ModalTune: Fine-Tuning Slide-Level Foundation Models with Multi-Modal Information for Multi-task Learning in Digital Pathology

Supplementary Material

In this Supplementary Material, we have provided additional details for the following:

- Hyperparameters (Sec. 7)
- Class Groupings (Sec. 8)
- Text Construction (Sec. 9)
- Text Embedding Analysis (Sec. 10)
- Ablations (Sec. 11)
- Experiments with TITAN (Sec. 12)
- Additional Modalities (Sec. 13)
- Qualitative Analysis (Sec. 14)

7. Hyperparameters

We display additional hyperparameters used to train ModalTune in Tab. 6.

8. Class Groupings

In the TCGA dataset, clinician-annotated cancer subtypes are highly detailed. However, predicting each individual subtype is infeasible due to their large number and the limited cases for certain subtypes. Therefore, we grouped the subtypes into broader categories and *Rare-set class* with the help of OncoTree code [37], as shown in Tab. 4. Subtypes in *Rare-set classes* were used for generating text embeddings but were excluded from baseline training due to their low sample size. Additionally, we merged subtypes with minor differences (e.g., "Squamous cell carcinoma, keratinizing" and "Squamous cell carcinoma, large cell, nonkeratinizing" were grouped as "Squamous cell carcinoma"). The final subtype bins for different cancer types are presented in Tab. 4.

For survival prediction task, we categorized survival durations into four bins, ensuring an approximately equal number of patients in each. The bin limits were then used to generate textual descriptions, as illustrated in Tab. 5.

9. Text Construction

We construct the embedded text prompts used to train ModalTune from clinical tables in `.csv` format available with all TCGA slides. This is done by firstly cleaning the table entries and converting them to natural language to take better advantage of semantic relationships in text. For example, we convert node status 0 (N0) to: "cancer has not spread to lymph nodes". For text related to tumor, node, metastastasis (TNM) staging, we also bin sub-categories of stages into a single stage to reduce variability of text embeddings (e.g. T1a, T1b, and T1c become "tumor stage 1"). We use the cancer subtype texts obtained after the preprocessing steps described in (Sec. 8).

We then describe the type of event (censored or an event occurred) along with a description of the bin as shown in Tab. 5. For example, a patient censored at 144 months would have the status: "The patient was censored between

Cancer Type	Class Groupings
BRCA	0: Infiltrating duct carcinoma 1: Lobular carcinoma Rare-set: infiltrating duct and lobular carcinoma, infiltrating (duct/lobular) mixed with other types of carcinoma, mucinous adenocarcinoma, metaplastic carcinoma, medullary carcinoma, intraductal papillary adenocarcinoma with invasion, tubular adenocarcinoma, adenoid cystic carcinoma, cribriform carcinoma
GBMLGG	0: Glioblastoma 1: Mixed glioma, Oligodendroglioma, Astrocytoma, Oligodendroglioma anaplastic, Astrocytoma anaplastic
NSCLC	0: Lung adenocarcinoma 1: Lung squamous cell carcinoma Rare-set: lung bronchiolo-alveolar carcinoma, lung papillary adenocarcinoma, lung acinar cell carcinoma, lung basaloid squamous cell carcinoma, lung solid carcinoma, lung signet ring cell carcinoma, lung papillary squamous cell carcinoma, lung micropapillary carcinoma
RCC	0: Papillary renal cell carcinoma 1: Renal clear cell carcinoma 2: Chromophobe renal cell carcinoma
COADREAD	0: Colon adenocarcinoma 1: Rectal adenocarcinoma Rare-set: colon mucinous adenocarcinoma, rectal mucinous adenocarcinoma, rectal adenocarcinoma in tubolovillous adenoma, rectal tubular adenocarcinoma, colon papillary adenocarcinoma
BLCA	0: Transitional cell carcinoma 1: Papillary transitional cell carcinoma

Table 4. Cancer subtype groupings for different cancer types

Cancer Type	Duration Bins
BRCA	0: before 15 months 1: between 15 and 27 months 2: between 27 and 55 months 3: between 55 and 283 months
GBMLGG	0: before 8 months 1: between 8 and 17 months 2: between 17 and 31 months 3: between 31 and 211 months
NSCLC	0: before 12 months 1: between 12 and 22 months 2: between 22 and 39 months 3: between 39 and 238 months
RCC	0: before 16 months 1: between 16 and 34 months 2: between 34 and 62 months 3: between 62 and 169 months
COADREAD	0: before 12 months 1: between 12 and 21 months 2: between 21 and 36 months 3: between 36 and 148 months
BLCA	0: before 11 months 1: between 11 and 18 months 2: between 18 and 30 months 3: between 30 and 163 months

