# M2U: MULTIMODAL-TO-UNIMODAL CROSS-MODAL GEOMETRY TRANSFER FOR CANCER PROGNOSIS

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Multimodal models typically assume that the same modalities are available during training and deployment. In practice, however, some modalities may be available in training cohorts but missing or prohibitively costly for new patients. We study this multimodal-to-unimodal setting in cancer prognosis, where paired histopathology and molecular omics are available during training, but predictions at deployment must be made from histopathology alone. Existing approaches typically distill predictions or features, reconstruct the missing modality, or align paired samples. Here, we introduce M2U, a framework that transfers patient-to-patient similarity structure from the privileged molecular space into the histopathology space and makes this structure queryable at deployment. Its Mapping Block introduces X-GeoSCL, a cross-modal objective that uses molecular representations to define omics neighborhoods and trains histopathology representations to reproduce these soft relational targets rather than merely matching individual pairs. Its Reasoning Block uses a query histopathology representation to retrieve patients from a train-only paired gallery, reads their stored molecular embeddings through cross-attention, and combines the recovered molecular evidence with direct histopathology predictions through a learned gate. We evaluate M2U on four TCGA cancer cohorts: kidney renal clear cell carcinoma, lung adenocarcinoma, brain lower-grade glioma, and pancreatic adenocarcinoma. The complete M2U framework achieves the highest mean AUC among all deployment-compatible methods on three of the four cohorts and outperforms both evaluated multimodal-to-unimodal baselines on all four. Within the Mapping Block, X-GeoSCL achieves the highest mean downstream AUC among the evaluated mapping objectives across all four cohorts. Ablations further show that the Mapping and Reasoning Blocks play complementary roles: transferring molecular neighborhood geometry enables cross-modal retrieval, but effectively exploiting this structure requires reading the retrieved molecular representations and integrating them with direct histopathology predictions. M2U can be applied to other settings in which a privileged modality is available during training but unavailable at deployment.

## 1 INTRODUCTION

Multimodal models that combine whole-slide histopathology (WSI) with molecular omics can improve cancer prognosis by integrating morphological and molecular evidence (Chen et al., 2021; 2022; Xu and Chen, 2023; Zhou and Chen, 2023; Jaume et al., 2024a; Song et al., 2024). Their standard deployment assumption, however, is restrictive: the same modalities used during training must also be available for each new patient. In many settings, histology is available while molecular profiling is absent, delayed, or acquired only for a subset of patients. This motivates *multimodal-to-unimodal* learning: paired modalities are available during training, but prediction is made from the deployment modality without query-patient omics.

Existing approaches commonly transfer predictions, distill features, reconstruct the missing modality, or maintain a memory of multimodal examples (Guo et al., 2026; Qu et al., 2025). We study a different transferable object: the *relational geometry* induced by the privileged molecular representation. Molecular embeddings do not only describe individual patients; their relative distances define neighborhoods that can encode continuous structure discarded by binary prognostic labels. If this



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.

structure can be transferred into WSI space, a WSI query can retrieve training patients with relevant molecular context even though its own omics are unavailable.

We introduce M2U, a two-stage framework for this setting. The **Mapping Block** uses X-GeoSCL, a soft cross-modal contrastive objective whose targets are Gaussian-weighted neighborhoods in a frozen molecular representation. Unlike pairwise contrastive or similarity-preserving distillation objectives, X-GeoSCL uses the privileged modality to define a normalized neighborhood distribution over multiple patients and trains the deployment representation to reproduce this relational structure. The **Reasoning Block** retrieves a query's nearest mapped WSI examples from a train-only paired gallery and reads their corresponding molecular embeddings with cross-attention. A scalar gate then combines this retrieval-based prediction with a direct WSI predictor. Thus, the method separates two questions that are often conflated: how privileged structure is *recovered* from WSI, and how the recovered evidence is *routed* into the final prediction.

We evaluate the framework on four distinct TCGA cancer cohorts, kidney renal clear cell carcinoma (KIRC), lung adenocarcinoma (LUAD), brain lower-grade glioma (LGG), and pancreatic adenocarcinoma (PAAD), which exhibit materially different unimodal and retrieval behaviors. M2U has the highest mean AUC among the evaluated WSI-only and multimodal-to-unimodal methods on KIRC, LUAD, and PAAD, and is the strongest multimodal-to-unimodal method on LGG. The reasoning analysis reveals two broad operating patterns. On KIRC and LUAD, the retrieval reader is stronger than the unimodal WSI predictor and receives high gate weight. On PAAD and LGG, the unimodal WSI predictor is stronger than the retrieval reader and the learned retrieval weight falls to about 0.49; PAAD still obtains a modest fusion gain, whereas LGG largely preserves an already strong WSI signal. Crucially, transferring privileged geometry is not sufficient to obtain strong predictions with a simple neighbor-label reader: the transferred structure must also be recovered through the retrieved molecular representations and appropriately combined with direct WSI evidence.

Our contributions are:

1. **Cross-modal geometry transfer.** We formulate molecular patient-to-patient similarity structure as privileged supervision and introduce X-GeoSCL, a cross-modal objective that transfers soft molecular neighborhoods into WSI representations rather than reducing supervision to paired correspondence or prediction mimicry.
2. **Retrieval-based multimodal-to-unimodal reasoning.** We introduce M2U, which uses a WSI query to retrieve patients from a train-only paired gallery, reads their stored molecular embeddings through cross-attention, and integrates the recovered molecular evidence with a direct WSI predictor, without requiring query-patient omics.
3. **Consistent evidence across distinct cancer cohorts.** Across four TCGA cohorts, X-GeoSCL achieves the highest mean downstream AUC among the evaluated mapping objectives, while the complete M2U framework outperforms both deployment-matched multimodal-to-unimodal baselines in mean AUC on all four cohorts and achieves the best deployment-compatible mean AUC on three.
4. **A preservation, recovery, routing view of privileged transfer.** Ablations separate three complementary roles: X-GeoSCL preserves privileged molecular structure in the WSI space, cross-attention recovers information from the retrieved molecular representations, and the gate routes this evidence relative to the direct WSI prediction. Patient-level analyses further reveal heterogeneous benefits across cohorts, motivating patient-adaptive routing.

