ES4SURV: AN EVIDENCE-SEEKING MULTIMODAL LARGE LANGUAGE MODEL FOR WHOLE-SLIDE SURVIVAL PREDICTION

Anonymous authors
Paper under double-blind review

ABSTRACT

In digital pathology, accurate clinical predictions require integrating both global tissue context and local morphological details. While multimodal large language models (MLLMs) show strong potential, existing frameworks compress thousands of slide patches into a fixed set of global tokens, often discarding fine-grained, clinically decisive evidence. To overcome this limitation, we present ES4Surv, an autoregressive whole-slide image (WSI) MLLM that dynamically bridges macro- and micro-level features for survival analysis. ES4Surv first constructs a global slide representation, then uses that context to guide an adaptive, query-driven finding of high-resolution patch features aligned with predefined histologic concepts. These targeted patches are reinjected into the model, forming an interpretable reasoning pathway from global slide understanding to localized evidence and final prognosis. Across five TCGA benchmarks comprising 2,938 patients, ES4Surv consistently outperforms existing state-of-the-art slide-level models in term of macro disease-specific survival.

1 INTRODUCTION

Cancer survival prediction estimates a patient's probability of surviving beyond a given time, and histopathology whole slide images (WSI) provide rich morphological evidence for this task. However, a WSI contains an enormous amount of visual information, not all of which is equally relevant to survival prediction. Therefore, pathologists do not examine every region uniformly; they first establish the global tissue context, and then selectively inspect regions to identify prognostic patterns. This process is not simply information aggregation, it is a top-down process to find evidence, i.e., localized patterns involving malignant tissue, necrosis, stroma, immune infiltration, or their spatial interactions (Xue et al., 2026; Chakraborty et al., 2024). Crucially, WSIs also differ across patients in cellular morphology, tissue composition, and spatial organization. Therefore, this evidence seeking process should be tailored to each patient's slide, using its global tissue context to determine which components and local patterns to prioritize, rather than assuming that the same types of evidence are equally informative for every patient. Mirroring the workflow of pathologists, an effective survival model must move beyond passive slide-level aggregation; it must dynamically seek out local evidence, tailoring both what it inspects and where it searches to each patient.

WSI survival models have long sought to identify local morphological patterns associated with survival. For example, multiple instance learning (MIL) and related methods learn to prioritize informative patches or model their relationships during slide-level aggregation (Ilse et al., 2018; Shao et al., 2021; Lu et al., 2021; Yao et al., 2020; Chen et al., 2021a). Pathology foundation models such as UNI, CONCH, and GigaPath further strengthen this paradigm by providing transferable visual representations learned through large-scale pretraining (Chen et al., 2024; Lu et al., 2024; Xu et al., 2024). Building on such pretrained visual understanding, WSI multimodal large language models (MLLMs) introduce an additional opportunity: their autoregressive decoder forms a task-conditioned context after observing the slide and instruction, which could in principle be used to determine what visual information should be examined next. However, current WSI MLLMs mostly process a slide in two steps: they first compress thousands of patches into a small set of visual tokens, and then use these tokens for reasoning or prediction (Liang et al., 2025; Sun et al., 2025; Alawode et al., 2026). **Although efficient, these models must decide what visual information to**

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.

keep before they know what evidence will matter for the final prediction. Once decoding begins, these models can only access the compressed visual representation, with no mechanism to revisit the original high-resolution patches. Task-conditioned patch selection has been explored (Huang et al., 2026), but the selection is still made before the MLLM begins its subsequent reasoning.

To overcome this limitation, our key insight is to emulate the top-down diagnostic workflow of pathologists by enabling the MLLM to inspect the WSI across scales: first at coarse granularity to establish global tissue context, and then, conditioned on that context, at high resolution to find diagnostic local evidence for survival prediction. We introduce **ES4Surv**, an autoregressive WSI MLLM that actively seeks relevant tissue evidence to support survival prediction for the current patient (Fig. 1). Guided by its global slide understanding, the model selectively examines local tissue regions for potentially informative detail beyond the initial slide summary. Its pretrained knowledge therefore guides not only how available evidence is interpreted, but also where to look for additional evidence. The selected high-resolution patches are returned to the same MLLM, which continues decoding with both the global context and the newly acquired local evidence before predicting survival. In this way, evidence acquisition is conditioned on what the model has already understood about the slide **during** the reasoning process, rather than being completed entirely **before** reasoning begins.

The main technical challenge is how to use the MLLM's global understanding to search thousands of patches for local evidence and then incorporate the found evidence into survival prediction. We introduce **Query-Conditioned Adaptive Grouping and Patch Seeking (QAG-Seek)**, where decoder-generated queries participate in both *forming* candidate patch groups and *selecting* evidence within them. Multiple queries capture complementary tissue patterns, and each query is paired with a patch group. These groups are adaptively determined based on the global context of each slide. Meanwhile, they are aligned with predefined histologic concepts, while its actual patch membership remains slide-specific. The model then ranks the groups and selects the most relevant original patches from the retained groups. These patches are returned to the same MLLM as additional visual tokens, after which decoding resumes and a dedicated survival token produces the discrete time-to-event distribution. ES4Surv couples context-conditioned evidence seeking with autoregressive survival prediction in a single prediction process: the model first acquires local evidence, then predicts survival from both the global slide context and the selected fine-grained evidence.

Beyond improving survival prediction, ES4Surv makes the evidence-acquisition process directly interpretable. The WSI patches are clustered into concept-aligned groups. Scores on the groups indicate which tissue components are prioritized. Within each group, we know which are the exact patches used by the MLLM to make the final prediction. These records therefore trace how global slide context is translated into local evidence that is actually incorporated into the survival prediction, instead of post-hoc attribution of the final output. Across $2{,}938$ patients from five TCGA cohorts, ES4Surv achieves a macro disease-specific survival C-index of 0.738, outperforming the matched no-evidence-seeking baseline in all five cohorts and obtaining the highest macro C-index among the compared methods.

Our main contributions are:

- **Decoding-time evidence seeking for WSI survival prediction.** To our knowledge, ES4Surv is the first framework to enable MLLM-guided local patch selection during autoregressive decoding for WSI survival prediction. The model uses its global slide understanding to acquire finer local evidence and combines both levels of information for prediction.
- **Query-conditioned adaptive grouping and patch seeking.** Our proposed algorithm QAG-Seek lets MLLM-generated queries shape both candidate groups and patch selection, adapting to the current WSI and prediction context. Aligning each group with a predefined histologic concept gives the searches interpretable roles.
- **An inspectable evidence path.** ES4Surv makes visible which tissue groups the model prioritizes and which local patches it incorporates into its prediction context, connecting global slide understanding to identifiable evidence used during survival prediction.

2 Related Work

MIL and Multimodal Survival Architectures: The gigapixel resolution of WSIs makes direct end-to-end processing difficult. Multiple instance learning (MIL) addresses this through a local-to-global formulation: a patch encoder represents local morphology, and an aggregator combines the patch features into a slide-level representation (Ilse et al., 2018; Lu et al., 2021; Shao et al., 2021). Survival-specific models adapt this aggregation to prognosis through attention, spatial relations, recurring visual patterns, or language-guided priors (Yao et al., 2020; Chen et al., 2021a; Wang et al., 2023; Liu et al., 2025). These methods increasingly use pathology foundation models as pretrained patch encoders (Chen et al., 2024; Xu et al., 2024; Lu et al., 2024).

Multimodal Large Language Models for Whole-Slide Images: LLaVA-style models connect vision and language by converting image features into visual tokens consumed by the LLM (Liu et al., 2023). This conversion becomes a central design problem for WSIs, which yield thousands of patch features. SlideChat uses LongNet to contextualize the patch sequence and projects the resulting tokens to the LLM, whereas WSI-LLaVA reduces the LongNet output to a fixed-length visual-token sequence (Chen et al., 2025; Liang et al., 2025). CPath-Omni follows the same broad encode-then-tokenize paradigm, using multiscale region features and query-based token compression in its SlideParser (Sun et al., 2025). Other approaches change the visual evidence represented by these tokens: MLLM-HWSI supplies features across cell, patch, region, and slide scales (Alawode et al., 2026), while HistoSelect selects question-relevant patches before LLM decoding (Huang et al., 2026). ES4Surv retains both compact global WSI tokens and the original patch bank, then generates visual queries during decoding to identify additional local tokens before survival prediction.