Table 5. Duration bins for different cancer types. The text is used by ModalTune and the bin labels are used as labels by baselines

Parameter	Value
Slide encoder settings	
Embedding dim, D	768
Layers, L	12
Attention heads	16
Feedfoward dim	3072
Dilated attention segment lengths	$[1024, 2048, 4096, 8192, 16384]$
Dilated attention ratios	$[1, 2, 4, 8, 16]$
Activation function	GELU
Transcriptomics encoder settings	
Embedding dim, D_{gp}	256
# Compressed pathways, N_t	64
Layers	3
Feedforward expansion ratio	0.5
Dropout	0.25
Activation function	GELU
ModalTune settings	
Embedding dim, D	768
Adapter blocks, B	3
Adapter cross-attention heads	12
Adapter feedforward expansion ratio	0.25
Adapter dropout	0.1
Adapter initial gamma, γ_0^i	0
Final output dim, D_{final}	256
Text embedding dim, D_{text}	512
Training settings	
Epochs	30
Optimizer	AdamW
Max LR	1e-4
LR scheduler	LinearWarmupCosineAnnealing
Warmup epochs	10
Weight decay	0.0005
Batch size	1
Inference settings	
Logistic regression max iters	200
Logistic regression solver	liblinear
CPH penalizer	0.1

Table 6. Additional ModalTune hyperparameters.

55 and 283 months".

Task-specific text prompts contain information that is directly relevant to the task. For subtype classification ($j = 3$), we included the cancer site and the cancer subtype. For survival prediction ($j = 2$), we included the cancer site, TNM stages, and the survival status of the patient as described above. For the general task ($j = 1$), we merged the two prompts (only mentioning cancer site a single time). We chose to include TNM staging information for the survival and general tasks to better estimate and *delineate* risk between patients. We found staging to be prognostic, improving performance over solely relying on survival duration bins (Tab. 7). In any cases where TNM stage is not available (like the full patient cohort of GBMLGG), we simply omit stage-related text from the prompt. Example text generated for the low-risk patient in Fig. 3 is displayed in Fig. 4.

General prompt ($j = 1$)
"Cancer location: **breast**; Cancer diagnosis: **infiltrating duct carcinoma**; Overall stage: **stage one**; Tumor stage status: **tumor stage one**; Lymph node status: **cancer has not spread to lymph nodes**; Distant metastasis status: **no metastasis detected**; Survival status: **The patient was censored between 55 and 283 months**"

Survival prompt ($j = 2$)
"Cancer location: **breast**; Overall stage: **stage one**; Tumor stage status: **tumor stage one**; Lymph node status: **cancer has not spread to lymph nodes**; Distant metastasis status: **no metastasis detected**; Survival status: **The patient was censored between 55 and 283 months**"

Diagnosis prompt ($j = 3$)
"Cancer location: **breast**; Cancer diagnosis: **infiltrating duct carcinoma**"

Figure 4. Example text prompts generated for the low-risk patient in Fig. 3. Text in bold and red are directly obtained from clinical tables in TCGA after being clearned and converted to natural language.

10. Text Embedding Analysis

We analyze performance across different tasks using ($j = 1$) general task embeddings to confirm whether the text embeddings capture task-relevant information. Logistic regression is used for cancer subtyping and duration bin prediction, while cox proportional hazards model is applied for survival prediction. We report the mean and standard deviation of balanced accuracy for classification tasks and the C-index for survival prediction, averaged over three random seeds. We used the same splits as those used for other experiments. Overall, as shown in Tab. 7, we observe near-perfect performance across all tasks. Adding stage information to the general embedding slightly improves the C-index. Overall, as shown in Tab. 7, we observe near-perfect performance across all tasks. Adding stage information to the general embedding slightly improves the C-index. Random projections preserve the tight clusters, maintaining performance across multiple seeds and exhibiting similar performance in all of the tasks with minor degradations. This does not impact ModalTune; in fact, it enhances its performance in both cancer subtype prediction and survival prediction, as shown in Tab. 8.

