A. Complementary Details

- **Evaluation Metrics:** Our primary evaluation metric is accuracy. For each dataset, we conduct three independent runs with different random seeds and report the mean accuracy along with the standard deviation to account for performance variability and ensure statistical robustness. Furthermore, to assess the statistical significance of the reported performance improvements across models and datasets, we use Fisher's Combined Probability Test. In addition, we evaluate the computational cost in terms of floating point operations (FLOPs) that provide a hardware-agnostic estimate of inference efficiency for a fair comparison of computational complexity.
- **Hardware:** All training was performed on NVIDIA RTX A6000 and NVIDIA A100 GPUs.
- **Training Details:** The fixed hyperparameters we used are shown in Table 1. For the hyperparameter search, we focus on LoRA's hyperparameters, i.e., LoRA rank ($r \in \{8, 16, 32\}$), LoRA scale ($\alpha \in \{2, 4, 8\}$), and LoRA targets (a combination of Linear, 2D Convolutional, and Embedding layers), in addition to λ.

Table 1. Training Hyperparameters

Parameter	**Value**
Optimizer	Adam
Learning Rate	1×10^{-3}
Weight Decay	1×10^{-4}
Batch Size	64
Epochs (PCAM, CRC)	1
Epochs (BACH)	25

B. Datasets

In this work, we evaluate our method on three publicly available histopathology image datasets: PCAM, BACH, and CRC. These datasets are chosen to represent a diverse set of tissue types, magnification levels, and classification difficulties. PCAM has the most number of patches, while BACH has a small number of patches , 400. For PCAM, we used the offical train-test-validation set and for BACH and CRC we use 10% of the train set as validation. A summary of the dataset characteristics is provided in Table 2. All datasets are used exclusively for image classification, with no textual metadata. To enable modality pairing, external histopathology reports are sourced from the TCGA dataset, which provides rich textual descriptions for related tissue types.

Algorithm 1 Modality Pairing

```
Require: Image dataset $\mathcal{D}_I = \{(I_i, y_i)\}_{i=1}^{N}$;
    external text dataset $\mathcal{D}_T = \{T_j\}_{j=1}^{M}$
Ensure: Paired dataset $\mathcal{D}' = \{(I_i, T_{\psi(i)}, y_i)\}_{i=1}^{N}$
Filter $\mathcal{D}_T$ with organ-specific keywords
for each $T_j \in \mathcal{D}_T$ do
    $\mathbf{t}_j \leftarrow \frac{f_t(T_j)}{\|f_t(T_j)\|}$ {Normalized text embeddings}
end for
for each $I_i \in \mathcal{D}_I$ do
    $\mathbf{v}_i \leftarrow \frac{f_v(I_i)}{\|f_v(I_i)\|}$ {Normalized image embedding}
    $j^* \leftarrow \arg\max_j \frac{\mathbf{v}_i \cdot \mathbf{t}_j}{\|\mathbf{v}_i\| \|\mathbf{t}_j\|}$
    Add $(I_i, T_{j^*}, y_i)$ to $\mathcal{D}'$
end for
return $\mathcal{D}'$
```

C. Models

We evaluate our method using a variety of vision transformer models with diverse capacities, ranging from lightweight backbones, such as ViT-S/16, to large-scale multimodal frameworks, such as CONCH. Table 3 summarizes the models and their parameter sizes. For DINO ViT-L/14, ViT-S/16, ViT-S/8, ViT-B/16, and ViT-B/8 we use the pathology fine-tuned version [4]. Additionally, we employ state-of-the-art histology specialized multimodal vision-language models: CONCH [7] and QUILTNet [5]. While these are trained multimodally, we also extract and evaluate their vision backbones separately to isolate the contribution of vision-only learning from paired training.

D. Algorithms

We outline the full training pipeline of `CLIP-IT` in Algorithms 1–3. First, each histology image is pseudo-paired with the most semantically relevant external report using CLIP-based similarity (Algorithm 1). These pseudo-pairs are then used to train a multimodal model that learns to predict textual features from image representations and combine them for classification (Algorithm 2). Finally, after training, the text encoder is discarded, and the resulting model retains only the vision branch and learned mappings, enabling efficient unimodal inference while benefiting from multimodal supervision (Algorithm 3).