Interpretability and Evidence Acquisition: To identify the prediction-driving information, post-hoc methods attribute model predictions to inputs or features (Selvaraju et al., 2017; Lundberg & Lee, 2017). Computational pathology methods also examine tile/patch weights, language concepts, prototypes, and molecular features (Ilse et al., 2018; Liu et al., 2025; Chen et al., 2021b; Zhang et al., 2026). A complementary approach is to explore how evidence enters the prediction process. CoVLM adopts this approach by returning features from selected image regions to the language model before generation continues (Li et al., 2024). ES4Surv **is the first MLLM to adopt** an evidence-seeking perspective for whole-slide survival analysis, directly tackling the domain-specific challenges posed by gigapixel WSIs.

3 Method

ES4Surv predicts survival in two stages: seeking local evidence and then using that evidence for prediction (Fig. 1). In Stage 1, the MLLM uses compressed visual tokens summarizing the WSI and a survival instruction to generate visual queries that guide the search for prognostic evidence. Each query helps gather high-resolution patches into a slide-specific candidate group and ranks patches within that group. The model then selects the most relevant groups and their highest-ranked patches from those groups as evidence. Organizing the search this way helps gather complementary tissue evidence that the global summary may not fully preserve.

In Stage 2, the selected patch features return to the same MLLM, which combines them with the global context to predict the patient's survival. In particular, we project the selected high-resolution patch features into visual tokens and provide them to the same MLLM alongside the global WSI context and text. A survival head maps the MLLM's post-evidence-seeking hidden state (*i.e.*, the "*sur*" token in Fig.1) to an event-time distribution.

While Stage 2 is straightforward, for the rest of this section, we focus on describing the first stage.

3.1 Evidence Seeking Stage

In this stage, we first summarize the whole slide to establish global context. Guided by this summary, the model then searches for fine-grained local evidence to predict survival. To achieve an effective search, we leverage contextualized queries aligned with different tissue contexts, group patches adaptively according to these queries (Sec. 3.1.1), and finally rank the groups and patches within these groups to select the most relevant patches for survival prediction (Sec. 3.1.2).

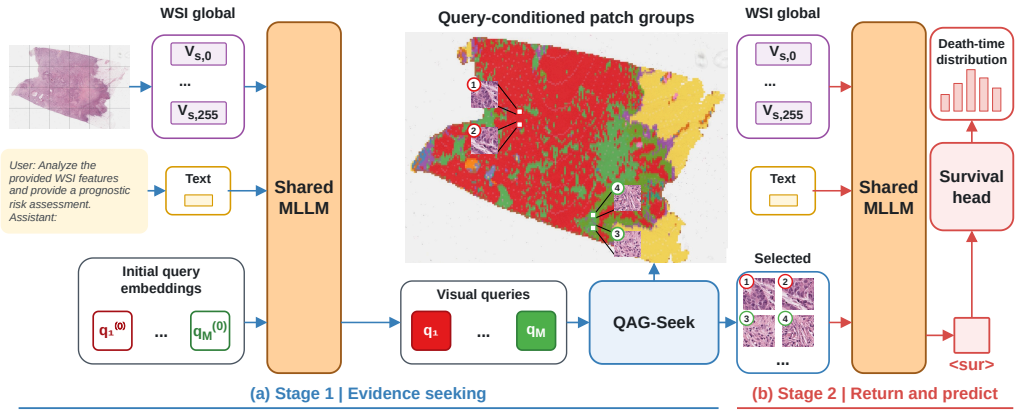


Figure 1: Overview of ES4Surv. In **(a)**, **Stage 1**, the MLLM generates visual queries from 256 global WSI tokens and text and query tokens; the visual queries cluster the WSI patch bank into groups and search for relevant local context. In **(b)**, **Stage 2**, the MLLM combines the text, global WSI tokens, and selected context to predict survival.

We begin by extracting patch-level representations from the whole-slide image using a pretrained pathology encoder. Following existing whole-slide vision-language paradigms (Liang et al., 2025), a slide encoder aggregates these patch representations into a compact sequence of global visual tokens that summarize the slide. The model processes these summary tokens alongside a survival instruction prompt to generate targeted search vectors, which we call **visual queries**. Rather than relying on static queries that fail to account for inter-patient heterogeneity, our queries are dynamically conditioned on the patient's global slide context and prompt. Specifically, a set of shared initial query embeddings ($Q^{(0)} = \{q_m^{(0)}\}_{m=1}^{M}$) is contextualized by the slide summary tokens and mapped into the patch feature space, forming M slide-specific queries ($Q = \{q_m\}_{m=1}^{M}$). This ensures each query adaptively reflects what evidence the model needs to seek for the specific patient under examination.

To ground these queries in local evidence, we structure the search into two sequential stages: *adaptive grouping* and *patch selection*. This design directly prevents queries from redundantly collapsing onto the same dominant visual hotspot. Instead of performing static or unconditional clustering, which would group patches purely by generic visual similarity, our grouping process is query-driven and adaptive. Guided by the visual queries, the group centers are iteratively updated and aligned with what the model seeks, dynamically partitioning the patch bank into concept-aligned histological sub-regions tailored to the slide's unique tissue architecture. Once these adaptive groups are established, we evaluate their prognostic relevance, rank the patches within each group, and return the most informative local features back to the model for multimodal survival prediction. The following subsections detail each component of this pipeline.

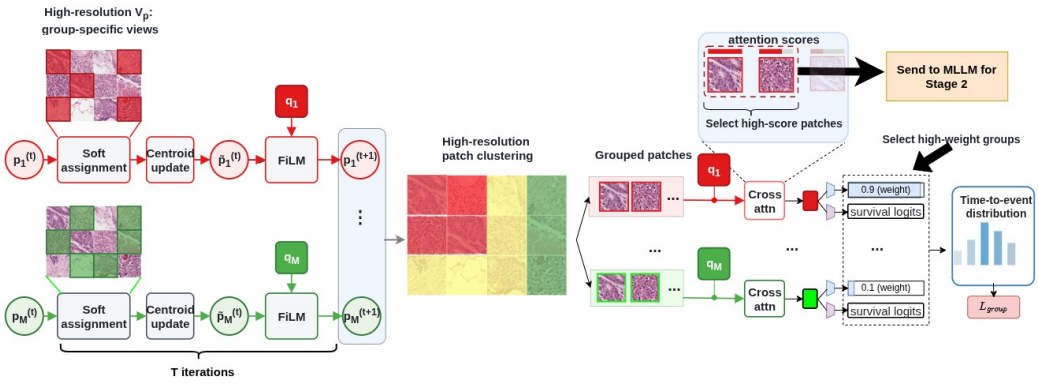
3.1.1 Query-Conditioned Adaptive Grouping

We start with pairing each query q_m with a learnable initial group center $p_m^{(0)}$. The query guides the refinement of this center, which in turn determines which patches enter the group. At each iteration, patches receive soft assignments based on their cosine similarity to the centers. Their features are pooled into group summaries $\widetilde{p}_m^{(\ell)}$, providing information about the tissue available in the current slide.

Each query then steers its center's update through feature-wise linear modulation (FiLM) (Perez et al., 2018):

$$p_m^{(\ell+1)} = \mathrm{Norm}\Big[(1 + \tanh \gamma_\ell(q_m)) \odot \mathrm{Project}_\ell\left(p_m^{(\ell)} + \widetilde{p}_m^{(\ell)}\right) + \tanh \beta_\ell(q_m)\Big], \tag{1}$$

where γ_ℓ and β_ℓ map the query to scale and shift vectors, and Norm denotes ℓ_2 normalization. The updated centers determine the next patch assignments. Repeating this process lets the queries progressively shape their candidate groups using the slide's patch features.