11. Ablation Studies

In this section, we investigate multiple key design choices in our study: impact of different Modal Adapters, the effect of Modal Adapters, the effect of text embeddings, the impact of training solely on general prompts, the impact of different text encoders and Projectors, illustrated in Tab. 8.

11.1. Modal Adapters

Different Modal Adapters (LoRA): We compared ModalTune by extending LoRA [26], instead of using the ViT

Tasks	Text Embedding	Text Embedding after random projection
Cancer Subtyping	1.000 ± 0.000	1.000 ± 0.000
Duration Bins	1.000 ± 0.000	0.998 ± 0.002
Survival prediction w/o stage	0.966 ± 0.000	0.968 ± 0.000
Survival prediction	0.972 ± 0.000	0.970 ± 0.001

Table 7. Text embedding performance on different tasks on TCGA BRCA. We report balanced accuracy for cancer subtyping (2 class classification) and duration bins (4 class classification), and C-index for Survival prediction tasks

Ablation	BRCA	GBMLGG	NSCLC	RCC	Overall
Cancer Subtype Prediction					
Single Modal	0.887 ± 0.029	0.937 ± 0.012	0.926 ± 0.002	0.930 ± 0.009	0.920
No Text Embedding	0.885 ± 0.012	$\underline{0.998} \pm 0.002$	$\underline{0.956} \pm 0.002$	$\underline{0.954} \pm 0.005$	$\underline{0.948}$
Single Task Prompt	0.855 ± 0.005	0.995 ± 0.004	0.950 ± 0.005	0.939 ± 0.016	0.935
ABMIL (cat) w/ text emb.	0.874 ± 0.024	0.973 ± 0.021	0.934 ± 0.006	0.906 ± 0.012	0.922
ModalTune w/ LoRA	0.883 ± 0.026	$\underline{0.998} \pm 0.003$	$\mathbf{0.960} \pm 0.001$	0.920 ± 0.033	0.940
ModalTune w/ MedLlama v3.1	0.853 ± 0.011	0.993 ± 0.006	0.919 ± 0.009	0.918 ± 0.013	0.921
No Projector	0.891 ± 0.024	0.997 ± 0.002	0.948 ± 0.005	0.951 ± 0.011	0.947
Trainable Projector	0.612 ± 0.017	0.542 ± 0.023	0.768 ± 0.024	0.685 ± 0.078	0.652
Model-side Projector	$\underline{0.898} \pm 0.003$	0.993 ± 0.003	$\underline{0.956} \pm 0.006$	0.918 ± 0.044	0.941
ModalTune	$\mathbf{0.899} \pm 0.026$	$\mathbf{1.000} \pm 0.000$	$\underline{0.956} \pm 0.010$	$\mathbf{0.959} \pm 0.003$	**0.954**
Survival Prediction					
Single Modal	0.730 ± 0.025	0.821 ± 0.016	0.586 ± 0.013	0.689 ± 0.018	0.707
No Text Embedding	0.724 ± 0.024	0.881 ± 0.006	$\mathbf{0.631} \pm 0.010$	0.682 ± 0.022	0.730
Single Task Prompt	0.757 ± 0.014	0.872 ± 0.005	0.585 ± 0.018	$\underline{0.741} \pm 0.007$	0.739
ABMIL (cat) w/ text emb.	0.742 ± 0.016	0.869 ± 0.024	0.603 ± 0.033	0.710 ± 0.011	0.731
ModalTune w/ LoRA	0.756 ± 0.038	$\mathbf{0.894} \pm 0.008$	0.598 ± 0.011	0.728 ± 0.032	$\underline{0.744}$
ModalTune w/ MedLlama v3.1	0.752 ± 0.032	0.868 ± 0.011	0.603 ± 0.036	0.733 ± 0.012	0.739
No Projector	0.726 ± 0.007	0.868 ± 0.005	$\underline{0.612} \pm 0.028$	0.714 ± 0.016	0.730
Trainable Projector	0.693 ± 0.029	0.803 ± 0.016	0.610 ± 0.008	0.694 ± 0.027	0.700
Model-side Projector	$\underline{0.771} \pm 0.037$	$\underline{0.888} \pm 0.007$	0.594 ± 0.009	0.712 ± 0.019	0.742
ModalTune	$\mathbf{0.772} \pm 0.008$	0.879 ± 0.004	0.608 ± 0.023	$\mathbf{0.743} \pm 0.004$	**0.750**

Table 8. Ablations across different tasks and cancer types investigating key design choices of ModalTune. Best model in **bold**, second best is underlined

Adapter–based Modal Adapter, to handle transcriptomics and its interactions with the slide encoder. Overall, we observed that LoRA slightly underperformed ModalTune in both cancer subtype classification (1.5% drop) and survival prediction (0.8% drop), thereby motivating our choice of the Modal Adapter architecture.