E. Prompts

For the fine-tuning of multimodal models, specifically CONCH [7] and QUILTNet [5], we adopt a contrastive learning strategy. We follow the procedure outlined in their original works, where image and text pairs are jointly optimized using a contrastive loss function to bring semantically aligned vision-text pairs closer in the shared embedding space, while pushing apart unrelated pairs. This objec-

Table 2. Summary of the histopathology image datasets used in this study. The datasets vary in organ domain, number of patches, image resolution, magnification, and number of classes.

Dataset	Domain	Patch #	Patch Size	Magnification	Classes	Class Names
PCAM	Breast	$327,680$	96×96	$10\times$ ($0.97\,\mu$m/px)	2	0: Normal 1: Tumor
BACH	Breast	400	2048×1536	$20\times$ ($0.42\,\mu$m/px)	4	1: Normal 2: Benign 3: In situ carcinoma 4: Invasive carcinoma
CRC	Colorectal	$107,180$	224×224	$20\times$ ($0.50\,\mu$m/px)	9	0: Adipose tissue 1: Background 2: Debris (necrosis, mucus, hemorrhage) 3: Lymphocytes 4: Mucus 5: Smooth muscle 6: Normal colon mucosa 7: Cancer-associated stroma 8: Colorectal adenocarcinoma epithelium

Table 3. List of vision and multimodal models used in our experiments, along with their parameter sizes. We include both standard vision backbones (e.g., ViT-B/16, ViT-S/8) and large-scale multimodal models (e.g., CONCH, QUILTNet).

Model	# Parameters
CONCH [7]	395M
UNI [2]	303M
DINO ViT-L/14 [4]	303M
QUILTNet [5]	151M
CONCH Vision Backbone [7]	90M
QUILTNet Vision Backbone [5]	88M
ViT-B/16 [4]	86M
ViT-B/8 [4]	86M
ViT-S/16 [4]	22M
ViT-S/8 [4]	22M

tive encourages the models to learn rich cross-modal representations that align visual features with natural language descriptions. The text we used is the class name used with the prompts.

During inference, we evaluate the models using predefined textual prompts for each class. Given a set of candidate class labels, we replace ”CLASSNAME” with each label in templates to form descriptive textual prompts. These prompts are encoded using the text encoder and compared to the image embeddings to compute similarity scores. The final prediction is made using template ensemble averaging, where the logits or similarities across templates are aggregated for each class.

We use the official prompt templates provided by the authors of each model:

Algorithm 2 Multimodal Distillation

```
Require: Paired dataset $\mathcal{D}' = \{(I_i, T_i, y_i)\}_{i=1}^{N}$; multimodal model: $\mathcal{M}_{\theta_M} = \{f_t, f'_v, h_t, h_v, h_d, g\}$
Ensure: Trained multimodal model, $\mathcal{M}_{\theta_M}$
for each batch $(I^b, T^b, y^b)$ sampled from $\mathcal{D}'$ with $I^b = \{I_i^b\}_{i=1}^{N_B}$, $T^b = \{T_i^b\}_{i=1}^{N_B}$, $y^b = \{y_i^b\}_{i=1}^{N_B}$ do
    $\mathbf{v}^b \leftarrow f'_v(I^b)$ {Image features}
    $\mathbf{t}^b \leftarrow f_t(T^b)$ {Text features}
    $\hat{\mathbf{t}}^b \leftarrow h_d(\mathbf{v}^b)$ {Predicted text from image}
    $\hat{y}^b \leftarrow g(h_t(\hat{\mathbf{t}}^b), h_v(\mathbf{v}^b))$ {Predicted label}
    Update $\theta_M$:
        $\theta_M \leftarrow \theta_M - \eta \frac{1}{N_B} \sum_{i=1}^{N_B} \mathcal{L}(y_i^b, \hat{y}_i^b,\ t_i^b,\ \hat{t}_i^b)$
end for
return Trained $\mathcal{M}_{\theta_M}$
```