## 2 Related Work

**Histology and multimodal cancer prognosis.** Attention-based multiple-instance learning (MIL) is a standard approach for aggregating gigapixel WSIs (Ilse et al., 2018; Lu et al., 2021; Shao et al., 2021). Multimodal extensions combine histology with genomic or pathway representations through co-attention, optimal-transport matching, cross-modal translation, pathway-level interaction, and prototypical compression (Chen et al., 2021; 2022; Xu and Chen, 2023; Zhou and Chen, 2023; Jaume et al., 2024a; Song et al., 2024). These methods primarily optimize multimodal prediction when both modalities are available to the model. Our target setting instead uses paired WSI-omics examples for learning but does not observe omics for the query patient.

**Missing-modality and multimodal-to-unimodal learning.** General missing-modality methods such as MultiModN, ShaSpec, and HEALNet are designed to remain functional under different patterns of modality availability (Swamy et al., 2023; Wang et al., 2023; Hemker et al., 2024). Pathology-specific approaches include PathoGen-X, which translates and aligns histology features toward genomic space (Krishna et al., 2025); ProSurv, which uses modality-specific prototypes for cross-modal feature completion (Liu et al., 2025); MoMKD, which distills genomic supervision into histology through momentum memory (Guo et al., 2026); and M$^2$Surv, which uses a memory mechanism to compensate for incomplete pathology-genomic inputs (Qu et al., 2025). MoMKD and M$^2$Surv are the deployment-matched baselines available under our experimental protocol. M2U differs in what it transfers and how it uses it: the privileged modality defines a neighborhood distribution, and prediction reads continuous molecular embeddings retrieved from a train-only paired gallery rather than reconstructing the missing query modality.

**Cross-modal representation and relational transfer.** Cross-modal pathology representation learning has shown that molecular supervision can shape slide representations. TANGLE aligns WSI and transcriptomic embeddings through paired contrastive learning (Jaume et al., 2024b), while CPathomic combines cross-modal contrastive alignment with attention-based multimodal interaction (Li et al., 2025). More generally, InfoNCE learns paired correspondence (van den Oord et al., 2018), contrastive representation distillation and relational distillation transfer similarities between examples (Tian et al., 2020; Park et al., 2019; Tung and Mori, 2019), and soft contrastive estimation replaces hard instance targets with similarity-weighted distributions (Denize et al., 2023). X-GeoSCL builds on these principles but uses a frozen molecular representation to define a full Gaussian-weighted target neighborhood for each WSI anchor. Thus, the supervision preserves relative molecular relationships beyond the paired positive.

**Retrieval-augmented prediction.** Learning with privileged information formalizes the use of richer training-time information that is absent for a test example (Vapnik and Vashist, 2009). Non-parametric retrieval can augment parametric prediction, as in $k$-NN language models (Khandelwal et al., 2020), while attention-based support-set readers provide a learned alternative to uniform neighbor voting (Hou et al., 2019). M2U combines these ideas in a cross-modal setting: retrieval is keyed by the mapped WSI representation, but the reader operates on molecular embeddings associated with the retrieved training examples.

# 3 METHOD

## 3.1 PROBLEM FORMULATION

Let $\mathcal{D} = \{(W_i, O_i, y_i)\}_{i=1}^{N}$ denote paired training data. $W_i = \{h_{ip}\}_{p=1}^{P_i}$ is a variable-size WSI bag of frozen 1536-dimensional UNI patch features (Chen et al., 2024), $O_i$ is the raw multi-omics profile, and $y_i \in \{0, 1\}$ is the fixed-horizon prognosis label defined in Section 4. M2U uses two WSI encoders with the same AMIL architecture (Ilse et al., 2018) but independent parameters and objectives, together with a molecular SNN (Klambauer et al., 2017):

$$w_i^{\mathrm{uni}} = f_\theta^{\mathrm{uni}}(W_i), \qquad w_i^{\mathrm{map}} = f_\psi^{\mathrm{map}}(W_i), \qquad o_i = g_\phi(O_i) \in \mathbb{R}^{256}. \tag{1}$$

$f_\theta^{\mathrm{uni}}$ is trained with the prognosis objective and supplies the direct WSI prediction. $g_\phi$ is trained within each fold and then frozen, defining the privileged molecular representation space. The mapping encoder $f_\psi^{\mathrm{map}}$ is trained independently with X-GeoSCL. Thus, M2U does not overwrite the unimodal WSI representation: both WSI encoders are retained at inference, one for direct prediction and one for retrieval-based reasoning.

M2U does not reconstruct the query patient's raw molecular profile. Instead, it transfers relational structure from the frozen molecular space into the separately trained WSI mapping space. Prediction uses only the query WSI together with a *train-only paired gallery*; no held-out patient enters the gallery and no query-patient omics are used.

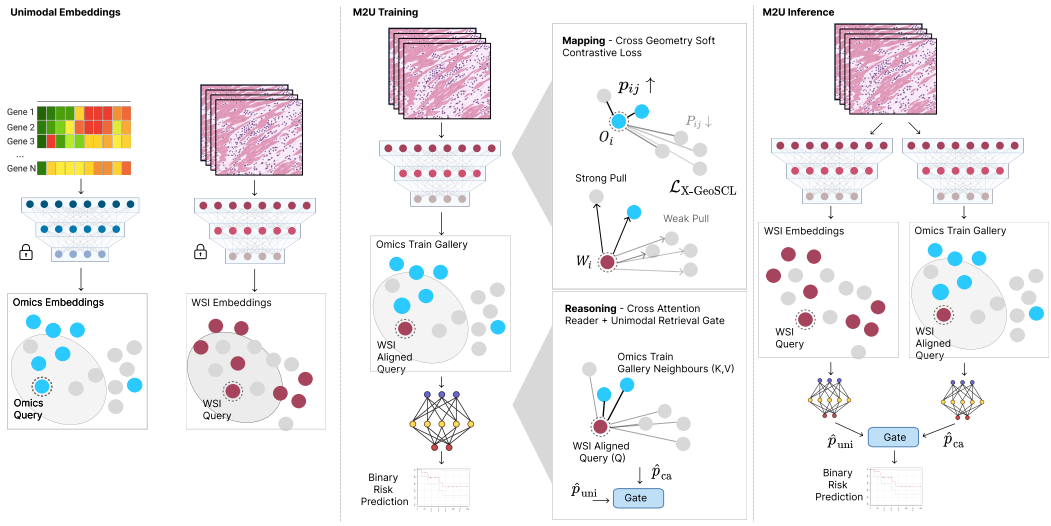



Figure 1: **M2U overview. Left:** Modality-specific encoders produce patient-level unimodal representations, here, WSI and molecular representations, whose relative distances induce patient neighborhoods. **Middle:** During training, the Mapping Block uses X-GeoSCL to transfer neighborhood structure from the privileged modality into an independently trained deployment-modality mapping space. In our application, molecular neighborhoods supervise the WSI mapping space. The resulting representations support retrieval from a train-only paired gallery, while the Reasoning Block learns to read the retrieved privileged-modality embeddings through cross-attention and combine the resulting prediction with the direct unimodal prediction through a learned gate. **Right:** At inference, the query's available modality, here, WSI, is used for both direct prediction and retrieval of paired training examples. Their stored molecular embeddings are read by the Reasoning Block and integrated with the direct WSI prediction. No query-patient omics are required.