Effect of Modal Adapters: To assess more deeply if the Modal Adapter architecture provides benefits in uni-modal fine-tuning, we evaluate the performance of ModalTune by replacing transcriptomics tokens from the genomic encoder with the same number and dimension of randomly initialized *trainable* embedding vectors ('Single Modal'). We find that overall, the model outperforms all other image-only models (Tab. 2) in subtype classification and has competitive performance in survival prediction. This effect is most pronounced when compared against the Gigapath fully fine-tuned model, where the model demonstrates superior performance across cancer subtype classifications and survival prediction (0.9%; 3.4%). Fine-tuning with the Modal Adapter setup requires updating fewer parameters than fully tuning Gigapath. Thus, this experiment demonstrates both the efficiency and effectiveness of the proposed architecture.

11.2. Multi-task using Texts

To examine the effect of multi-task learning, we train the Modal Adapter in a single-task manner, without embedding the tasks using text ('No Text Embedding'). We found cancer subtype prediction (0.6% drop) was less affected than survival prediction (2.7% drop), indicating the latter utilized *more* information from other tasks than the former. These findings indicate the utility of using text embeddings for multi-task learning and suggest that inter-task information is beneficial for downstream performance.

We also tested the utility of text embeddings for multi-task learning on other architectures such as ABMIL (cat) ('ABMIL (cat) w/ text emb.'), and found that compared to ABMIL (cat) trained on single tasks, there was an overall drop in performance for cancer subtype classification (1.6% drop), while performance for survival prediction remained similar. However, better architectures like ModalTune, in-

TITAN exp.	BRCA	GBMLGG	NSCLC	RCC	Overall
Cancer Subtype Prediction					
TITAN LP [17]	0.809	0.965	0.941	0.943	0.914
TITAN (Tuned) [17]	0.845 ± 0.007	0.948 ± 0.020	0.938 ± 0.002	**0.951** ± 0.003	0.920
TITAN (cat) [17]	0.849 ± 0.010	**0.998** ± 0.002	0.940 ± 0.018	0.941 ± 0.016	0.932
TITAN (KP) [17]	0.825 ± 0.011	**0.998** ± 0.003	**0.955** ± 0.003	0.949 ± 0.022	0.932
ModalTune TITAN	**0.872** ± 0.013	0.997 ± 0.002	0.950 ± 0.003	0.948 ± 0.012	**0.942**
Survival Prediction					
TITAN LP [17]	0.710	0.770	0.552	0.677	0.677
TITAN (Tuned) [17]	0.732 ± 0.009	0.832 ± 0.006	**0.620** ± 0.005	0.717 ± 0.002	0.725
TITAN (cat) [17]	0.745 ± 0.046	0.850 ± 0.010	0.604 ± 0.008	**0.729** ± 0.008	0.732
TITAN (KP) [17]	0.739 ± 0.018	**0.866** ± 0.018	0.571 ± 0.008	0.719 ± 0.008	0.724
ModalTune TITAN	**0.753** ± 0.012	0.858 ± 0.016	0.604 ± 0.012	0.725 ± 0.023	**0.735**

Table 9. Cancer subtype prediction balanced accuracy and survival prediction C-index scores across 4 cancer types for TITAN slide encoder. Best model in **bold**, second best is underlined. Here, LP refers to linear probing, cat refers to concatenation, and KP refers to Kronecker product.

Multimodal exp.	BRCA	GBMLGG	NSCLC	RCC	Overall
Cancer Subtype Prediction					
ModalTune	0.899 ± 0.026	**1.000** ± 0.000	0.956 ± 0.010	**0.959** ± 0.003	**0.954**
ModalTune w/ Clinical	**0.904** ± 0.020	0.998 ± 0.003	**0.959** ± 0.001	0.938 ± 0.010	0.950
Survival Prediction					
ModalTune	0.772 ± 0.008	0.879 ± 0.004	0.608 ± 0.023	0.743 ± 0.004	0.750
ModalTune w/ Clinical	**0.777** ± 0.012	**0.885** ± 0.013	**0.609** ± 0.016	**0.748** ± 0.019	**0.755**

Table 10. Experiments with incorporating clinical data alongside transcriptomics in ModalTune. Best model in **bold**.

