Immune class	$53.8_{\pm 2.9}$	$55.0_{\pm 2.1}$	$\mathbf{56.9}_{\pm 2.8}$	$55.4_{\pm 2.6}$	$55.9_{\pm 2.8}$	$53.1_{\pm 2.6}$	$53.5_{\pm 2.3}$	$54.3_{\pm 2.7}$
	PBRM1 mut.	$58.4_{\pm 3.1}$	$52.2_{\pm 3.2}$	$60.8_{\pm 3.5}$	$57.6_{\pm 3.4}$	$56.5_{\pm 4.1}$	$57.5_{\pm 2.8}$	$\mathbf{63.1}_{\pm 3.0}$	$58.9_{\pm 3.2}$
	VHL mut.	$61.4_{\pm 4.7}$	$\mathbf{65.2}_{\pm 3.6}$	$60.0_{\pm 4.3}$	$59.9_{\pm 4.1}$	$63.2_{\pm 5.0}$	$49.6_{\pm 4.0}$	$57.3_{\pm 3.5}$	$58.3_{\pm 4.1}$
CPTAC-GBM	EGFR mut.	$65.5_{\pm 3.9}$	$59.6_{\pm 3.2}$	$66.6_{\pm 3.7}$	$72.4_{\pm 2.3}$	$67.6_{\pm 4.7}$	$\mathbf{73.4}_{\pm 2.8}$	$70.5_{\pm 2.9}$	$73.0_{\pm 2.4}$
	Immune class	$67.8_{\pm 3.1}$	$72.4_{\pm 2.5}$	$\mathbf{79.7}_{\pm 2.2}$	$76.7_{\pm 2.1}$	$68.7_{\pm 3.4}$	$77.2_{\pm 2.6}$	$76.0_{\pm 2.3}$	$76.1_{\pm 2.1}$
	TP53 mut.	$67.6_{\pm 4.6}$	$78.0_{\pm 2.7}$	$83.5_{\pm 2.5}$	$\mathbf{86.0}_{\pm 2.6}$	$69.5_{\pm 4.7}$	$85.9_{\pm 2.0}$	$84.8_{\pm 2.4}$	$84.7_{\pm 2.6}$
CPTAC-HNSC	CASP8 mut.	$69.2_{\pm 6.4}$	$58.0_{\pm 5.2}$	$60.4_{\pm 6.5}$	$67.6_{\pm 5.1}$	$45.4_{\pm 7.3}$	$63.7_{\pm 4.8}$	$72.6_{\pm 4.9}$	$\mathbf{73.8}_{\pm 4.8}$
	Histologic grade†	$63.6_{\pm 3.4}$	$76.7_{\pm 3.2}$	$77.1_{\pm 2.6}$	$\mathbf{78.2}_{\pm 2.9}$	$63.2_{\pm 4.0}$	$76.0_{\pm 3.4}$	$76.5_{\pm 2.9}$	$73.6_{\pm 3.0}$
	Immune class	$62.1_{\pm 2.9}$	$64.4_{\pm 2.0}$	$69.3_{\pm 2.2}$	$68.4_{\pm 2.1}$	$63.1_{\pm 3.3}$	$66.0_{\pm 2.0}$	$70.3_{\pm 2.1}$	$\mathbf{73.0}_{\pm 2.4}$
CPTAC-OV	Immune class	$59.2_{\pm 3.6}$	$51.1_{\pm 3.0}$	$57.1_{\pm 3.1}$	$55.7_{\pm 3.2}$	$55.9_{\pm 3.7}$	$58.1_{\pm 3.3}$	$59.1_{\pm 3.9}$	$\mathbf{59.4}_{\pm 3.6}$
CPTAC-PDA	Immune class	$56.5_{\pm 2.8}$	$53.9_{\pm 2.4}$	$\mathbf{58.9}_{\pm 2.6}$	$50.2_{\pm 2.6}$	$58.5_{\pm 3.1}$	$54.2_{\pm 2.7}$	$57.2_{\pm 2.6}$	$49.1_{\pm 2.1}$
	SMAD4 mut.	$58.6_{\pm 4.6}$	$41.5_{\pm 4.1}$	$49.0_{\pm 4.3}$	$57.4_{\pm 3.5}$	$52.5_{\pm 4.6}$	$52.9_{\pm 4.2}$	$55.3_{\pm 4.1}$	$\mathbf{60.5}_{\pm 4.0}$

Table 8: Survival analysis results on 3 CPTAC overall-survival tasks. We report c-index with 95% confidence intervals. The best result for each task is bolded.

