VCRE-FIB: VIEW-CONDITIONED REGIONAL EVIDENCE FOR FINE-GRAINED ULTRASOUND GRADING OF *Schistosoma japonicum*-ASSOCIATED LIVER FIBROSIS

Anonymous authors
Paper under double-blind review

ABSTRACT

Accurate assessment of *Schistosoma japonicum*-associated liver fibrosis is essential for disease management and long-term follow-up in endemic regions. Ultrasound provides non-invasive imaging, but complex local echogenic patterns and anatomical structures make fine-grained grading challenging. Existing deep learning methods can predict fibrosis scores, yet directly incorporating acquisition views and regional cues into grading while retaining spatial information for inspection remains an open problem. Here we present VCRE-Fib, a view-conditioned regional evidence framework that integrates anatomical context, local information, and global image assessment for fine-grained ultrasound grading. The framework forms view-conditioned local grading evidence before spatial pooling, uses weak localization to guide its aggregation, and combines it with global predictions. Image-only inference jointly returns a fibrosis score, acquisition view, and candidate abnormal-region map. We developed and evaluated the method on a re-curated cohort of 108,709 ultrasound images from 6,373 patients across 35 centers. On a patient-disjoint test set of 4,107 images from 240 patients across four centers, VCRE-Fib reduced the prespecified composite grading risk by 7.115% relative to SFibAI trained and evaluated on the same data split. Image-level mean absolute error decreased from 0.391 to 0.378, alongside lower patient-max, patient-median, and center-balanced risks. The full model also achieved lower composite grading risk than variants that separately removed view conditioning or weak localization. These results support incorporating anatomical context and regional evidence into ultrasound grading while exposing spatial predictions for inspection alongside severity estimates.

1 INTRODUCTION

Liver fibrosis caused by *Schistosoma japonicum* is a persistent concern in disease management in endemic regions and can be accompanied by portal hypertension and severe hepatosplenic complications (Colley et al., 2014; McManus et al., 2018). In endemic settings, including China, assessment of established hepatic morbidity complements infection control (McManus et al., 2010; Lo et al., 2022). Longitudinal evidence shows that severe parenchymal fibrosis can persist for years despite antiparasitic treatment (Carlton et al., 2010). Repeatable imaging assessment of disease severity therefore provides an important foundation for population screening, assessment of morbidity, and long-term follow-up.

B-mode ultrasound non-invasively depicts hepatic parenchymal and portal abnormalities, providing an imaging tool suitable for repeated assessment. However, the ultrasound appearances of *S. japonicum*-associated fibrosis have complex spatial patterns: interseptal fibrosis can produce network-like or fish-scale-like echoes, whereas portal abnormalities have a different anatomical distribution. The Basel ultrasonography protocol emphasizes distinguishing these patterns (Richter et al., 2025). These features make local echoes, structural distribution, and anatomical context relevant to severity assessment. For automated grading, this raises two linked requirements: extracting information that

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.

supports fine-grained scoring from complex images, and retaining relevant spatial cues as outputs that can be compared with the original image.

Computational approaches have progressed from ultrasound texture analysis to radiomics and deep convolutional models, providing a technical foundation for reducing variation in manual interpretation (Yeh et al., 2003; Guo et al., 2024; Lee et al., 2020). SFibAI specifically represents ultrasound grading of *S. japonicum*-associated liver fibrosis using 36 ordered score levels from 0.0 to 3.5 at intervals of 0.1, while retaining their correspondence to the four clinical grades (Xu et al., 2026). This representation provides a more detailed description of severity than coarse categories and establishes a basis for learning fine-grained scores directly from images. We retain this task representation and grading loss, use SFibAI as the direct baseline, and investigate how anatomical views and local regions can be organized and used within the scoring process.

An accurate score derived from whole-image features does not directly identify the acquisition view recognized by the model, the regions exhibiting relevant appearances, or how those regions contribute to the final prediction. Global representations can contain this information, but a score alone leaves users unable to inspect view predictions and regional responses separately. Acquisition-view labels and local lesion annotations provide complementary supervision: the former describes anatomical context, whereas the latter identifies locations of interest. Using these signals for grading requires an explicit computational connection between anatomical context, local features, and image-level judgment, while accommodating annotations that cover only part of an abnormal region.