(a) Adaptive Grouping (b) Evidence Selection

Figure 2: Query-conditioned adaptive grouping and evidence selection. **(a)** Starting from learned prototypes, each center is iteratively updated using soft patch assignments and FiLM conditioning on its paired visual query, yielding query-conditioned groups. **(b)** Each query summarizes its group via cross-attention; the resulting group and patch attention scores are then used to select the final evidence patches, while group summaries are also trained by the auxiliary survival prediction.

After L iterations, we assign each patch to its most similar center, producing non-overlapping groups $\{\mathcal{P}_m\}_{m=1}^{M}$. Appendix A provides the assignment and pooling details.

Query guidance makes the groups responsive to the evidence sought in each slide, but does not by itself give each group a consistent tissue meaning across slides. We want the groups to provide recognizable, complementary sources of evidence, so that their selection can be interpreted in histologic terms. We therefore assign each group a predefined histologic concept and encourage its final center to match that concept during training:

$$\mathcal{L}_{\text{con}} = -\frac{1}{M}\sum_{m=1}^{M}\log\frac{\exp\left(\text{sim}(p_m^{(L)}, a_{\sigma(m)})/\tau_{\text{con}}\right)}{\sum_{c=1}^{C}\exp\left(\text{sim}(p_m^{(L)}, a_c)/\tau_{\text{con}}\right)}. \tag{2}$$

Here, a_c is the representation of concept c built from training-patch exemplars, $\sigma(m)$ assigns group m to its concept through a fixed one-to-one mapping, $C = M$, and τ_{con} is a temperature parameter. This alignment gives each group a stable histologic reference while the query and current slide determine which patches belong to it. Concept exemplars are used only during training. Appendix D gives the detail on how to find the examplar

3.1.2 Group Routing and Patch Selection

After adaptive grouping, each query q_m is paired with a slide-specific candidate group $\mathcal{P}_m$. The next step is to determine which groups should be examined further and which patches within them should be selected. We use the same query that helped form a group to evaluate the evidence contained within it. For each nonempty group $m \in \mathcal{M}^+ = \{m : |\mathcal{P}_m| > 0\}$, we compute

$$(x_m, \alpha_m) = \text{CrossAttn}\left(q_m, \{v_j : j \in \mathcal{P}_m\}\right), \tag{3}$$

where x_m is the resulting group representation and $\alpha_m = \{\alpha_{mj}\}_{j\in\mathcal{P}_m}$ are the normalized attention weights over its member patches. Thus, α_{mj} provides a within-group ranking of the local evidence associated with query q_m.

To prioritize groups, a learned routing function $A_{\text{route}} : \mathbb{R}^{d_v} \to \mathbb{R}$ maps each group representation to a scalar logit. We normalize these logits across the nonempty groups:

$$\pi_m = \frac{\exp(A_{\text{route}}(x_m))}{\sum_{n\in\mathcal{M}^+}\exp(A_{\text{route}}(x_n))}, \qquad m \in \mathcal{M}^+. \tag{4}$$

The resulting π_m gives the relative priority of group m, while α_{mj} ranks patches within that group.

We train these priorities to identify evidence useful for survival prediction through an auxiliary group-level survival objective. Each group representation produces event-time logits through ϕ_{group},

which are aggregated using the routing weights:

$$\hat{y}_{\text{group}} = \text{softmax}\left(\sum_{m \in \mathcal{M}^+} \pi_m \, \phi_{\text{group}}(x_m)\right), \qquad \mathcal{L}_{\text{group}} = \mathcal{L}_{\text{surv}}\left(\hat{y}_{\text{group}}\right). \tag{5}$$

Because x_m is produced through query-to-patch attention and π_m is computed from x_m, this objective jointly supervises the group representations, within-group patch attention, and group routing. It therefore encourages π_m to prioritize groups and α_{mj} to emphasize patches that support survival prediction.

Finally, we convert these soft priorities into a fixed finding budget. We retain up to the top p nonempty groups according to π_m and, within each retained group, the top k patches according to α_{mj}. This yields at most $p \times k$ original patch features from the high-resolution bank. Importantly, the process returns the individual features v_j rather than pooled group representations x_m, preserving the identity and WSI location of each selected patch. These features are subsequently returned to the MLLM as additional visual context for the final survival prediction. The auxiliary group-level prediction is used to train evidence selection; the final survival prediction is made only after the selected patches re-enter the MLLM.

3.2 Training Details

Putting everything together, we train the model with four loss terms:

$$\mathcal{L} = \lambda_{\text{LM}}\mathcal{L}_{\text{LM}} + \lambda_{\text{surv}}\mathcal{L}_{\text{surv}} + \lambda_{\text{group}}\mathcal{L}_{\text{group}} + \lambda_{\text{con}}\mathcal{L}_{\text{con}}, \tag{6}$$

where $\mathcal{L}_{\text{LM}}$ is the teacher-forced token loss, $\mathcal{L}_{\text{surv}}$ supervises the final event-time prediction, $\mathcal{L}_{\text{group}}$ trains the groups to identify prognostic evidence, and $\mathcal{L}_{\text{con}}$ aligns them with predefined histologic concepts. The λ terms are nonnegative loss weights. Appendix A gives the training details.

4 Experiments

Experiment Settings. We evaluate disease-specific survival (DSS) prediction on five TCGA cohorts: BLCA ($n = 380$), BRCA ($n = 1{,}045$), GBMLGG ($n = 569$), LUAD ($n = 457$), and UCEC ($n = 487$), totaling 2,938 patients, 3,650 diagnostic whole-slide images, and 597 DSS events. We use CLAM Lu et al. (2021) and Trident for tissue segmentation and non-overlapping patch extraction. GigaPath encodes patches for the MIL baselines and WSI-LLaVA variants, while VLSA uses its native CONCH representation. Performance is measured by the patient-level Harrell C-index. Target discretization, data splits, multi-slide aggregation, and metric computation are provided in Appendix B.

Compared Methods. We compare ES4Surv with MIL-based survival models ABMIL (Ilse et al., 2018), DeepAttnMISL (Yao et al., 2020), TransMIL (Shao et al., 2021), and PatchGCN (Chen et al., 2021a), as well as recent methods Surformer (Wang et al., 2023), SlotSPE (Zhang et al., 2026), and VLSA (Liu et al., 2025). For SlotSPE, we use the histopathology-only setting. We additionally adapt WSI-LLaVA (Liang et al., 2025) for survival prediction as a matched MLLM baseline. It predicts from the global WSI tokens alone, without selecting or re-entering high-resolution local patches.

Results. Table 1 shows that ES4Surv achieves the highest macro C-index, 0.738. Its matched global-context-only WSI-LLaVA baseline reaches 0.726, demonstrating that the pretrained MLLM representation already transfers competitively to survival prediction relative to SlotSPE (0.722), VLSA (0.718), and Surformer (0.717).

Adding decoder-guided local evidence improves over WSI-LLaVA in all five cohorts. ES4Surv attains the highest point estimates on BRCA, LUAD, and UCEC, while VLSA and SlotSPE remain strongest on BLCA and GBMLGG, respectively. The 0.012 macro improvement indicates that selecting and re-entering high-resolution patches complements the initial global WSI context. The associated group assignments and selected patches also make this additional evidence available for inspection.

5 Ablation Studies

Multiple Queries and Grouped Search. To test whether multiple queries require organized candidate groups, we compare three 32-patch retrieval schemes: one query over the full patch bank, eight

Table 1: Patient-level DSS C-index across five cohorts. Values are mean ± standard deviation over five folds; Overall is the fold-wise macro average. Bold denotes the highest point estimate per column.