	COADREAD	BLCA
Cancer Subtype Prediction		
TITAN LP	0.556	0.675
TITAN Sup. (cat)	0.585 ± 0.026	0.694 ± 0.013
TITAN Cls. (cat)	0.511 ± 0.018	0.526 ± 0.044
TITAN Surv. (cat)	0.522 ± 0.020	0.597 ± 0.018
ModalTune TITAN	**0.583** ± 0.089	**0.691** ± 0.016
Survival Prediction		
TITAN LP	0.562	0.615
TITAN Sup. (cat)	0.593 ± 0.036	0.679 ± 0.015
TITAN Cls. (cat)	0.483 ± 0.038	**0.617** ± 0.017
TITAN Surv. (cat)	0.549 ± 0.051	0.609 ± 0.032
ModalTune TITAN	**0.581** ± 0.062	0.611 ± 0.063

Table 11. Generalization study on OOD datasets using different TITAN-based models, compared with TITAN Sup. (cat) trained directly on the OOD data. Best OOD model in **bold**, second best is underlined.

corporating interaction terms, were able to achieve substantially improved performance when trained with text embeddings.

11.3. Task-Prompts

Here we investigate the role of using a multi-task prompt formulation versus simply pooling all tasks together into a general prompt, and performing single-task training. We do so by comparing our baseline model trained using both general and task-specific text embeddings ($T = 3$, 'ModalTune') versus a model trained solely on a general prompt ($T = 1$, 'Single Task Prompt'). Our results indicate that training with a single task prompt worsens overall model performance (2.0% drop in subtype prediction, 1.5% drop in survival prediction), potentially due to the regularization effects introduced by additional constraints that maximize the KL divergence between individual task-specific text vectors.

11.4. Text encoders

To evaluate the performance of ModalTune when using a different text embedding LLM, we tested Llama-3-8B-UltraMedical [75]. We observed a major drop in performance compared to ModalTune (3.4% in subtype prediction, 1.5% in survival prediction). We hypothesize several reasons for this decline. First, Llama-3-8B-UltraMedical is a general-purpose model trained on large-scale medical text datasets, whereas CONCH was trained in a contrastive manner using histopathology-related text datasets against image patches. This specialized training likely made CONCH a better fit for our use case, leading to superior performance. Similar findings were also reported in [30], where specialized models outperformed the generic model in the molecular status prediction task. Additionally, Llama-3-8B-UltraMedical is a generative model, requiring mean-

pooling after encoding to obtain a single 4096-dimensional text representation. In contrast, CONCH directly outputs a more compact text representation (512-dimensions), which may reduce the chance of overfit and hence improve ModalTune training.

11.5. Projectors

We found ModalTune to perform best when using a frozen and randomly-initialized Projector ('ModalTune'), which we explore here. Removing the Projector ('No Projector') simply requires adjusting the final output dimension D_{final} to 512, matching the dimensionality of text embeddings, D_{text}. This adjustment resulted in a drop in performance (0.7% in cancer subtype, 2.7% in survival prediction). We expect this occurred because the noise introduced by the random Projector has a regularizing effect on training, reducing model overfit on specific cancer sites. This has also been explored by Arani et. al. [3], where the introduction of noise in the knowledge distillation framework had positive effects. We additionally explore *training* the randomly-initialized Projector ('Trainable Projector'), which results in severe degradations in performance on both tasks. We believe this is due to model collapse, where the KL divergence loss function could be easily minimized by having the projector and the Modal Adapter output trivial solutions. The impact on survival prediction is less pronounced, which we attribute to the C-index metric being dependent only on *relative* ordering of risk scores. To avoid model collapse while tuning the Projector, we attach it to the end of the Modal Adapter instead of the text embeddings ('Model-side Projector'). We found best results when using a trained linear projection, though it still results in slightly inferior performance.

While unorthodox, these findings do align with prior studies highlighting the utility of randomly-initialized and fixed projectors in extracting non-trivial features in various scenarios. Of particular relevance to ModalTune, random projectors are effective feature extractors, reducing dimensionality and producing powerful representations [2, 55, 76]. Additionally, random projectors largely preserve inter-sample distances, as discussed in [22, 76], i.e., they maintain smaller distances between samples of the same class and larger distances between samples from different classes. This is evident empirically through the performance of linear regression on text embeddings (Tab. 7) with and without random projectors and theoretically from the Johnson-Lindenstrauss lemma [34], as discussed in Boutsidis et al. [6]. Given that clusters in the dataset are largely preserved and we opt to perform simple linear probing on extracted features, we expect the fixed random Projector to be a viable, generalizable, and effective projection method for ModalTune.