### 3.2 Mapping Block: X-GeoSCL

The Mapping Block trains $f_\psi^{\text{map}}$ to preserve relational structure from the frozen molecular representation. We $\ell_2$-normalize the mapped WSI and molecular embeddings, writing $\hat{w}_i^{\text{map}} = \text{norm}(f_\psi^{\text{map}}(W_i))$ and $\hat{o}_i = \text{norm}(o_i)$. Following the probabilistic-neighborhood construction of stochastic neighbor embedding (Hinton and Roweis, 2002), each molecular anchor defines a Gaussian neighborhood distribution

$$p_{ij} = \frac{\exp(-\|\hat{o}_i - \hat{o}_j\|_2^2/\sigma^2)}{\sum_{k\neq i}\exp(-\|\hat{o}_i - \hat{o}_k\|_2^2/\sigma^2)}, \qquad p_{ii} = 0, \tag{2}$$

where $\sigma$ is the median pairwise molecular $\ell_2$ distance in the batch, analogous to the standard median bandwidth heuristic for Gaussian kernels (Gretton et al., 2012). This converts continuous molecular distances into a normalized distribution over neighboring patients rather than selecting discrete positive pairs.

The mapped WSI anchor induces a distribution over the same molecular support,

$$q_{ij} = \frac{\exp((\hat{w}_i^{\text{map}})^\top \hat{o}_j/\tau)}{\sum_{k\neq i}\exp((\hat{w}_i^{\text{map}})^\top \hat{o}_k/\tau)}, \tag{3}$$

where $\tau$ controls the sharpness of the WSI-induced distribution. X-GeoSCL transfers the privileged neighborhood distribution by minimizing

$$\mathcal{L}_{\text{X-GeoSCL}} = -\frac{1}{B}\sum_{i=1}^{B}\sum_{j\neq i} p_{ij}\log q_{ij}. \tag{4}$$

For a fixed batch, $-\sum_j p_{ij}\log q_{ij} = H(p_i) + \text{KL}(p_i\|q_i)$. Because $p_i$ depends only on frozen molecular embeddings, minimizing Eq. 4 with respect to $\psi$ is equivalent to minimizing the KL

term for those fixed targets. The probabilistic-neighborhood construction is related to SNE, but here a frozen privileged modality defines the target geometry for a separate deployment-modality representation. Unlike pairwise InfoNCE (van den Oord et al., 2018), X-GeoSCL therefore supervises the WSI encoder with the full soft molecular neighborhood distribution. We compare against InfoNCE and Soft-InfoNCE (Denize et al., 2023) under the same downstream reader.

### 3.3 Reasoning Block and unimodal-retrieval routing

After mapping, $f_\psi^{\mathrm{map}}$ is frozen and the training fold forms a paired gallery

$$\mathcal{B} = \{(\hat{w}_j^{\mathrm{map}}, \hat{o}_j, y_j)\}_{j \in \mathrm{train}}. \tag{5}$$

For a query $W_q$, M2U computes $\hat{w}_q^{\mathrm{map}} = \mathrm{norm}(f_\psi^{\mathrm{map}}(W_q))$ and retrieves its $K$ nearest gallery entries by cosine similarity. Retrieval is keyed only by WSI; molecular information enters only through the stored molecular embeddings of the retrieved training patients. A cross-attention reader uses the mapped WSI as query and retrieved molecular embeddings as keys and values:

$$\alpha_q = \mathrm{softmax}\left(\frac{Q(\hat{w}_q^{\mathrm{map}})K(\hat{O}_{\mathcal{N}_q})^\top}{\sqrt{d}}\right), \qquad z_q = \sum_{j \in \mathcal{N}_q} \alpha_{qj} V(\hat{o}_j). \tag{6}$$

The reader forms $u_q = [\hat{w}_q^{\mathrm{map}}; z_q; \hat{w}_q^{\mathrm{map}} - z_q; \hat{w}_q^{\mathrm{map}} \odot z_q]$ and predicts $\hat{p}_{\mathrm{CA}} = h_{\mathrm{CA}}(u_q)$. This is richer than uniform $K$-NN label voting because it operates on the continuous privileged representations rather than reducing the neighborhood to binary labels.

The direct WSI branch independently predicts $\hat{p}_{\mathrm{WSI}} = h_{\mathrm{uni}}(f_\theta^{\mathrm{uni}}(W_q))$. After the cross-attention reader is trained and frozen, a cohort-level scalar gate is fitted on validation predictions:

$$\hat{p} = (1-\lambda)\hat{p}_{\mathrm{WSI}} + \lambda \hat{p}_{\mathrm{CA}}, \qquad \lambda \in [0,1]. \tag{7}$$

The gate is soft fusion rather than a hard selector. Importantly, its two inputs originate from independently trained WSI encoders with different objectives, so the routing analysis compares genuinely distinct sources of evidence.

**Algorithm 1** M2U training and inference without query-patient omics

**Require:** Paired training data $\{(W_i, O_i, y_i)\}$ and neighborhood size $K$

1: Train $f_\theta^{\mathrm{uni}}$ with its prognosis head; train $g_\phi$, compute $o_i$, and freeze $g_\phi$.
2: Train an independent AMIL encoder $f_\psi^{\mathrm{map}}$ with X-GeoSCL (Eqs. 2-4).
3: Freeze $f_\psi^{\mathrm{map}}$ and build the train-only gallery $\mathcal{B}$ (Eq. 5).
4: Train and freeze the cross-attention reader; fit $\lambda$ on validation predictions.
**Inference:**
5: From query WSI $W_q$, compute $\hat{p}_{\mathrm{WSI}}$ and $\hat{w}_q^{\mathrm{map}}$.
6: Retrieve $K$ mapped-WSI neighbors, read their stored molecular embeddings with Eq. 6, and obtain $\hat{p}_{\mathrm{CA}}$.
7: Return the gated prediction from Eq. 7.