Model	BLCA	BRCA	GBMLGG	LUAD	UCEC	Overall
ABMIL (Ilse et al., 2018)	0.589 ± 0.035	0.589 ± 0.148	0.749 ± 0.073	0.539 ± 0.073	0.678 ± 0.095	0.629 ± 0.068
TransMIL (Shao et al., 2021)	0.515 ± 0.064	0.619 ± 0.060	0.770 ± 0.047	0.597 ± 0.081	0.715 ± 0.094	0.643 ± 0.023
PatchGCN (Chen et al., 2021a)	0.565 ± 0.038	0.669 ± 0.091	0.799 ± 0.038	0.574 ± 0.072	0.739 ± 0.105	0.669 ± 0.016
DeepAttnMISL (Yao et al., 2020)	0.505 ± 0.024	0.585 ± 0.050	0.695 ± 0.034	0.520 ± 0.031	0.620 ± 0.111	0.585 ± 0.032
Surformer (Wang et al., 2023)	0.625 ± 0.040	0.705 ± 0.072	0.814 ± 0.025	0.642 ± 0.050	0.798 ± 0.069	0.717 ± 0.010
VLSA (Liu et al., 2025)	$\mathbf{0.632 \pm 0.031}$	0.718 ± 0.039	0.806 ± 0.038	0.615 ± 0.034	0.819 ± 0.048	0.718 ± 0.010
SlotSPE (Zhang et al., 2026) (histology only)	0.626 ± 0.060	0.711 ± 0.053	$\mathbf{0.823 \pm 0.028}$	0.649 ± 0.032	0.799 ± 0.066	0.722 ± 0.017
WSI-LLaVA (Liang et al., 2025)	0.623 ± 0.052	0.736 ± 0.022	0.807 ± 0.038	0.652 ± 0.060	0.810 ± 0.069	0.726 ± 0.019
ES4Surv	0.626 ± 0.054	$\mathbf{0.767 \pm 0.057}$	0.816 ± 0.020	$\mathbf{0.662 \pm 0.042}$	$\mathbf{0.821 \pm 0.081}$	$\mathbf{0.738 \pm 0.016}$

Table 2: Ablation of adaptive and query-conditioned grouping.

Variant	Slide-adaptive	Query-conditioned	Overall
Fixed CONCH13	No	No	0.722 ± 0.016
Adaptive w/o FiLM	Yes	No	0.724 ± 0.013
ES4Surv	Yes	Yes	$\mathbf{0.738 \pm 0.016}$

Table 3: DSS C-index versus patches per selected group k. All variants use $M = 13$ and $p = 4$; values are mean ± standard deviation over five folds.

k	BLCA	BRCA	GBMLGG	LUAD	UCEC	Overall
4	0.619 ± 0.032	0.738 ± 0.055	$\mathbf{0.817 \pm 0.023}$	0.631 ± 0.042	0.762 ± 0.085	0.713 ± 0.013
8	$\mathbf{0.626 \pm 0.054}$	$\mathbf{0.767 \pm 0.057}$	0.816 ± 0.020	$\mathbf{0.662 \pm 0.042}$	$\mathbf{0.821 \pm 0.081}$	$\mathbf{0.738 \pm 0.016}$
16	0.604 ± 0.035	0.762 ± 0.059	0.810 ± 0.017	0.624 ± 0.046	0.796 ± 0.058	0.719 ± 0.016
32	0.608 ± 0.028	0.705 ± 0.051	0.815 ± 0.029	$\mathbf{0.663 \pm 0.045}$	0.812 ± 0.068	0.721 ± 0.008

unrestricted queries, and the proposed grouped search. A single query concentrates 59.22% of its attention on the top patch, while the remaining selected patches receive only 1.32% on average. Eight unrestricted queries reduce top-1 attention to 39.29% but retrieve only 8.67 unique patches among 32 slots. Grouped search reduces top-1 attention to 19.83% and retrieves 30.80 unique patches, showing that grouping promotes broader, less redundant evidence retrieval.

Adaptive and Query-Conditioned Grouping. To test how the candidate groups in Sec. 3.1.1 should be formed, we compare fixed concept groups, slide-adaptive grouping without FiLM, and the complete query-conditioned adaptive model. All variants retain 13 concept-aligned groups, the same backbone and patient folds, and the same maximum 4×8 patches budget. This comparison isolates the effects of adapting group membership to the current WSI and allowing the decoder queries to condition that adaptation.

Table 2 shows a progression from 0.722 with fixed groups to 0.724 with slide-adaptive grouping and 0.738 after adding query-conditioned refinement. The small gain from slide-specific centers and the larger gain from query conditioning indicate that the complete benefit arises when group formation adapts to both the current WSI and the decoder's globally informed queries.

Number of Selected Patches. We keep $M = 13$ candidate groups and select the top $p = 4$ groups, then vary the number of selected patches per group $k \in \{4, 8, 16, 32\}$. These settings provide at most 16, 32, 64, and 128 local patch tokens, respectively. Table 3 shows that $k = 8$ gives the highest Overall C-index, 0.738 ± 0.016. Selecting more patches does not yield a monotonic improvement, suggesting that a larger local context can also introduce weakly relevant evidence. We therefore use $k = 8$ elsewhere.

Anchored Conditioning vs Direct Query Centers. We test whether the decoder queries can themselves serve as group centers. In this variant, query q_m replaces learned prototype anchor $p_m^{(0)}$, and the concept-alignment objective is applied directly to the post-MLLM queries. Because these queries are contextualized by the same decoder and jointly optimized for the same survival target, this direct separation is difficult: their mean pairwise cosine reaches 0.994 ± 0.001 and their effective rank is only 1.040 ± 0.006. Thus, nominally different queries converge to nearly the same search direction.

ES4Surv instead uses learned prototypes to establish distinct group identities before conditioning their slide-specific updates on the queries through FiLM. This anchored design reduces query pairwise cosine to 0.815 ± 0.044 and raises query effective rank to 2.079 ± 0.180; the final prototype centers also remain separated, with pairwise cosine 0.819 ± 0.029 and effective rank 1.717 ± 0.109. The difference carries into the resulting partition: the effective number of occupied groups increases from 3.475 ± 1.713 to 4.975 ± 1.424, while the dominant-group patch fraction falls from 0.632 ± 0.214 to 0.425 ± 0.128. These results support separating stable group identity from slide-specific search

intent: the prototypes serve as diverse anchors, while FiLM allows the MLLM queries to adapt those anchors without collapsing them into a common survival direction.

Group Scoring and Within-Group Patch Selection. Table 4 tests the two selection levels defined in Sec. 3.1.2: group scoring and within-group patch ranking through matched inference-time interventions on the same trained model, folds, and 4×8 patches budget. We randomize patches within the selected groups, randomize the selected groups while retaining patch ranking, or replace patient-specific group selection with the four group identities selected most frequently across the dataset.

The complete strategy achieves 0.738 ± 0.016. Random patch selection reduces the C-index to 0.714 ± 0.021, while random group selection reduces it to 0.716 ± 0.022, showing that both group scoring and within-group ranking contribute to evidence seeking. Using the same dataset-wide top-four groups for every patient reaches 0.729 ± 0.025: common tissue groups carry useful evidence, but patient-specific group selection remains stronger. Together, the interventions show that ES4Surv works best when it customizes both which tissue groups to inspect and which local patches to select from them.

Survival Readout without local evidence. Table 5 compares two ways to predict from the same initial visual context. Mean pooling applies the survival head to the mean of the image-conditioned decoder hidden states; the survival-token model uses the hidden state of the generated <sur> token. Neither model include local evidence. The survival-token readout achieves 0.726 ± 0.019, compared with 0.709 ± 0.027 for mean pooling, a macro gain of 0.017. It also retains an autoregressive prediction interface generating <sur> marks when the decoder produces a survival prediction instead of applying the head unconditionally to pooled hidden states.

Table 4: Selection interventions under a 4×8 finding patches budget. Macro DSS C-index over five folds.

Selection strategy	Overall
Top-4 groups by score, top-8 patches	$\mathbf{0.738 \pm 0.016}$
Top-4 groups by score, random-8 patches	0.714 ± 0.021
Random-4 groups, top-8 patches	0.716 ± 0.022
Dataset-wide top-4 groups, top-8 patches	0.729 ± 0.025

Table 5: Survival-readout ablation without local evidence. Macro DSS C-index over five folds.