Multitask learning can jointly learn grading and spatial tasks through shared representations (Caruana, 1997), but representation sharing alone does not specify how predicted views and localization outputs enter the scoring computation. Post-hoc attribution methods such as Grad-CAM can highlight regions associated with a prediction (Selvaraju et al., 2017), while generating these visualizations leaves the original scoring pathway unchanged. We therefore ask: *can acquisition views and local positions be organized as regional evidence that directly participates in prediction, improving fine-grained grading while retaining spatial outputs for human inspection?* Our design principle is to introduce view conditioning before spatial aggregation, allowing local features to form grading evidence within their anatomical context before combining them with whole-image judgment.

To implement this design, we propose VCRE-Fib (View-Conditioned Regional Evidence for Fibrosis Grading), illustrated in Figure 1. The framework preserves a global scoring pathway and uses view labels and local lesion boxes to learn view prediction and weak localization. It generates view-conditioned regional grading evidence before spatial pooling, uses weak localization to guide regional aggregation, adds the resulting correction to the global prediction, and mixes the conditional outputs according to predicted view probabilities. This structure explicitly connects the learning, use, and presentation of spatial information: view predictions determine the mixture of conditional outputs, localization responses participate in regional aggregation, and both remain available as inspectable outputs. Inference requires only an image and returns a fibrosis score, an acquisition view, and a localization map.

We compare VCRE-Fib with SFibAI and two independently trained module-deletion variants under a common data split and evaluation protocol. On a test set of 4,107 images from 240 patients across four centers, the full model achieves the lowest prespecified composite risk in this comparison ($R_{\text{final}} = 0.186865$), a 7.115% reduction relative to SFibAI, together with the lowest image-level, patient-max, patient-median, and center-balanced COR. The module-deletion comparison also shows that lower grading risk does not coincide with uniformly better spatial-task metrics, motivating joint assessment of scoring performance and spatial output quality. These results support further investigation of view-conditioned regional evidence in this task and provide a concrete framework linking fine-grained grading with inspectable spatial information. Its practical benefit for clinician review and clinical decisions remains to be evaluated.

2 Related Work

Schistosomiasis-related fibrosis and ultrasound grading. Schistosome infection and established organ morbidity require complementary assessment: parasitological, serological, and molecular tests address infection, while imaging characterizes hepatosplenic abnormalities (Weerakoon et al., 2015; Sah et al., 2015). Liver biopsy provides histological information but is invasive, motivating