Overall, we find that using Modal Adapters, combining tasks with a text embedding, using multiple task prompts, and employing a fixed, randomly initialized Projector are all key components of ModalTune's success in improving the fine-tuning of SLFMs.

12. Experiments with TITAN

To demonstrate the ability of ModalTune to extend to other Transformer-based SLFMs, we perform experiments interfacing it with the TITAN [17] SLFM. To do so, we first re-extract patch features using the method described in the original work. We then perform analogous comparisons to Gigapath-ModalTune (Tab. 9). We found TITAN to be a much stronger standalone model than Gigapath, obtaining strong results with only linear probing or fine-tuning. Nonetheless, we find TITAN benefits from the bulk transcriptomics modality over uni-modal fine-tuning (1.3% in subtype classification, 1.0% in survival prediction), and slightly improves when tuned using the ModalTune pipeline (1.1%, 0.4%).

We additionally probe the generalizability of ModalTune TITAN similarly to the generalization study in Sec. 4.4 (Tab. 11). We found the ModalTune framework to greatly benefit OOD prediction performance on COADREAD and BLCA, with an average of 13.9% improvement in subtype prediction and 2.9% improvement in survival prediction than the next best OOD baseline. Furthermore, ModalTune demonstrated generalizability, performing only 0.4% worse than a fully-supervised TITAN (cat) network in subtype prediction and 6.3% worse in survival prediction. Thus, even though we obtained modest improvements from in-domain validation, we emphasize ModalTune maintains SLFM generalization better than conventional tuning methods.

13. Additional Modalities

To validate ModalTune's extensibility, we integrated clinical data (m_2) for TCGA by incorporating available features: patient age (all); TNM staging (BRCA, NSCLC, RCC); treatment type (BRCA only); and hormone receptor status (BRCA only). A 2-layer MLP encodes m_2 into $\mathbb{R}^{1\times D}$ for concatenation with transcriptomics [Sec. 3.2]. Overall, in our experiments shown in Tab. 10, the addition of clinical data marginally improved ModalTune's performance, primarily for survival prediction, while in the case of RCC subtype classification, it even led to a degradation, possibly because the clinical features for RCC are less relevant for the subtype classification task. This highlights ModalTune's ability to integrate salient features across modalities, though incorporating additional modalities remains a direction for future work.

14. Qualitative Analysis

14.1. t-SNE Analysis

After training ModalTune and ModalTune Pan-Cancer, we extract embedding vectors from combined train, validation, and testing datasets for every cancer site. For standard ModalTune, we extract embeddings using the best model per cancer site. For ModalTune Pan-Cancer, we simply use the overall best model. t-SNE plots of the extracted embeddings, along with text embeddings, are visualized in Fig. 5.

Notably, regardless of cancer sites being trained separately or together in a pan-cancer setup, embedding vectors distinctly cluster into individual sites. This may partially explain why we found minimal benefit in in-domain datasets from the pan-cancer experiments, as there is not much shared information between sites. We see much better separation in the former when comparing GBMLGG for standard ModalTune versus ModalTune Pan-Cancer. We expect this is due to issues with convergence mentioned in Sec. 4.5, where the best pan-cancer model had not yet converged on GBMLGG. In all other cases, embeddings are clearly clustered into groups based on primary diagnosis. In contrast, while text embeddings remain well separated for vital status and survival duration, separation is not as clear for embeddings from ModalTune. This is likely due to the inherently noisy nature of survival prediction. Since text prompts are directly created from clinical data and can only take discretized values, text embeddings are markedly more sparse than those generated from ModalTune.

14.2. Kaplan Meier Analysis

Although a high c-index risk model is preferred, it is equally important for the model to stratify patients into two distinct groups to aid clinicians in making treatment decisions, allowing them to choose between more or less aggressive interventions based on the patient's risk group. We used Kaplan-Meier curves on the test set to visualize this stratification, comparing high-risk and low-risk groups. The two groups were then assessed using a log-rank test to measure differences between their survival distributions, with a significance threshold set at $\alpha = 0.05$. In Fig. 6, we compare ModalTune against the best-performing survival models from image-only, genomics-only, and multi-modal categories, as well as Gigapath (cat). ModalTune consistently maintains significance in patient stratification across all four cancer types. ModalTune is the only model whose stratification was significant for NSCLC. Interestingly, the Kaplan-Meier curves for both ModalTune and Gigapath (cat) show strong similarities in patient stratification and their pattern, with ModalTune achieving improved stratification through better integration of transcriptomics information.