Readout	Overall
Mean pooling	0.709 ± 0.027
Survival token	$\mathbf{0.726 \pm 0.019}$

6 Analyses

Inspectable Evidence Acquisition. ES4Surv exposes the evidence-acquisition path at both the group and patch levels. At the group level, concept alignment assigns each adaptive group a recognizable histologic role, while slide-specific assignments reveal which tissue regions instantiate that role for the current WSI. Routing scores then indicate which of these groups the model prioritizes for further inspection. In Fig. 3, for example, the four retained routes correspond to malignant tissue, benign tissue, stroma, and smooth muscle. Viewing their routing priorities together with the assignment map allows one to assess whether the model searches tissue regions consistent with the intended concepts. At the patch level, ES4Surv reveals the specific local evidence founded from each prioritized group. Each selected patch remains tied to its original WSI location, making it possible to inspect whether the founded morphology is relevant, redundant, or inconsistent with the group's intended semantic role. In Fig. 3, the eight patches shown for each retained group are precisely the patch features returned to the MLLM. These features enter the decoder as additional visual context before the final survival prediction, so the displayed patches are part of the model's prediction pathway rather than post-hoc highlights.

Together, the concept labels, group assignments, routing scores, and selected patches provide an inspectable trace from the semantic role of a search, to where it is instantiated on the slide, to the local evidence ultimately supplied to the decoder. This transparency concerns evidence acquisition rather than attribution: routing scores indicate group priority, not signed effects on risk, and the selection of a patch does not establish its causal contribution to the final prediction.

Group-Level Routing Analysis Within this 39-WSI sample, Table 6 shows nonuniform selection across the concept vocabulary. Malignant tissue is selected in 87.2% of WSIs and receives the largest mean routing weight (0.368). Plasma cells and smooth muscle are also frequently retained, with

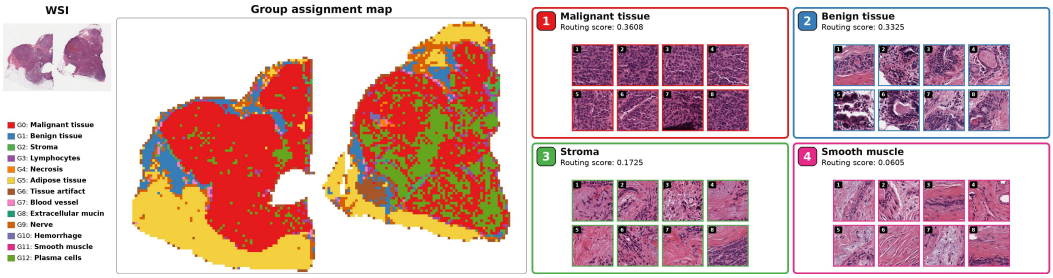


Figure 3: Inspectable evidence acquisition for one BRCA WSI. The assignment map shows the 13 concept-aligned groups formed for this slide. The model routes to four groups—malignant tissue, benign tissues, stroma, and smooth muscle—and find eight original patches from each, shown on the right. We also show each group's routing score.

Table 6: Group-level assignment and routing statistics on 39 fold-0 WSIs. "Patch share" is the percentage of patches assigned to each group, "Patches/WSI" the average count, "Selection rate" the percentage of WSIs in which the group is routed, and "Route" the mean routing weight π_m.

#	Concept	Patch share	Patches/ WSI	Sel. rate	Route	#	Concept	Patch share	Patches/ WSI	Sel. rate	Route
0	Malignant tissue	18.8	2244.1	87.2	.368	7	Blood vessel	16.9	2015.3	10.3	.039
1	Benign tissue	2.8	339.5	35.9	.061	8	Extracellular mucin	9.0	1074.9	51.3	.085
2	Stroma	0.9	102.5	35.9	.069	9	Nerve	4.2	501.8	5.1	.014
3	Lymphocytes	1.2	139.2	15.4	.032	10	Hemorrhage	2.0	243.1	5.1	.004
4	Necrosis	1.4	163.3	7.7	.011	11	Smooth muscle	19.2	2297.7	61.5	.174
5	Adipose tissue	7.4	880.9	7.7	.013	12	Plasma cells	10.1	1207.6	76.9	.131
6	Tissue artifact	6.2	743.9	0.0	.000						

selection rates of 76.9% and 61.5%, respectively, while extracellular mucin is selected in 51.3% of WSIs. Selection is not determined by group size alone: blood vessels account for 16.9% of assigned patches but are selected in only 10.3% of WSIs, whereas stroma accounts for 0.9% of patches but is selected in 35.9%. The artifact group is never selected despite containing 6.2% of assigned patches. These statistics characterize the evidence the model requests, rather than the effect of a tissue type on survival.

7 Limitations

ES4Surv makes evidence acquisition inspectable, but its decoder still combines the local features with the initial WSI context through high-dimensional transformations. Recording that a patch enters the model neither fully explains the final prediction nor establishes that patch's causal necessity or sufficiency. Concept alignment is defined by training exemplars rather than blinded pathologist annotation, and the predefined 13-concept vocabulary may not cover all prognostic morphology.

8 Conclusion

We propose ES4Surv, which reframes WSI survival prediction as an evidence-acquisition-then-prediction process. ES4Surv moves WSI survival prediction beyond a fixed, compressed view of the slide. Rather than requiring the model to make a prediction from global context alone, ES4Surv allows it to seek additional local evidence as prediction unfolds. The found evidence remains tied to identifiable tissue regions, making it possible to inspect not only the final prediction, but also what the model chose to examine before making it.

Across five TCGA cohorts, ES4Surv achieves a macro DSS C-index of 0.738, compared with 0.726 for the same global-context-only model. Our ablations further show that performance depends on how evidence is organized, prioritized, and selected, rather than simply on providing the model with additional patches. Together, these results point toward a broader direction for WSI modeling: instead of compressing a gigapixel slide once and predicting from what remains, models can actively return to the slide for the local evidence they need.

AI USE STATEMENT

Large language models were used as writing and review assistants to reorganize prose, improve clarity, identify ambiguous claims, and suggest technical consistency checks, code inspection and debugging. All AI-assisted technical suggestions and code changes were reviewed and validated by the authors. All reported measurements were produced by the authors' executable experimental pipelines and verified against saved outputs; no experimental values were generated or altered by generative AI. Separately, MedGemma 1.5 was used as the explicitly documented offline verifier during concept-exemplar bank construction, as described in Appendix D. The authors verified the data processing, equations, citations, implementation, results, and claims, and take full responsibility for the final submission.

ETHICS STATEMENT

This study uses publicly available, de-identified TCGA data and does not involve new patient data collection. ES4Surv is intended for research purposes only and is not a clinically validated decision-support system. Its predictions may reflect biases in the underlying datasets and may not generalize across populations or institutions. The selected tissue evidence should also not be interpreted as causal explanation or a substitute for expert pathological assessment.

REPRODUCIBILITY STATEMENT

We support reproducibility by documenting the complete experimental and implementation protocol. Section 4 specifies the cohorts, preprocessing, representations, comparison methods, and evaluation metric. Appendix A gives the model architecture, tensor dimensions, adaptive-grouping operations, concept alignment, group scoring, patch selection, and survival readout. Appendix B describes patient-level splitting and aggregation, discrete-time target construction, censoring-aware training, and C-index computation. Appendix C reports the optimization, checkpoint-selection, and baseline-training procedures, including random seeds and model-specific settings. Finally, Appendix D documents the fold-specific CONCH–MedGemma concept-bank construction and provides visual examples of the accepted exemplars. We will provide the anonymized implementation, experiment configurations, fold definitions, and evaluation scripts with the supplementary material.

REFERENCES

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

Souradeep Chakraborty, Rajarsi Gupta, Oksana Yaskiv, Constantin Friedman, Natallia Sheuka, Dana Perez, Paul Friedman, Gregory Zelinsky, Joel Saltz, and Dimitris Samaras. Decoding the visual attention of pathologists to reveal their level of expertise . In *proceedings of Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, volume LNCS 15003. Springer Nature Switzerland, October 2024.