Additional encoder, optimization, temperature, and architectural details are provided in Appendix A.2.

## 4 Experimental Setup

We evaluate four TCGA cohorts with paired WSI and multi-omics data (TCGA Research Network et al., 2013): KIRC (447 patients, 162 events), LUAD (412, 173), LGG (445, 114), and PAAD (165, 99). Molecular input combines RNA expression, copy-number alterations, and somatic mutations (approximately 15k features). All results use five-fold patient-level cross-validation. For each fold, the binary prognosis horizon $\tau$ is the Kaplan-Meier median estimated from training-fold event times only (Kaplan and Meier, 1958); observed events at or before $\tau$ are high-risk, patients known to survive beyond $\tau$ are low-risk, and patients censored before $\tau$ are excluded. The horizon,

preprocessing, learned models, and paired retrieval gallery are derived from training data only. We use $K = 10$ for retrieval. Appendix E reports a neighborhood-purity sensitivity check across $K \in \{3, 5, 7, 10, 15, 20\}$; this diagnostic tests neighborhood composition rather than predictive performance.

WSI-only baselines are WSI $K$-NN retrieval, AMIL (Ilse et al., 2018), and TransMIL (Shao et al., 2021). Deployment-matched multimodal-to-unimodal baselines are MoMKD (Guo et al., 2026) and M$^2$Surv (Qu et al., 2025); omics retrieval and the SNN predictor are privileged-information references. For mapping, we compare X-GeoSCL with InfoNCE (van den Oord et al., 2018) and Soft-InfoNCE (Denize et al., 2023) under the same downstream reader. AUC is the primary metric. IPCW-Brier (Graf et al., 1999), AUPRC, and balanced accuracy are secondary metrics; IPCW-Brier provides censoring-aware supporting evidence at the same prediction horizon. Full implementation, leakage-control, and secondary-metric details are provided in the appendix.

# 5 Results

## 5.1 Main predictive performance

Table 1 compares deployment-compatible WSI-only and multimodal-to-unimodal methods. M2U attains the highest mean AUC among these methods on KIRC (0.683), LUAD (0.542), and PAAD (0.747). On LGG, where AMIL is already strong (0.819), M2U reaches 0.802 and remains the strongest evaluated multimodal-to-unimodal method. Across all four cohorts, M2U exceeds both MoMKD and M$^2$Surv in mean AUC.

Table 1: **Main predictive performance across four TCGA cohorts.** AUC is mean±SEM over five folds. Bold marks the best deployment-compatible result; omics-only rows are privileged references.

| Setting | Method | KIRC | LUAD | LGG | PAAD |
|---|---|---|---|---|---|
| WSI-only | WSI retrieval | 0.568 ±0.028 | 0.479 ±0.023 | 0.791 ±0.026 | 0.657 ±0.114 |
| | AMIL (Ilse et al., 2018) | 0.576 ±0.021 | 0.429 ±0.052 | **0.819** ±0.032 | 0.733 ±0.093 |
| | TransMIL (Shao et al., 2021) | 0.649 ±0.013 | 0.522 ±0.042 | 0.645 ±0.081 | 0.543 ±0.086 |
| M2U | MoMKD (Guo et al., 2026) | 0.663 ±0.035 | 0.439 ±0.046 | 0.633 ±0.054 | 0.707 ±0.088 |
| | M$^2$Surv (Qu et al., 2025) | 0.681 ±0.034 | 0.488 ±0.054 | 0.617 ±0.076 | 0.702 ±0.101 |
| | M2U (ours) | **0.683** ±0.013 | **0.542** ±0.052 | 0.802 ±0.022 | **0.747** ±0.100 |
| Privileged ref. | Omics retrieval | 0.711 ±0.011 | 0.606 ±0.015 | 0.891 ±0.044 | 0.834 ±0.035 |
| | SNN (Klambauer et al., 2017) | 0.748 ±0.020 | 0.619 ±0.024 | 0.891 ±0.032 | 0.798 ±0.037 |

Figure 2 provides censoring-aware supporting evidence. M2U has lower mean IPCW-Brier than the strongest WSI-only comparator on KIRC (0.240 vs. 0.364), LUAD (0.236 vs. 0.316), and PAAD (0.191 vs. 0.222). The paired bootstrap 95% confidence interval for the Brier improvement excludes zero on KIRC and LUAD, while PAAD is borderline at the reported precision. On LGG, WSI retrieval is slightly lower (0.126 vs. 0.135), consistent with the already-strong unimodal WSI result. Secondary metrics show the same broad pattern: M2U improves AUPRC over AMIL and WSI retrieval on all four cohorts and improves balanced accuracy on KIRC, LUAD, and PAAD, while LGG remains WSI-strong (Appendix B).

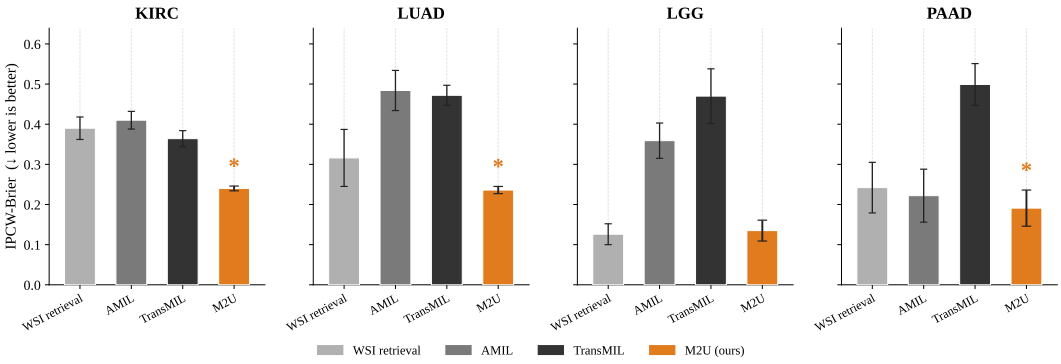



Figure 2: **Censoring-aware fixed-horizon prediction error.** IPCW-Brier (lower is better) for WSI-only baselines and M2U. Error bars show fold-level SEM. Stars indicate that the patient-level bootstrap 95% confidence interval for M2U improvement over the strongest WSI-only comparator excludes zero.