Cohort	Task	CHIEF	GigaPath	PRISM	Madeleine	FEATHER	TITAN	CARE	**FACET**
CPTAC-LUAD	OS	$45.2_{\pm 9.6}$	$49.8_{\pm 9.1}$	$66.3_{\pm 12.2}$	$51.9_{\pm 11.4}$	$\mathbf{66.7}_{\pm 7.4}$	$59.1_{\pm 13.1}$	$65.7_{\pm 13.3}$	$66.5_{\pm 16.1}$
CPTAC-HNSC	OS	$48.1_{\pm 9.7}$	$\mathbf{61.6}_{\pm 17.8}$	$52.5_{\pm 18.1}$	$52.6_{\pm 19.2}$	$52.5_{\pm 23.9}$	$59.1_{\pm 21.4}$	$54.1_{\pm 15.2}$	$59.3_{\pm 15.7}$
CPTAC-PDA	OS	$51.0_{\pm 13.1}$	$50.6_{\pm 11.0}$	$55.8_{\pm 7.6}$	$56.9_{\pm 10.9}$	$51.3_{\pm 10.2}$	$55.7_{\pm 9.8}$	$53.8_{\pm 9.0}$	$\mathbf{57.3}_{\pm 9.3}$

Table 9: Architectural ablations on the pretraining validation set. Performance is reported as mean AUC across 25 classification tasks.

Layers	Hidden dimension	Attention heads	AUC (%)
1	256	4	87.4 ± 3.9
1	512	4	87.1 ± 4.0
1	512	8	87.7 ± 3.8
1	768	4	86.7 ± 4.4
1	768	8	87.0 ± 4.3
2	256	4	87.6 ± 3.9
2	512	4	85.7 ± 4.8
2	512	8	85.4 ± 5.1
3	256	4	85.9 ± 5.3
3	512	4	85.8 ± 5.0
3	512	8	86.4 ± 4.7

Table 10: Pretraining hyperparameter ablations on the validation set. Performance is reported as mean AUC across 25 classification tasks.

Epochs	Batch size	Learning rate	AUC (%)
50	8	1×10^{-5}	87.6 ± 3.9
30	8	1×10^{-5}	87.9 ± 3.6
30	4	1×10^{-5}	87.2 ± 4.5
30	16	1×10^{-5}	86.8 ± 4.2
50	8	5×10^{-5}	87.7 ± 4.0
50	8	1×10^{-4}	87.4 ± 4.6
30	8	5×10^{-4}	87.2 ± 5.2

of H&E images[3]. These annotations are used only to characterize the tissue types receiving attention and are not provided to FACET during training or inference.

E.3 Clinical Context Case Study

Prior studies have demonstrated that integrating patient-level clinical context with WSI-derived representations can improve predictive performance (Cai et al., 2024; Shirae et al., 2025; Schmauch et al., 2024). In this case study, we evaluate whether FACET can incorporate clinical context through its task description without modifying the model parameters. This highlights a practical advantage of FACET's task-conditioning interface: clinically informative patient-level priors can be introduced directly through text, without requiring an additional fusion architecture or retraining. We use smoking status as a clinically relevant prior because of its established association with `EGFR` and `TP53` mutation prevalence in lung cancer (Devarakonda et al., 2021; Mounawar et al., 2007). Here, smoking status is treated as an additional source of predictive information alongside the WSI, rather than to establish a causal connection between smoking and morphology.

For patients with known smoking status, the original task description is augmented with either "The patient is a smoker." or "The patient is not a smoker.". For patients with unknown smoking status, the original task description is left unchanged. Thus, the experiment modifies only the task representation while keeping the slide, model weights, and downstream evaluation procedure fixed.

The smoking-status distribution is shown in Table 11. In both cohorts, smoking status is available for the majority of slides, and slides from non-smokers constitute the larger group. The broadly similar class distributions provide a comparable setting for examining the effect of augmenting the task description with smoking status.

[3] `https://huggingface.co/datasets/Aignostics/OpenTME/blob/main/data/stomach_cancer/thumbnails/TCGA-HU-A4GQ-01Z-00-DX1.699A0E57-FE02-4329-9826-3EFC2DECD408/tissue_segmentation.png`

Table 11: Smoking-status distribution across slides for the clinical-context case study. Slides with missing smoking status use the original task description without additional clinical context.

Cohort	Smoker	Non-smoker	Missing
CPTAC-GBM	69	303	107
CPTAC-LUAD	122	238	48