Richard J. Chen, Ming Y. Lu, Muhammad Shaban, Chengkuan Chen, Tiffany Y. Chen, Drew F. K. Williamson, and Faisal Mahmood. Whole slide images are 2D point clouds: Context-aware survival prediction using patch-based graph convolutional networks. In *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, pp. 339–349. Springer International Publishing, 2021a. doi: 10.1007/978-3-030-87237-3_33.

Richard J. Chen, Ming Y. Lu, Wei-Hung Weng, Tiffany Y. Chen, Drew F. K. Williamson, Trevor Manz, Maha Shady, and Faisal Mahmood. Multimodal co-attention transformer for survival prediction in gigapixel whole slide images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4015–4025, 2021b.

Richard J Chen, Tong Ding, Ming Y Lu, Drew FK Williamson, Guillaume Jaume, Andrew H Song, Bowen Chen, Andrew Zhang, Daniel Shao, Muhammad Shaban, et al. Towards a general-purpose foundation model for computational pathology. *Nature medicine*, 30(3):850–862, 2024.

Ying Chen, Guoan Wang, Yuanfeng Ji, Yanjun Li, Jin Ye, Tianbin Li, Ming Hu, Rongshan Yu, Yu Qiao, and Junjun He. Slidechat: A large vision-language assistant for whole-slide pathology image understanding. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5134–5143. IEEE, 2025.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021. URL https://arxiv.org/abs/2106.09685.

Wentao Huang, Weimin Lyu, Peiliang Lou, Qingqiao Hu, Xiaoling Hu, Shahira Abousamra, Wenchao Han, Ruifeng Guo, Jiawei Zhou, Chao Chen, and Chen Wang. Act like a pathologist: Tissue-aware whole slide image reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6972–6981, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Huang_Act_Like_a_Pathologist_Tissue-Aware_Whole_Slide_Image_Reasoning_CVPR_2026_paper.html.

Maximilian Ilse, Jakub Tomczak, and Max Welling. Attention-based deep multiple instance learning. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2127–2136. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/ilse18a.html.

Junyan Li, Delin Chen, Yining Hong, Zhenfang Chen, Peihao Chen, Yikang Shen, and Chuang Gan. CoVLM: Composing visual entities and relationships in large language models via communicative decoding. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=PHGxChm1l5.

Yuci Liang, Xinheng Lyu, Wenting Chen, Meidan Ding, Jipeng Zhang, Xiangjian He, Song Wu, Xiaohan Xing, Sen Yang, Xiyue Wang, et al. Wsi-llava: A multimodal large language model for whole slide image. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22718–22727. IEEE, 2025.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

Pei Liu, Luping Ji, Jiaxiang Gou, Bo Fu, and Mao Ye. Interpretable vision-language survival analysis with ordinal inductive bias for computational pathology. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=trj2Jq8riA.

Ming Y Lu, Drew FK Williamson, Tiffany Y Chen, Richard J Chen, Matteo Barbieri, and Faisal Mahmood. Data-efficient and weakly supervised computational pathology on whole-slide images. *Nature biomedical engineering*, 5(6):555–570, 2021.

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

Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html.

Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron Courville. FiLM: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018. doi: 10.1609/aaai.v32i1.11671.

Andrew Sellergren, Chufan Gao, Fereshteh Mahvar, Timo Kohlberger, Fayaz Jamil, Madeleine Traverse, Alberto Tono, Bashir Sadjad, Lin Yang, Charles Lau, Liron Yatziv, Tiffany Chen, Bram Sterling, Kenneth Philbrick, Richa Tiwari, Yun Liu, Madhuram Jajoo, Chandrashekar Sankarapu, Swapnil Vispute, Harshad Purandare, Abhishek Bijay Mishra, Sam Schmidgall, Tao Tu, Anil Palepu, Chunjong Park, Tim Strother, Rahul Thapa, Yong Cheng, Preeti Singh, Kat Black, Yossi

Matias, Katherine Chou, Avinatan Hassidim, Kavi Goel, Joelle Barral, Tris Warkentin, Shravya Shetty, Dale Webster, Sunny Virmani, David F. Steiner, Can Kirmizibayrak, and Daniel Golden. Medgemma 1.5 technical report, 2026. URL https://arxiv.org/abs/2604.05081.

Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 618–626, 2017. doi: 10.1109/ICCV.2017.74.

Zhuchen Shao, Hao Bian, Yang Chen, Yifeng Wang, Jian Zhang, Xiangyang Ji, et al. Transmil: Transformer based correlated multiple instance learning for whole slide image classification. *Advances in Neural Information Processing Systems*, 34:2136–2147, 2021.

Yuxuan Sun, Yixuan Si, Chenglu Zhu, Xuan Gong, Kai Zhang, Pingyi Chen, Ye Zhang, Zhongyi Shui, Tao Lin, and Lin Yang. Cpath-omni: A unified multimodal foundation model for patch and whole slide image analysis in computational pathology. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10360–10371. IEEE, 2025.

Zhikang Wang, Qian Gao, Xiaoping Yi, Xinyu Zhang, Yiwen Zhang, Daokun Zhang, Pietro Liò, Chris Bain, Richard Bassed, Shanshan Li, et al. Surformer: An interpretable pattern-perceptive survival transformer for cancer survival prediction from histopathology whole slide images. *Computer Methods and Programs in Biomedicine*, pp. 107733, 2023.

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

Ruoyu Xue, Suryakant Singh, Souradeep Chakraborty, Pierre Marza, Oksana Yaskiv, Constantin Friedman, Natallia Sheuka, Paul Friedman, Bharat Ramlal, Beatrice Knudsen, Rajarsi Gupta, Joel Saltz, Prateek Prasanna, Gregory Zelinsky, and Dimitris Samaras. Pathologist Attention–Aligned Report Generation for Prostate Histopathology . In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2026*, volume LNCS 16891. Springer Nature Switzerland, September 2026.

Jiawen Yao, Xinliang Zhu, Jitendra Jonnagaddala, Nicholas Hawkins, and Junzhou Huang. Whole slide images based cancer survival prediction using attention guided deep multiple instance learning networks. *Medical Image Analysis*, 65:101789, 2020. doi: 10.1016/j.media.2020.101789.

Yilan Zhang, Li Nanbo, Changchun Yang, Jürgen Schmidhuber, and Xin Gao. Structural prognostic event modeling for multimodal cancer survival analysis. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=WqCRSn2WAY.

A Model Specification

A.1 Visual inputs and decoder

For each WSI, we load precomputed GigaPath slide tokens $G \in \mathbb{R}^{N_s \times 768}$, and a separate bank $X = \{x_j\}_{j=1}^{T}$ of raw GigaPath patch features $x_j \in \mathbb{R}^{1536}$. The slide tokens are zero-padded to width 1024 and mapped to the 4096-dimensional decoder space by a trainable linear–GELU–linear projector with widths $1024 \to 4096 \to 4096$. They enter the prompt at `<image>`. The patch bank is not inserted into this first decoder pass. If its stored length exceeds 15,000, adaptive average pooling along token order reduces it to 15,000 features. A pooled patch-bank entry need not correspond to a single original tile; we use *patch-bank feature* for either case.

The WSI-LLaVA-7B decoder has 32 layers, width 4096, 32 attention heads, and a maximum sequence length of 2048. We adapt its attention and MLP projections (`q_proj`, `k_proj`, `v_proj`, `o_proj`, `gate_proj`, `up_proj`, and `down_proj`) with LoRA (Hu et al., 2021) of rank 8, scale 16, and dropout 0.05; the remaining pretrained decoder weights are frozen. The global projector and the evidence-seeking components described below are trainable.