### 5.2 Cross-modal geometry transfer

Keeping the downstream CA+Gate reader fixed, Table 2 compares InfoNCE (van den Oord et al., 2018), Soft-InfoNCE (Denize et al., 2023), and X-GeoSCL. X-GeoSCL achieves the numerically highest mean downstream AUC in all four cohorts: 0.683 on KIRC, 0.542 on LUAD, 0.802 on LGG, and 0.747 on PAAD. The magnitude of the difference varies across cohorts, including only a small margin over InfoNCE on PAAD (0.747 vs. 0.745).

Table 2: **Cross-modal mapping quality across four TCGA cohorts.** Mapping objectives use the same downstream CA+Gate reader; reported metrics are downstream AUC, neighborhood KL, and Purity@10.

| Metric | Mapping objective | KIRC | LUAD | LGG | PAAD |
|---|---|---|---|---|---|
| AUC ↑ | InfoNCE | 0.661 ±0.018 | 0.504 ±0.036 | 0.787 ±0.032 | 0.745 ±0.101 |
| | Soft-InfoNCE | 0.657 ±0.027 | 0.526 ±0.044 | 0.785 ±0.034 | 0.714 ±0.113 |
| | X-GeoSCL (ours) | **0.683** ±0.013 | **0.542** ±0.052 | **0.802** ±0.022 | **0.747** ±0.100 |
| KL ↓ | InfoNCE | 0.251 ±0.100 | 0.559 ±0.049 | **0.334** ±0.062 | 0.431 ±0.065 |
| | Soft-InfoNCE | 0.203 ±0.020 | 0.354 ±0.072 | 0.365 ±0.087 | 0.546 ±0.114 |
| | X-GeoSCL (ours) | **0.098** ±0.024 | **0.215** ±0.120 | 0.352 ±0.091 | **0.310** ±0.094 |
| Purity@10 | InfoNCE | 0.528 ±0.012 | 0.540 ±0.023 | 0.676 ±0.051 | 0.500 ±0.010 |
| | Soft-InfoNCE | 0.519 ±0.011 | 0.523 ±0.014 | 0.682 ±0.048 | 0.533 ±0.018 |
| | X-GeoSCL (ours) | 0.522 ±0.011 | 0.524 ±0.015 | 0.683 ±0.046 | 0.503 ±0.016 |

The geometry diagnostics qualify the AUC result. X-GeoSCL substantially lowers neighborhood KL on KIRC and LUAD and also gives the lowest KL on PAAD, while Purity@10 changes little across objectives. LGG is an informative counterexample: InfoNCE has the lowest KL (0.334 vs. 0.352), yet X-GeoSCL has the highest downstream AUC. Thus, KL measures how closely WSI neighborhoods reproduce the privileged geometry but is not a standalone performance proxy; likewise, label purity does not explain the downstream ranking. The result supports the central distinction between preserving continuous molecular structure and merely improving discrete neighbor-label agreement.

### 5.3 Recovering and routing transferred information

Figure 3 separates geometry transfer from the mechanism used to read and route the transferred information. $K$-NN label voting over the mapped neighborhood is weak, whereas cross-attention

provides the main recovery step by reading the retrieved molecular embeddings; the final gate then combines this readout with the direct WSI prediction.

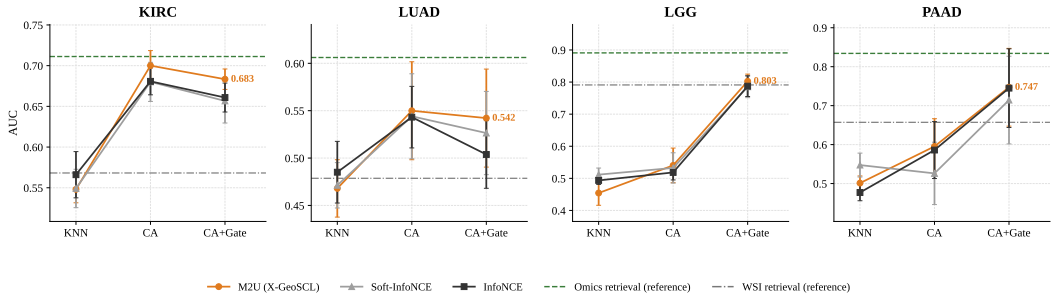



Figure 3: **Reasoning and routing ablation.** AUC as reasoning progresses from $K$-NN label voting to the cross-attention reader (CA) and CA+Gate. Horizontal lines show WSI- and omics-retrieval references.

The branch scores and gate weights reveal two broad routing patterns (Appendix C). On KIRC and LUAD, CA is stronger than AMIL (0.700 vs. 0.576; 0.550 vs. 0.429) and receives high retrieval weight ($\lambda = 0.826, 0.812$); adding the weaker WSI branch slightly lowers final AUC to 0.683 and 0.542. On PAAD and LGG, AMIL is stronger than CA and $\lambda \approx 0.49$. PAAD nevertheless improves from AMIL 0.733 and CA 0.595 to 0.747 after fusion, whereas on LGG gating raises CA from 0.540 to 0.802, largely preserving the strong AMIL predictor (0.819). Similar gate weights therefore need not imply similar outcomes: branch strength and complementarity also matter.

Patient-level behavior is likewise heterogeneous. Relative to AMIL, M2U improves high-risk accuracy from 0.689 to 0.753 on KIRC and from 0.831 to 0.938 on LUAD; PAAD instead improves low-risk accuracy from 0.602 to 0.821 while reducing high-risk accuracy, and LGG is approximately unchanged. The benefiting subgroup is therefore not fixed across cohorts, motivating patient-adaptive routing; full subgroup results are reported in Appendix D.

## 6 Discussion

**Cross-modal geometry as privileged information.** The central result is not simply that multimodal supervision can improve a unimodal predictor; the *form* of that supervision matters. X-GeoSCL gives the strongest downstream mapping performance on all four cohorts, while neighborhood label purity changes little and can favor weaker mappings. LGG further shows that neighborhood KL is a diagnostic rather than a performance proxy: InfoNCE attains lower KL than X-GeoSCL although X-GeoSCL gives higher downstream AUC. Together, these observations support continuous molecular neighborhood structure-not discrete neighbor labels alone-as the useful privileged signal. The result also clarifies the role of the Gaussian target distribution: it is a mechanism for preserving relative molecular relationships, not a claim that any single geometric diagnostic must monotonically predict downstream accuracy.