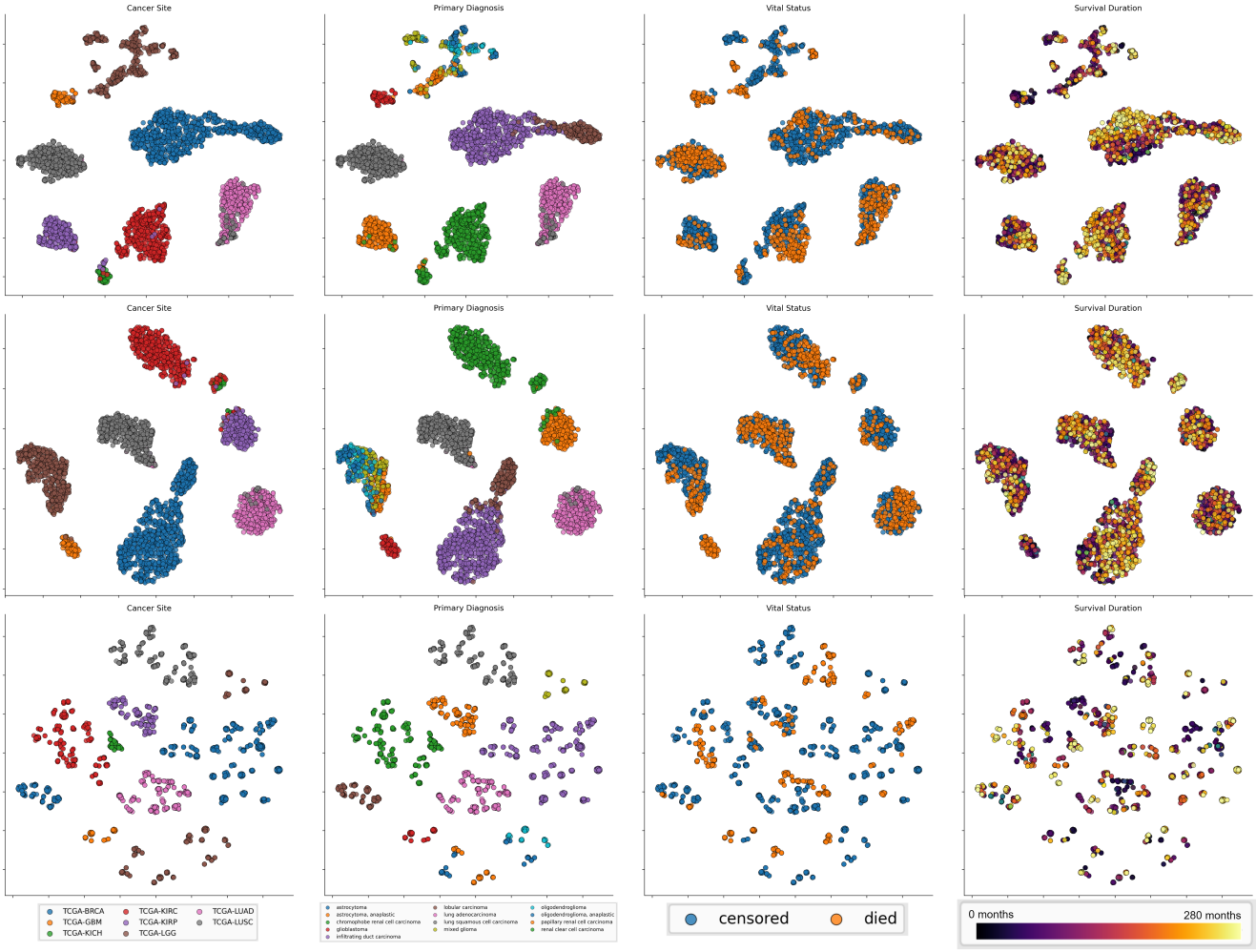

Figure 5. t-SNE plots generated for embeddings extracted using ModalTune (**first row**), ModalTune Pan-Cancer (**second row**), and text embedding vectors (**third row**). From left to right, each data point is colored by **cancer site**, **primary diagnosis**, **vital status**, and **survival duration**.