- **CONCH Prompts**: 22 natural language templates that reflect varied clinical and descriptive phrasings, such as:
  - *”an H&E image of CLASSNAME.”*
  - *”CLASSNAME.”*
  - *”a photomicrograph showing CLASSNAME.”*
  - *”a photomicrograph of CLASSNAME.”*
  - *”an image of CLASSNAME.”*
  - *”an image showing CLASSNAME.”*
  - *”an example of CLASSNAME.”*
  - *”CLASSNAME is shown.”*
  - *”this is CLASSNAME.”*
  - *”there is CLASSNAME.”*
  - *”a histopathological image showing CLASSNAME.”*
  - *”a histopathological image of CLASSNAME.”*
  - *”a histopathological photograph of CLASSNAME.”*

Algorithm 3 CLIP-IT

```
Require: Image dataset $\mathcal{D}_I = \{(I_i, y_i)\}_{i=1}^{N}$; text dataset $\mathcal{D}_T = \{T_j\}_{j=1}^{M}$; a vision encoder $f_v'$; text encoder text-encoder $f_t$
Ensure: Trained unimodal model:
        $\mathcal{M}_{\theta_U} = \{f_v', h_t, h_v, h_d, g\}$
$\mathcal{D}' \leftarrow$ Modality Pairing($\mathcal{D}_I, \mathcal{D}_T$)
Fine-tune text-encoder $f_t$ using text-label pairs $(T_i, y_i)$ from $\mathcal{D}'$
Initialize multimodal model:
        $\mathcal{M}_{\theta_M} = \{f_t, f_v', h_t, h_v, h_d, g\}$
for each epoch do
    $\mathcal{M}_{\theta_M} \leftarrow$ Multimodal Distillation( $\mathcal{D}'$, $\mathcal{M}_{\theta_M}$)
end for
Extract unimodal model: $\mathcal{M}_{\theta_U} \leftarrow \{f_v', h_t, h_v, h_d, g\}$
return $\mathcal{M}_{\theta_U}$ for unimodal inference
```

  - *"a histopathological photograph showing CLASSNAME."*
  - *"shows CLASSNAME."*
  - *"presence of CLASSNAME."*
  - *"CLASSNAME is present."*
  - *"an H&E stained image of CLASSNAME."*
  - *"an H&E stained image showing CLASSNAME."*
  - *"an H&E image showing CLASSNAME."*
  - *"CLASSNAME, H&E stain."*
  - *"CLASSNAME, H&E."*
- **QUILTNet Prompts**:
  - *"a histopathology slide showing CLASSNAME"*
  - *"histopathology image of CLASSNAME"*
  - *"pathology tissue showing CLASSNAME"*
  - *"presence of CLASSNAME tissue on image"*

The final prediction is made by selecting the class with the highest average similarity across its corresponding prompt variants.

F. Complementary Results

F.1. Reports or prompts?

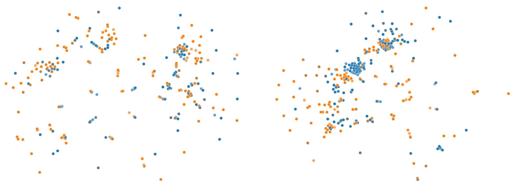
To answer this question, we compare the performance of a classification layer on top of our text encoder, with different texts. On one side we have the reports we pair using CLIP-IT, and on the other side, we have the prompts used for using our text-encoder (Conch text-encoder) in a CLIP-based approach. To do so, instead of reports, we pair the images of each dataset with prompts and descriptions introduced in [8]. As shown in Table 4, the reports are better to use than prompts, having significantly higher accuracy.

F.2. Feature Visualization

To further analyze this effect, we visualize the joint text-image embeddings using t-SNE for the UNI model on PCAM, with and without LoRA. As shown in Figure 2, the embeddings without LoRA remain more scattered and class-overlapping, indicating poor class separability. In contrast, the embeddings after LoRA fine-tuning form distinct and well-separated clusters based on the class labels. This demonstrates that since the text and image embeddings are not fully aligned, not using LoRA will add noise to the fusion process.

Table 4. Comparison of text classification accuracy on PCAM and BACH datasets using different types of textual inputs. We compare the prompts used in CONCH, and TQx prompt of [8] with the structured reports used in CLIP-IT.

Method	PCAM	BACH
Conch Prompt	62.99	36.26
TQx	64.91	43.06
CLIP-IT Text	**70.29**	**55.50**



(a) Text No LoRA (b) Text With LoRA

Figure 1. t-SNE plot of text embeddings with and without LoRA, for UNI-PCAM (blue: normal and yellow: tumor).