**Recovery and routing are distinct.** Geometry transfer does not make label voting sufficient. Cross-attention recovers information from continuous retrieved molecular embeddings, and the gate controls how strongly it affects the direct WSI prediction. On KIRC and LUAD, CA is stronger than AMIL and receives high retrieval weight. On PAAD and LGG, AMIL is stronger and the retrieval weight is about 0.49; PAAD still gains from fusion (0.733 to 0.747), whereas LGG recovers from CA 0.540 to 0.802 while largely preserving an already strong WSI predictor (0.819). The important distinction is therefore not the gate value in isolation, but the relative quality and complementarity of the two branches. This separates two failure modes that are otherwise easy to conflate: transferred geometry may be difficult to read, or recovered molecular evidence may add little beyond what WSI already captures.

**Patient-level heterogeneity.** The cohort-level gate summarizes a behavior that is heterogeneous within a cohort. KIRC and LUAD improve mainly on high-risk accuracy, PAAD improves low-risk accuracy while losing high-risk accuracy, and LGG is approximately unchanged across risk groups. These patterns do not support a universal patient subgroup that benefits from privileged transfer. Instead, they motivate patient-adaptive routing as a direct extension of the current cohort-level fusion mechanism.

**Implications for multimodal-to-unimodal learning.** M2U separates three requirements that are often conflated in multimodal-to-unimodal learning: privileged structure must be preserved, recovered by an appropriate reader, and integrated relative to the unimodal evidence already available at deployment. This perspective shifts the problem from transferring a single multimodal prediction function to transferring relational information that can be queried when the privileged modality is unavailable for the test patient.

**Limitations and future directions.** M2U currently uses a single cohort-level routing weight, and therefore cannot adapt the balance between direct WSI evidence and retrieved molecular evidence to individual patients. The heterogeneous subgroup behavior, particularly in PAAD, motivates patient-adaptive routing as a natural extension. Second, we evaluate only the WSI←omics transfer direction; although the formulation is not inherently directional, transfer in the reverse direction remains untested. Finally, the present study considers a single privileged modality. Extending M2U to multiple privileged modalities will require principled mechanisms for constructing and reconciling potentially different relational geometries.

## 7 Conclusion

We introduced M2U, a multimodal-to-unimodal framework that transfers patient-to-patient similarity structure from a privileged modality into the deployment-modality representation and makes this structure queryable at inference. We instantiated M2U for cancer prognosis, using molecular neighborhood geometry as privileged supervision for histopathology-based prediction. Across four cancer cohorts, M2U achieves the highest mean AUC among all evaluated deployment-compatible methods on KIRC, LUAD, and PAAD and outperforms both evaluated multimodal-to-unimodal baselines on all four cohorts. Component ablations highlight the complementary roles of its two blocks: X-GeoSCL transfers soft molecular neighborhood structure into the WSI mapping space, while retrieval and cross-attention read molecular evidence from paired training patients, and the learned gate integrates this evidence with direct WSI predictions. Together, these results support cross-modal neighborhood geometry as a useful form of privileged supervision and characterize preservation, recovery, and routing as complementary stages of multimodal-to-unimodal learning. More broadly, M2U provides a general framework for addressing deployment gaps in which a privileged modality is available during training but unavailable at inference.

## AI Use Statement

Generative AI tools were used to assist with manuscript drafting and language editing, literature search and retrieval, and refinement of presentation. All AI-assisted content was reviewed and verified by the authors, who take responsibility for the final manuscript, including its claims, citations, analyses, and reported results.

## Ethics Statement

This study uses de-identified, publicly available TCGA data and involves no new patient recruitment or intervention. Model routing weights are analyzed as internal aggregation behavior and are not interpreted as causal biological or clinical evidence.

## Reproducibility Statement

Architecture, preprocessing, training, endpoint construction, and evaluation procedures are described in the main text and appendix. The appendix reports secondary metrics and uncertainty, full mapping-and-reasoning ablations, patient-level routing heterogeneity, and retrieval-$K$ sensitivity. Code will be released upon publication.

## References


M. Ilse, J. M. Tomczak, and M. Welling. Attention-based deep multiple instance learning. In *ICML*, 2018.

M. Y. Lu, D. F. K. Williamson, T. Y. Chen, et al. Data-efficient and weakly supervised computational pathology on whole-slide images. *Nature Biomedical Engineering*, 5:555-570, 2021.

Z. Shao, H. Bian, Y. Chen, Y. Wang, J. Zhang, X. Ji, and Y. Zhang. TransMIL: Transformer based correlated multiple instance learning for whole slide image classification. In *NeurIPS*, 2021.

R. J. Chen, M. Y. Lu, W.-H. Weng, et al. Multimodal co-attention transformer for survival prediction in gigapixel whole slide images. In *ICCV*, 2021.

R. J. Chen, M. Y. Lu, D. F. K. Williamson, et al. Pan-cancer integrative histology-genomic analysis via multimodal deep learning. *Cancer Cell*, 40(8):865-878, 2022.

Y. Xu and H. Chen. Multimodal optimal transport-based co-attention transformer with global structure consistency for survival prediction. In *ICCV*, 2023.

F. Zhou and H. Chen. Cross-modal translation and alignment for survival analysis. In *ICCV*, 2023.

A. H. Song, R. J. Chen, G. Jaume, A. J. Vaidya, A. S. Baras, and F. Mahmood. Multimodal prototyping for cancer survival prediction. In *ICML*, 2024.

G. Jaume et al. Modeling dense multimodal interactions between biological pathways and histology for survival prediction. In *CVPR*, 2024.

G. Jaume, L. Oldenburg, A. Vaidya, R. J. Chen, D. F. K. Williamson, T. Peeters, A. H. Song, and F. Mahmood. Transcriptomics-guided slide representation learning in computational pathology. In *CVPR*, 2024.

T. Li, X. Zhou, J. Xue, L. Zeng, Q. Zhu, R. Wang, H. Yu, and J. Xia. Cross-modal alignment and contrastive learning for enhanced cancer survival prediction. *Computer Methods and Programs in Biomedicine*, 263:108633, 2025.

R. J. Chen et al. Towards a general-purpose foundation model for computational pathology. *Nature Medicine*, 2024.

K. Hemker, N. Simidjievski, and M. Jamnik. HEALNet: Multimodal fusion for heterogeneous biomedical data. In *NeurIPS*, 2024.