A.2 Initial query embeddings and visual queries

The survival instruction asks the model to analyze the WSI and provide a prognostic risk assessment. Its assistant trajectory contains `<ref>`, an evidence-context span, and `<sur>`. When `<ref>` is generated, it introduces $M = 13$ learned initial query embeddings $Q^{(0)} = \{q_m^{(0)}\}_{m=1}^{M}$. Each $q_m^{(0)}$ is the sum of a shared `<ref>` embedding and a learned, orthogonally initialized query-specific vector. In the first decoder pass, the global tokens and instruction precede these initial query embeddings. Their causal states $h_m^{(0)} \in \mathbb{R}^{4096}$ yield the visual queries

$$q_m = \mathcal{A}_q(h_m^{(0)}) \in \mathbb{R}^{4096}, \qquad m = 1, \dots, M. \tag{7}$$

The trainable query projector $\mathcal{A}_q$ uses layer normalization, linear layers of widths $4096 \to 16384 \to 8192 \to 4096$, GELU between the linear layers, and dropout 0.1. These slide-conditioned visual queries are distinct from both the initial query embeddings and the learned group centers. Training teacher-forces this sequence; inference generates `<ref>` and `<sur>` autoregressively.

A.3 Query-conditioned adaptive grouping

A trainable patch encoder $\mathcal{E}_x : \mathbb{R}^{1536} \to \mathbb{R}^{4096}$ maps each raw patch-bank feature to $k_j = \mathcal{E}_x(x_j)$. It uses layer normalization and two 4096-wide linear layers with GELU and dropout 0.1. We initialize $M = 13$ trainable group centers $p_m^{(0)} \in \mathbb{R}^{4096}$ and normalize them before assignment. Each initial group center is paired with one visual query q_m.

The same bias-free projection $W_g : \mathbb{R}^{4096} \to \mathbb{R}^{1536}$ is used for patch assignment and concept guidance. At refinement step ℓ, normalized raw features and projected centers give

$$\begin{aligned} z_{jm}^{(\ell)} &= \mathrm{sim}\left(x_j, W_g p_m^{(\ell)}\right), \qquad s_{jm}^{(\ell)} = \mathrm{softmax}_m\left(z_{jm}^{(\ell)}/0.3\right), \\ u_{jm}^{(\ell)} &= [s_{jm}^{(\ell)} - 1/M]_+, \qquad \mu_m^{(\ell)} = \frac{\sum_{j=1}^{T} u_{jm}^{(\ell)} k_j}{\max(\sum_{j=1}^{T} u_{jm}^{(\ell)}, \epsilon)}. \end{aligned} \tag{8}$$

Here sim is cosine similarity and $\epsilon > 0$ is a numerical floor. The first assignment uses the learned initial group centers before query conditioning. Each of the $L = 4$ blocks recomputes these soft assignments, forms the weighted summary, and applies a residual FFN update followed by FiLM:

$$\begin{aligned} v_m^{(\ell)} &= \mathrm{LN}_\ell\left(p_m^{(\ell)} + \mu_m^{(\ell)}\right), \qquad \widetilde{p}_m^{(\ell)} = \mathrm{LN}_\ell\left(v_m^{(\ell)} + \mathrm{FFN}_\ell(v_m^{(\ell)})\right), \\ p_m^{(\ell+1)} &= \mathrm{Norm}\Big[\widetilde{p}_m^{(\ell)} \odot (1 + 0.1 \tanh \gamma(q_m)) + 0.1 \tanh \beta(q_m)\Big]. \end{aligned} \tag{9}$$

Each block has its own layer normalizations and width-8192 FFN; the FiLM map producing γ and β is shared across blocks. Its last linear layer is initialized to zero. There is no self-attention between group centers. Group competition occurs through the assignment softmax. After the fourth block, we recompute $z_{jm}^{(L)}$ and assign each patch-bank feature to $\arg\max_m z_{jm}^{(L)}$. Empty groups are masked from group routing and evidence selection.

A.4 CONCEPT ALIGNMENT, GROUP ROUTING, AND EVIDENCE SELECTION

We align the 13 groups, in fixed order, with malignant tissue, benign tissue, stroma, lymphocytes, necrosis, adipose tissue, tissue artifact, blood vessel, extracellular mucin, nerve, hemorrhage, smooth muscle, and plasma cells. For every fold, a concept-exemplar bank is constructed using training WSIs only. The 50 raw GigaPath exemplars for concept c are averaged to obtain the frozen target $a_c \in \mathbb{R}^{1536}$. Appendix D describes the CONCH–MedGemma construction procedure and visualizes the accepted exemplars. With $\sigma(m)$ the fixed concept paired with group m, the concept loss is

$$\mathcal{L}_{\text{con}} = -\frac{1}{M}\sum_{m=1}^{M} \log \frac{\exp\{\text{sim}(W_g p_m^{(L)}, a_{\sigma(m)})/0.07\}}{\sum_{c=1}^{M} \exp\{\text{sim}(W_g p_m^{(L)}, a_c)/0.07\}}. \tag{10}$$

The frozen targets supervise final centers through W_g; neither the targets nor their screening models are used at inference.

For each nonempty hard group $\mathcal{P}_m$, query q_m attends to encoded keys k_j for $j \in \mathcal{P}_m$. Its normalized attention α_{mj} ranks member patches, and the *raw* group summary is $r_m = \sum_{j\in\mathcal{P}_m} \alpha_{mj} x_j \in \mathbb{R}^{1536}$. A trainable projector maps r_m to the group representation $x_m \in \mathbb{R}^{4096}$. One shared routing function and one shared 11-way survival head give

$$\begin{aligned} s_m &= f_{\text{route}}(x_m), \qquad \pi_m = \text{softmax}_{m\in\mathcal{M}_+}(s_m), \\ \hat{y}_{\text{group}} &= \text{softmax}\left(\sum_{m\in\mathcal{M}_+} \pi_m \phi_{\text{group}}(x_m)\right), \qquad \mathcal{M}_+ = \{m : |\mathcal{P}_m| > 0\}. \end{aligned} \tag{11}$$

Here, s_m is the routing score and π_m is its normalized routing weight. The routing function consists of layer normalization followed by a scalar linear layer. The group projector has widths $1536 \to 4096 \to 4096$; the survival head has widths $4096 \to 2048 \to 11$ with ReLU and dropout 0.1. The auxiliary prediction uses all nonempty groups and trains group routing; it is not combined with the final decoder prediction.

The top $p = 4$ nonempty groups by s_m (equivalently π_m) each contribute up to $k = 8$ patches ranked by α_{mj}. This yields at most 32 selected patch-bank features, with no replacement for empty or underfilled groups. A linear–GELU–linear projector maps each selected raw feature from width 1536 to 4096. These evidence tokens replace the static 13-token context span in a complete second pass through the same decoder. The global WSI tokens remain visible; downstream attention to the initial query embeddings is masked after evidence insertion. The main 11-way survival head reads the causal state immediately *before* `<sur>`, which is the state used to predict that token. Thus the final reported distribution uses the returned evidence, whereas $\hat{y}_{\text{group}}$ is an auxiliary training output.

B EVALUATION PROTOCOL DETAILS

Data splits and patient-level aggregation. All methods use the same five-fold patient-level splits and DSS labels. The event indicator is $e = 1 - c_{\text{DSS}}$, where c_{DSS} denotes censoring. Training examples are slide-level: each diagnostic slide carries its patient's outcome, and all slides from the same patient remain in the same fold. WSI-LLaVA processes slides individually; for patients with multiple slides, we average their predicted incidence distributions before computing one patient-level risk score. We report the mean and standard deviation over folds. The overall score in each fold is the unweighted mean of the five cohort-specific C-indices.

Discrete-time targets. We predict an 11-bin incidence distribution. For each cohort and fold, bin boundaries are constructed from uncensored DSS times of training patients only, then applied unchanged to validation and test patients; the final interval is open-ended. Cohort-wise binning maps different absolute time scales to a common 11-position ordinal target space. ES4Surv first shares parameters and an 11-way survival head across cohorts in pooled training, then independently specializes a child model for each cohort and fold. Cohort identity is not a textual or learned model input; it determines interval labels and which cohort-specific checkpoint is used. The output is therefore interpreted on the corresponding cohort's time scale rather than as a survival curve calibrated across cancers. Training uses interval labels, whereas evaluation uses the original continuous DSS times and event indicators.