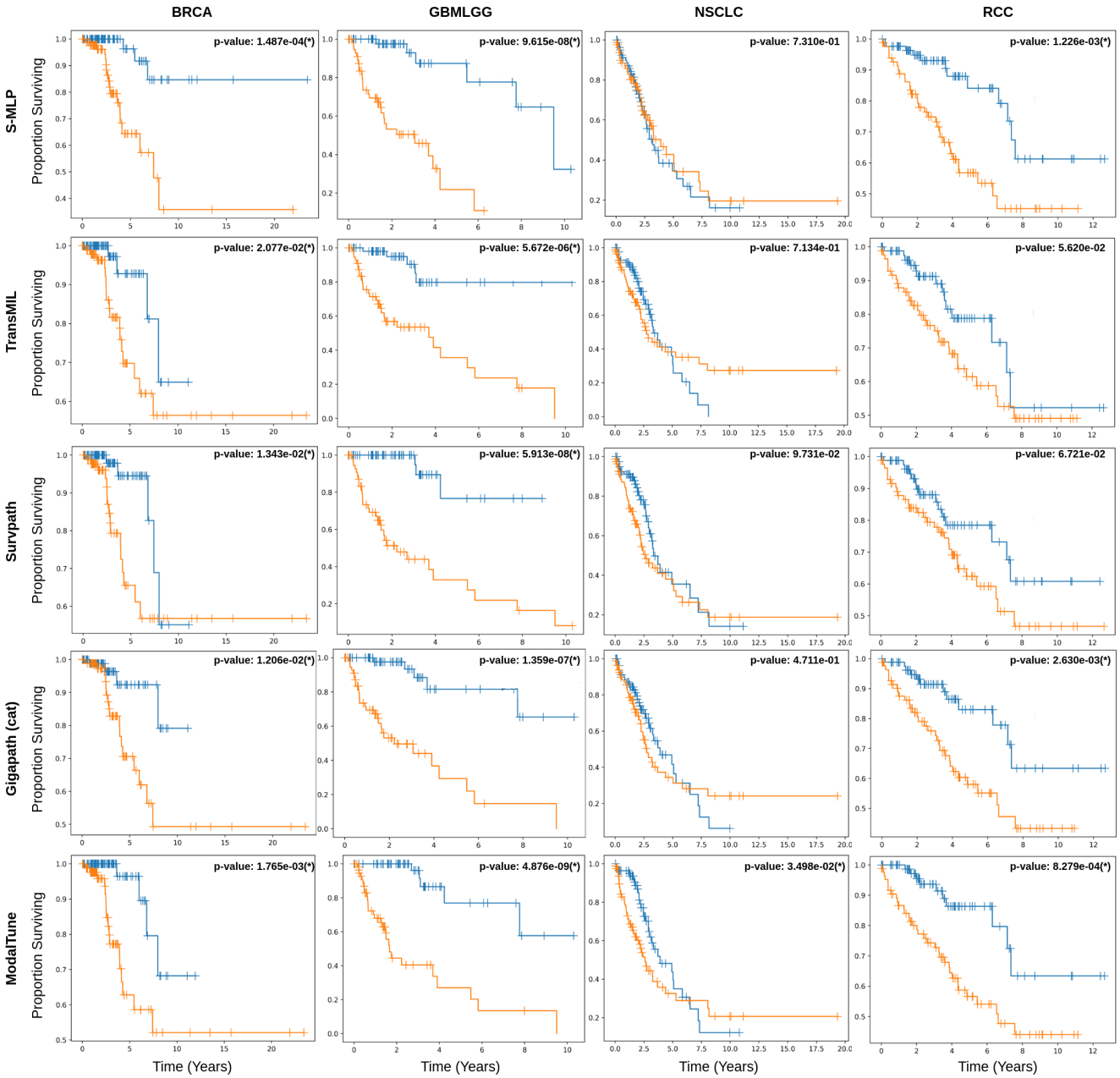

Figure 6. Kaplan-Meier curves of ModalTune and the best-performing survival model baselines across four cancer types. Patient groups are stratified based on the median of model-estimated risk scores on the test set, with orange representing the low-risk group and blue representing the high-risk group. A log-rank test with a significance threshold of $\alpha = 0.05$ was used to assess differences between the two distributions