V. Swamy, M. Satayeva, J. Frej, et al. MultiModN: Multimodal, multi-task, interpretable modular networks. In *NeurIPS*, 2023.

H. Wang, Y. Chen, C. Ma, J. Avery, L. Hull, and G. Carneiro. Multi-modal learning with missing modality via shared-specific feature modelling. In *CVPR*, 2023.

A. Krishna, N. C. Kurian, A. Patil, A. Parulekar, P. Jeevan, and A. Sethi. PathoGen-X: A cross-modal genomic feature trans-align network for enhanced survival prediction from histopathology images. In *ISBI*, 2025.

F. Liu, L. Cai, Z. Wang, Z. Fan, J.-g. Yu, H. Chen, and Y. Zhang. Prototype-guided cross-modal knowledge enhancement for adaptive survival prediction. In *MICCAI*, 2025.

Y. Guo, H. Lu, O. C. Koyun, Z. Zhu, M. F. Demir, and M. N. Gurcan. Momentum memory for knowledge distillation in computational pathology. In *CVPR*, 2026.

M. Qu, G. Yang, D. Di, Y. Gao, T. Su, Y. Song, and L. Fan. Memory-augmented incomplete multimodal survival prediction via cross-slide and gene-attentive hypergraph learning. In *MICCAI*, 2025.

G. E. Hinton and S. T. Roweis. Stochastic neighbor embedding. In *NeurIPS*, 2002.

A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13:723-773, 2012.

A. van den Oord, Y. Li, and O. Vinyals. Representation learning with contrastive predictive coding. arXiv:1807.03748, 2018.

Y. Tian, D. Krishnan, and P. Isola. Contrastive representation distillation. In *ICLR*, 2020.

J. Denize, J. Rabarisoa, A. Orcesi, R. Hérault, and S. Canu. Similarity contrastive estimation for image and video soft contrastive self-supervised learning. *Machine Vision and Applications*, 34(6), 2023.

W. Park, D. Kim, Y. Lu, and M. Cho. Relational knowledge distillation. In *CVPR*, 2019.

F. Tung and G. Mori. Similarity-preserving knowledge distillation. In *ICCV*, 2019.

U. Khandelwal, O. Levy, D. Jurafsky, L. Zettlemoyer, and M. Lewis. Generalization through memorization: Nearest neighbor language models. In *ICLR*, 2020.

R. Hou, H. Chang, B. Ma, S. Shan, and X. Chen. Cross attention network for few-shot classification. In *NeurIPS*, 2019.

V. Vapnik and A. Vashist. A new learning paradigm: Learning using privileged information. *Neural Networks*, 22(5-6):544-557, 2009.

G. Klambauer, T. Unterthiner, A. Mayr, and S. Hochreiter. Self-normalizing neural networks. In *NeurIPS*, 2017.

E. L. Kaplan and P. Meier. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53(282):457-481, 1958.

The Cancer Genome Atlas Research Network, J. N. Weinstein, E. A. Collisson, et al. The Cancer Genome Atlas Pan-Cancer Analysis Project. *Nature Genetics*, 45:1113-1120, 2013.

E. Graf, C. Schmoor, W. Sauerbrei, and M. Schumacher. Assessment and comparison of prognostic classification schemes for survival data. *Statistics in Medicine*, 18(17-18):2529-2545, 1999.

# A Experimental details

## A.1 Endpoint and leakage controls

For each outer fold, the prediction horizon $\tau$ is the Kaplan-Meier median (Kaplan and Meier, 1958) estimated from training-fold event times only. With observed time $t_i$ and event indicator $\delta_i$,

$$y_i = \begin{cases} 1, & \delta_i = 1 \text{ and } t_i \le \tau, \\ 0, & t_i > \tau, \\ \text{excluded}, & \delta_i = 0 \text{ and } t_i \le \tau. \end{cases}$$

Patients censored after $\tau$ are retained as low-risk because survival beyond the horizon is observed; patients censored before $\tau$ have unknown status at the horizon and are excluded.

All outer splits are patient-level. The horizon, preprocessing statistics, molecular encoder, both WSI encoders, reader, and paired gallery are derived without held-out test patients. The gallery contains training-fold patients only, and the gate is fitted on validation predictions with the reader frozen. For IPCW-Brier evaluation (Graf et al., 1999), the censoring distribution is estimated from the corresponding training fold and applied unchanged to held-out patients.

## A.2 Implementation details

**Representations and unimodal encoders.** Slides are represented by precomputed frozen UNI patch features (256×256 patches at 20×, 1536 dimensions); foundation-model weights are not fine-tuned. Both WSI branches use the same AMIL architecture and output 256-dimensional patient embeddings, but have independent parameters: the unimodal branch is trained with BCE prognosis supervision, whereas the mapping branch is trained independently with its mapping objective. Molecular input contains RNA expression, copy-number alterations, and somatic mutations (approximately 15k features). A four-layer SNN with ELU activations and AlphaDropout 0.25 produces a 256-dimensional embedding; it is trained per fold with AdamW (learning rate $10^{-4}$), batch size 64, BCE supervision, and early stopping (patience 50), then frozen.

**Mapping.** The mapping encoder uses AdamW with learning rate $10^{-4}$, weight decay $10^{-5}$, cosine learning-rate annealing, and early stopping with patience 10. Batch size is 64 (32 for PAAD). The contrastive temperature is $\tau = 0.07$ for all mapping losses; X-GeoSCL sets $\sigma$ to the median pairwise molecular $\ell_2$ distance in each batch. All within-batch pairs are used, requiring $O(B^2)$ similarities per batch.

**Reasoning and gate.** The reader uses two cross-attention heads with head dimension 32 and 256-dimensional output. The interaction MLP has hidden dimension 256 and dropout 0.1 and is optimized with AdamW at $5 \times 10^{-5}$ with early stopping (patience 5). The sigmoid-parameterized gate is initialized at 0.8 and fitted on validation BCE with the reader frozen using Adam at learning rate $10^{-2}$. All experiments were run on one NVIDIA RTX A6000 (48 GB).

## B SECONDARY METRICS AND UNCERTAINTY

Table 3 complements the primary AUC analysis with censoring-aware IPCW-Brier, AUPRC, and balanced accuracy.