Censoring-aware survival objective. Let $\bar{t}$ denote the interval containing the event or last-follow-up time t, and let $e \in \{0, 1\}$ indicate whether the event was observed. For the predicted incidence distribution $\hat{\mathbf{y}} = [\hat{y}_1, \ldots, \hat{y}_K]$, we use

$$\mathcal{L}_{\text{surv}} = -e \log \hat{y}_{\bar{t}} - (1 - e) \log \left(\sum_{k=\bar{t}}^{K} \hat{y}_k \right). \tag{12}$$

C-index computation. From the predicted incidence distribution $\hat{\mathbf{y}} = [\hat{y}_1, \ldots, \hat{y}_K]$, we derive

$$S_k = 1 - \sum_{j=1}^{k} \hat{y}_j, \qquad r = -\sum_{k=1}^{K} S_k.$$

Let $\psi(r_i, r_j) = \mathbf{1}[r_i > r_j] + \frac{1}{2}\mathbf{1}[r_i = r_j]$. Harrell's concordance index is

$$\widehat{C} = \frac{\sum_{i,j} e_i \mathbf{1}[t_i < t_j] \psi(r_i, r_j)}{\sum_{i,j} e_i \mathbf{1}[t_i < t_j]}, \tag{13}$$

where comparable pairs are defined using continuous DSS times rather than interval indices.

C Model and Baseline Training

C.1 ES4Surv training

Each cross-validation fold is trained in two stages. First, one pooled model is optimized on the union of the five training cohorts for one epoch. We select its checkpoint using the unweighted macro validation C-index across the five cohorts. The fold-matched pooled checkpoint then initializes five independent one-epoch cohort-specific models. No cohort-specific model initializes another. For each fold and cohort, we report the better validation C-index obtained by the pooled checkpoint and its matched cohort-specific checkpoint. This model selection uses validation data and is not an additional held-out test estimate.

The exact training objective is

$$\mathcal{L} = \mathcal{L}_{\text{LM}} + 5\mathcal{L}_{\text{surv}} + \mathcal{L}_{\text{group}} + 0.02\mathcal{L}_{\text{con}}. \tag{14}$$

Here, $\mathcal{L}_{\text{LM}}$ is the teacher-forced token loss, $\mathcal{L}_{\text{surv}}$ supervises the final prediction, $\mathcal{L}_{\text{group}}$ supervises the auxiliary group prediction, and $\mathcal{L}_{\text{con}}$ aligns the final group centers with the fold-specific concept targets.

Table 7: Optimization settings for ES4Surv and the matched WSI-LLaVA baselines. Each cohort-specific model starts independently from its fold-matched pooled checkpoint.

Setting	Pooled stage	Cohort stage
Epochs	1	1
Batch size / accumulation	$4/16$	$4/16$
Main / visual-projector LR	$10^{-4}/2\times10^{-5}$	$10^{-4}/2\times10^{-5}$
Warmup ratio	0.02	0.01
Evaluation / save interval	20 steps	20 steps
LoRA rank / scale / dropout	$8/16/0.05$	$8/16/0.05$

Both stages use AdamW ($\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-8}$), weight decay 0.01, gradient clipping at 1.0, a constant schedule after warmup, BF16 with TF32, gradient checkpointing, DeepSpeed ZeRO-3, and seed 42. Trainable parameters include the LoRA adapters, stage-control tokens, initial query embeddings, visual and query projectors, group centers, refinement blocks, FiLM, routing function, and survival heads. The GigaPath features, dense decoder weights outside LoRA, and fold-specific concept targets are frozen. At inference, we retain the 15,000-entry patch-bank cap, deterministic assignments, and the same top-4 groups by top-8 patches selection used during training.

C.2 MATCHED WSI-LLAVA BASELINES

The mean-pooling and survival-token baselines use the same patient-disjoint folds, global GigaPath slide tokens, WSI-LLaVA-7B backbone, visual projector, LoRA configuration, 11-bin survival objective, and pooled-to-cohort training protocol as ES4Surv. They do not access the high-resolution patch bank and do not use adaptive groups, concept alignment, the auxiliary group loss, or patch re-entry. The mean-pooling baseline applies its survival head to the mean of the image-conditioned assistant-prefix states. The survival-token baseline instead reads the causal state immediately before `<sur>`, matching the final prediction interface of ES4Surv without local evidence seeking. For both baselines, reporting selects the better validation result between the pooled checkpoint and its independently specialized cohort checkpoint.

C.3 MIL AND VISION–LANGUAGE BASELINES

ABMIL, TransMIL, DeepAttnMISL, PatchGCN, Surformer, and the histology-only SlotSPE setting use the same patient-disjoint folds, GigaPath patch features, 11-bin incidence target, and censoring-aware likelihood as ES4Surv. For every model and fold, we train for five pooled multi-cohort epochs, select the pooled checkpoint by macro validation C-index, and independently fine-tune one model per cohort for five epochs. The best cohort-specific validation checkpoint supplies the reported prediction.

We preserve model-appropriate optimization settings rather than imposing one optimizer on unlike architectures. ABMIL, TransMIL, DeepAttnMISL, and PatchGCN use AdamW with learning rate 10^{-5}; SlotSPE uses Adam with learning rate 5×10^{-4}; and Surformer uses Adam with learning rate 3×10^{-4}. All use weight decay 0.01 and seed 42. VLSA is trained independently for each cohort and fold with its native CONCH representation and released cohort-specific configuration; its checkpoint is selected by validation C-index. Thus, all methods share the endpoint, patient splits, and evaluation protocol, while each model class retains its appropriate training recipe.

D CONCH–MEDGEMMA CONCEPT-EXEMPLAR BANK

D.1 OFFLINE BANK CONSTRUCTION

We construct a separate concept-exemplar bank for every cross-validation fold using only that fold's training WSIs. The vocabulary follows the 13 histologic concepts used by HistoSelect (Huang et al., 2026): malignant tissue, benign tissue, stroma, lymphocytes, necrosis, adipose tissue, tissue artifact, blood vessel, extracellular mucin, nerve, hemorrhage, smooth muscle, and plasma cells. Validation patients and slides are excluded from every step.

For each concept, we encode five H&E morphology prompts with CONCH-v1 (Lu et al., 2024). Training WSIs are visited in a deterministic seed-7 order. Each eligible patch is assigned to the concept with the highest average similarity across its five prompts. For each concept represented on a slide, we retain the top half of assigned patches by CONCH score, shuffle this candidate set deterministically, and inspect at most 50 candidates. Out-of-bounds coordinates are rejected. We also reject patches with estimated white-space fraction above 0.5, except for adipose tissue and tissue artifact, for which this filter would remove valid morphology.

For every remaining candidate, MedGemma-1.5-4B-IT (Sellergren et al., 2026) receives the single H&E patch and a concept-specific binary morphology question. A patch is accepted only when the response is exactly `Yes`. Tissue artifact bypasses this screen and is selected directly from the CONCH candidates because technical degradation is itself the target appearance. If a slide exhausts its candidate budget before yielding an accepted patch, construction proceeds to the next training WSI; this bounds the search on any one slide.

The process continues until each concept has one accepted patch from 10 distinct training WSIs in each of BLCA, BRCA, GBMLGG, LUAD, and UCEC. The resulting fold-specific bank contains 50 exemplars per concept and 650 records:

$$A = \{A_{c,d,r}\} \in \mathbb{R}^{13\times5\times10\times1536}, \qquad A_{c,d,r} \in \mathbb{R}^{1536}. \tag{15}$$

Each accepted CONCH coordinate is mapped to its corresponding raw GigaPath feature. For concept c, the frozen alignment target is the equally weighted mean of the five cohort means,

$$a_c = \frac{1}{5}\sum_{d=1}^{5}\left(\frac{1}{10}\sum_{r=1}^{10}A_{c,d,r}\right) \in \mathbb{R}^{1536}. \tag{16}$$

CONCH and MedGemma are used only to build these fold-specific targets; neither model is invoked during ES4Surv training or inference.