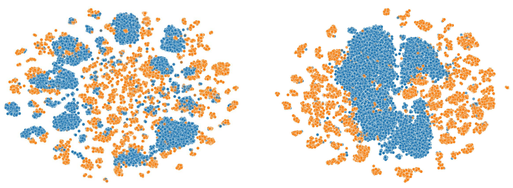
In order to further evaluate the influence of LoRA, we plot the t-SNE of text embeddings, with and without LoRA, as shown in Figure 1. The t-SNE shows that the classes form significantly more compact and separable clusters.

F.3. Text Statistics

To better understand the performance differences of CLIP-IT across datasets, we analyze the distribution of cosine similarity scores between the paired image and text embeddings for each dataset. Figure 4 shows the similarity histograms for PCAM, BACH, and CRC after the modality pairing step using the CLIP-based model. As shown in the figure, both PCAM and BACH have approximately Gaussian-like distributions centered around a mean cosine similarity of near 0.55. This suggests that for these datasets, the CLIP model successfully identifies semantically relevant textual descriptions for most patches, leading to high-quality pairings. In contrast, the distribution for CRC is broader and skewed toward lower values, with a mean around 0.40. This indicates that retrieved reports for CRC patches show lower average semantic alignment, explaining smaller downstream gains. This observation aligns with the experimental results: CLIP-IT yields the most significant performance improvements on PCAM and BACH, where the pairing quality is higher. For CRC, where the average similarity is lower, the benefit of using external reports as privileged information is reduced.

F.4. Distillation Ablation

To evaluate the effectiveness of our representation-level distillation, we compute the cosine similarity between the ground-truth text embeddings and the predicted (distilled) text features generated from the image branch. As shown in Figure 4, the distribution is sharply concentrated near 1, indicating high similarity between the original and distilled embeddings. This demonstrates that the text representations are accurately learned from the visual features, confirming the success of the distillation process in transferring semantic information from the text modality to the vision branch.



(a) Without LoRA

(b) With LoRA

Figure 2. t-SNE plot of text-image embeddings, (a) without and (b) with LoRA for PCAM (blue: normal and yellow: tumor). This demonstrates LoRA's importance in aligning the image and text feature spaces, as improved class separability in the fused embedding space reflects better multimodal representation learning.

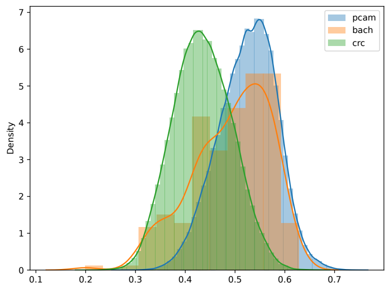


Figure 3. Histogram of cosine similarity scores between each histology image and its paired external report, as computed by the CLIP-based model during the modality pairing step. Distributions are shown for PCAM, BACH, and CRC datasets.

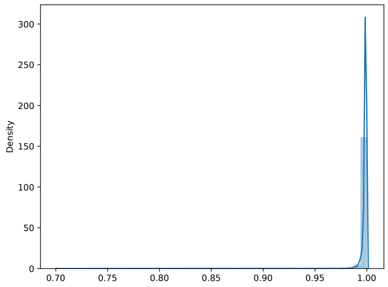


Figure 4. Histogram of cosine similarity between real text embedding and the distilled ones, showing that the text embedding was distilled perfectly to the vision.

F.5. Retrieval Backbone and Text Source

To further assess the flexibility of CLIP-IT, we ablate both the retrieval backbone and the source of textual data, as summarized in Table 5. Specifically, we compare three CLIP-style retrieval models: (i) a generic CLIP model trained on natural images, (ii) QUILTNet, and (iii) CONCH, which is domain-pretrained on histology image–text pairs. In addition, we test two different textual resources: TCGA pathology reports and synthetic text prompts generated from class labels. Results show that histology-pretrained retrieval backbones (e.g., CONCH, QUILT) consistently outperform generic CLIP, yielding gains of 1–2% across backbones. This highlights the importance of domain-specific retrieval for maximizing pseudo-pair quality. However, even generic CLIP achieves measurable improvements over unimodal baselines, confirming that CLIP-IT can still leverage external text supervision when only non-specialized retrieval models are available.