Table 3: **Secondary predictive metrics across four TCGA cohorts.** IPCW-Brier is lower-is-better; AUPRC and balanced accuracy are higher-is-better. Values are mean±SEM over five patient-level folds.

| Metric | Method | KIRC | LUAD | LGG | PAAD |
|---|---|---|---|---|---|
| IPCW-Brier ↓ | WSI retrieval | .390 ±.028 | .316 ±.071 | **.126** ±.026 | .242 ±.063 |
| | AMIL | .410 ±.022 | .484 ±.050 | .359 ±.044 | .222 ±.066 |
| | M2U (ours) | **.240** ±.006 | **.236** ±.009 | .135 ±.026 | **.191** ±.045 |
| AUPRC ↑ | WSI retrieval | .601 ±.021 | .618 ±.022 | .810 ±.046 | .553 ±.130 |
| | AMIL | .627 ±.014 | .625 ±.037 | .927 ±.024 | .697 ±.125 |
| | M2U (ours) | **.743** ±.013 | **.671** ±.042 | **.940** ±.017 | **.699** ±.130 |
| Bal. Acc. ↑ | WSI retrieval | .523 ±.025 | .447 ±.034 | .550 ±.050 | .590 ±.095 |
| | AMIL | .566 ±.031 | .472 ±.028 | **.688** ±.063 | .721 ±.089 |
| | M2U (ours) | **.597** ±.033 | **.503** ±.004 | .668 ±.061 | **.723** ±.085 |

**Bootstrap confidence intervals.** For the IPCW-Brier comparison in Figure 2, we bootstrap held-out patients while preserving paired predictions from M2U and the strongest WSI-only comparator for each cohort. We define $\Delta\text{Brier} = \text{Brier}_{\text{WSI}} - \text{Brier}_{\text{M2U}}$, so positive values favor M2U. The 95% confidence intervals are $[.061, .160]$ for KIRC, $[.020, .085]$ for LUAD, $[-.026, .019]$ for LGG, and $[.001, .066]$ for PAAD. Thus, the interval excludes zero for KIRC and LUAD; PAAD is borderline at the reported precision.

## C ADDITIONAL MAPPING AND REASONING ABLATIONS

Table 4: **Full mapping and reasoning ablation.** AUC is reported as mean±SEM. KNN uses label voting, CA reads retrieved molecular embeddings, and CA+Gate adds the direct WSI prediction.

| Cohort | Mapping objective | KNN | CA | CA+Gate |
|---|---|---|---|---|
| KIRC | InfoNCE | .566 ±.028 | .681 ±.017 | .661 ±.018 |
| | Soft-InfoNCE | .549 ±.024 | .680 ±.024 | .657 ±.027 |
| | X-GeoSCL | .548 ±.016 | **.700** ±.018 | **.683** ±.013 |
| LUAD | InfoNCE | .485 ±.033 | .543 ±.032 | .504 ±.036 |
| | Soft-InfoNCE | .471 ±.024 | .544 ±.045 | .526 ±.044 |
| | X-GeoSCL | .468 ±.030 | **.550** ±.052 | **.542** ±.052 |
| LGG | InfoNCE | .494 ±.012 | .518 ±.023 | .787 ±.032 |
| | Soft-InfoNCE | .512 ±.020 | .532 ±.047 | .785 ±.034 |
| | X-GeoSCL | .454 ±.038 | **.540** ±.054 | **.802** ±.022 |
| PAAD | InfoNCE | .477 ±.021 | .586 ±.073 | .745 ±.101 |
| | Soft-InfoNCE | .547 ±.031 | .526 ±.080 | .714 ±.113 |
| | X-GeoSCL | .501 ±.018 | **.595** ±.071 | **.747** ±.100 |

The learned retrieval weights are $\lambda = .826 \pm .050$ for KIRC, $.812 \pm .029$ for LUAD, $.487 \pm .075$ for LGG, and $.489 \pm .074$ for PAAD. Together with the corresponding branch AUCs in Figure 3, these values support the retrieval-dominant behavior of KIRC and LUAD and the WSI-dominant behavior of LGG and PAAD. The routing pattern is determined by branch strength and gate weight jointly, rather than by $\lambda$ alone.

## D PATIENT-LEVEL ROUTING HETEROGENEITY

Table 5: **Patient-level performance by risk subgroup.** Values are mean±SEM over five folds.

| Cohort | Model | AUC | Overall Acc. | Low-risk Acc. | High-risk Acc. | Low-risk F1 | High-risk F1 |
|---|---|---|---|---|---|---|---|
| KIRC | AMIL | .576 ±.021 | .579 ±.032 | .444 ±.070 | .689 ±.052 | .476 ±.048 | .641 ±.030 |
| | M2U | .683 ±.013 | .614 ±.041 | .446 ±.063 | .753 ±.062 | .503 ±.047 | .679 ±.043 |
| LUAD | AMIL | .429 ±.052 | .579 ±.071 | .113 ±.113 | .831 ±.169 | .077 ±.077 | .670 ±.115 |
| | M2U | .542 ±.052 | .636 ±.024 | .088 ±.088 | .938 ±.062 | .090 ±.090 | .763 ±.027 |
| LGG | AMIL | .819 ±.032 | .835 ±.037 | .447 ±.139 | .928 ±.038 | .494 ±.144 | .892 ±.028 |
| | M2U | .802 ±.022 | .829 ±.041 | .427 ±.131 | .928 ±.038 | .481 ±.141 | .888 ±.031 |
| PAAD | AMIL | .733 ±.093 | .705 ±.099 | .602 ±.169 | .840 ±.088 | .629 ±.165 | .712 ±.105 |
| | M2U | .747 ±.100 | .740 ±.074 | .821 ±.078 | .625 ±.178 | .784 ±.043 | .589 ±.175 |

The benefiting subgroup is not fixed across cancers. KIRC and LUAD primarily improve high-risk prediction, whereas PAAD primarily improves low-risk prediction; LGG remains approximately unchanged. This heterogeneity further motivates patient-adaptive rather than cohort-level routing.

## E RETRIEVAL-$K$ SENSITIVITY

Across $K \in \{3, 5, 7, 10, 15, 20\}$, $\Delta\mathrm{Purity@}K = \mathrm{Purity@}K_{\mathrm{omics}} - \mathrm{Purity@}K_{\mathrm{WSI}}$ remains positive in all four cohorts and changes modestly relative to fold-level variability, with no sharp dependence on $K$. We therefore use $K = 10$ as a representative neighborhood size rather than tuning $K$ per cohort. This analysis concerns neighborhood composition, not predictive-performance sensitivity.