To further examine the dependency on specific text sources, we tested CLIP-IT with synthetic reports generated in a TCGA-like style. Specifically, we provided 10 real TCGA reports per dataset as in-context examples to a large language model (ChatGPT) and generated 500 synthetic reports for each class label. These synthetic reports were then used in place of the real TCGA reports during training. Using synthetic text in place of real reports yields performance comparable to TCGA reports in some cases (e.g., UNI, ViT-S/16), but generally underperforms real clinical text, underscoring the value of semantically rich, naturally occurring reports.

Overall, these results demonstrate that CLIP-IT is robust across retrieval backbones and text sources: domain-specific models and authentic reports provide the strongest benefits, but the framework remains effective even under weaker supervision.

F.6. Extension to Survival Prediction

While our main experiments focus on histology image classification, we emphasize that CLIP-IT is task-agnostic by

Table 5. Ablation for CLIP model and text source.

Text data	**TCGA**				**Synthetic**
Model	Unimodal	CLIP	Quilt	CONCH	CONCH
UNI	94.24	94.93	95.10	95.49	95.39
DINO	88.88	90.35	91.34	92.32	89.91
VITS16	88.13	90.14	91.09	91.42	90.85
VITB16	88.43	89.82	92.84	90.93	90.14

design. The pseudo-pairing and distillation pipeline can be integrated with a variety of downstream objectives beyond classification. In response to reviewer feedback, we conducted a preliminary experiment on survival prediction using the Metastatic Breast Cancer (MBC) whole-slide image dataset [1, 3]. We employed UNI as the patch-level feature extractor and AttentionMIL [6] with a Cox head for survival analysis.

Across five folds, the multimodal CLIP-IT approaches improved survival prediction compared to the unimodal baseline, raising the mean C-index from 0.609 to 0.661 $(+0.052)$ and the mean iAUC from 0.633 to 0.699 $(+0.066)$ (Table 6). Improvements were consistent across four folds, with only one fold showing a slight decrease. These results indicate that incorporating textual embeddings provides measurable gains in survival discrimination. Overall, these findings support our claim that CLIP-IT generalizes beyond image classification and can enhance performance in more complex tasks such as survival analysis.

Table 6. Survival prediction results on the MBC dataset.

Fold	**C-index**		**iAUC**	
	Unimodal	CLIP-IT	Unimodal	CLIP-IT
Mean	0.609	**0.661**	0.634	**0.699**
Δ	**+0.052**		**+0.066**	

References

[1] E. N. Bergstrom, A. Abbasi, M. Díaz-Gay, L. Galland, S. Ladoire, S. M. Lippman, and L. B. Alexandrov. Deep learning artificial intelligence predicts homologous recombination deficiency and platinum response from histologic slides. *J. Clin. Oncol.*, 2024.

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

[3] L. Galland, E. Ballot, H. Mananet, R. Boidot, J. Lecuelle, J. Albuisson, L. Arnould, I. Desmoulins, D. Mayeur, C. Kaderbhai, S. Ilie, A. Hennequin, A. Bergeron, V. Derangère, F. Ghiringhelli, C. Truntzer, and S. Ladoire. Efficacy of platinumbased chemotherapy in metastatic breast cancer and hrd biomarkers:utility of exome sequencing. *NPJ BC*, 2022.

[4] Ioannis Gatopoulos, Nicolas Känzig, Roman Moser, Sebastian Otálora, et al. eva: Evaluation framework for pathology foundation models. In *Medical Imaging with Deep Learning*, 2024.

[5] Wisdom Oluchi Ikezogwo, Mehmet Saygin Seyfioglu, Fatemeh Ghezloo, Dylan Stefan Chan Geva, Fatwir Sheikh Mohammed, Pavan Kumar Anand, Ranjay Krishna, and Linda Shapiro. Quilt-1m: One million image-text pairs for histopathology. *arXiv preprint arXiv:2306.11207*, 2023.

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

[7] 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*, 30:863–874, 2024.

[8] Anh Tien Nguyen, Trinh Thi Le Vuong, and Jin Tae Kwak. Towards a text-based quantitative and explainable histopathology image analysis . In *MICCAI 2024*. Springer, 2024.