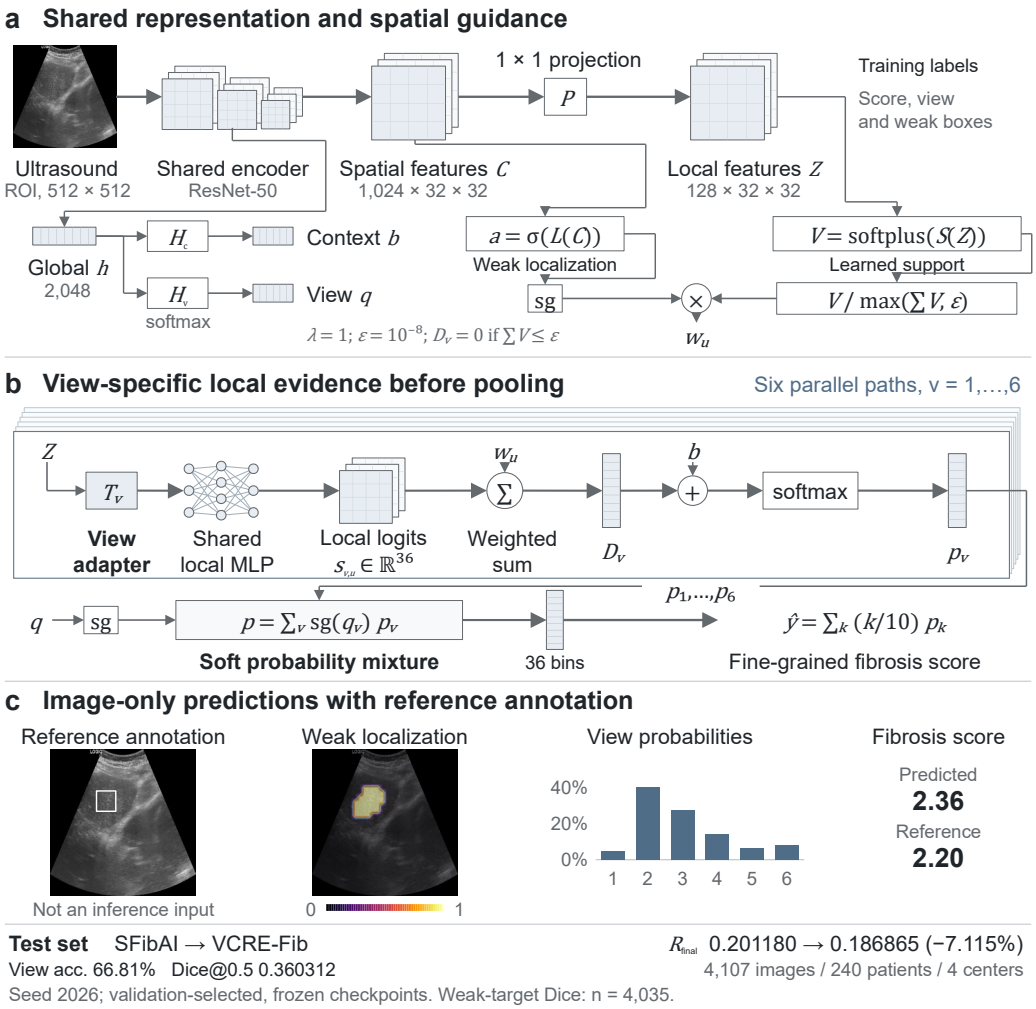

Figure 1: **VCRE-Fib: view conditioning before spatial aggregation.** (a) A shared encoder supplies global context, view probabilities, weak localization and local features. (b) Six view-conditioned paths produce regional logit corrections, weighted by localization and learned support. Conditional distributions share context logits and are mixed by detached view probabilities. (c) One test case and test results from frozen, validation-selected checkpoints. Feature cells are schematic; sg denotes stop-gradient. Reference boxes are not inference inputs, and localization does not represent exhaustive segmentation.

non-invasive assessment; the limitations of fibrosis tests also require care when interpreting their outputs (Bravo et al., 2001; Patel & Sebastiani, 2020). The target in this study is an expert-assigned ultrasound appearance score.

For *S. japonicum*, the anatomical distribution of fibrosis is particularly relevant. A longitudinal cohort linked ultrasound-detectable parenchymal fibrosis to infection history and documented limited reversal of severe changes after treatment (Carlton et al., 2010). The Basel protocol distinguishes portal and interseptal patterns when assessing Asian schistosomiasis (Richter et al., 2025). Together with the imaging literature (Sah et al., 2015), these findings motivate preserving anatomical context and visible regional appearances when predicting severity.

Computational assessment has progressed from B-mode texture features and support vector machines (Yeh et al., 2003) to multiparametric ultrasomics (Li et al., 2019), schistosomiasis-focused radiomics (Guo et al., 2024), and convolutional classification with METAVIR-derived targets (Lee et al., 2020).

SFibAI supplies the fine-grained score representation used here (Xu et al., 2026). Acquisition, labels, and cohorts differ across these studies. Our direct comparison therefore uses SFibAI on the same cohort splits, while investigating how view and lesion-location cues can participate in grading.

Anatomical context and regional evidence learning. Multitask learning shares representations across related objectives (Caruana, 1997); SonoNet demonstrates standard-plane recognition and anatomical localization in fetal ultrasound (Baumgartner et al., 2017). Regional aggregation offers a complementary way to connect local information to an image-level decision. Attention-based MIL learns aggregation weights (Ilse et al., 2018); pathology methods extend this idea with instance clustering in CLAM (Lu et al., 2021), instance relationships in TransMIL (Shao et al., 2021), and additive local contributions in Additive MIL (Javed et al., 2022). FiLM provides a general mechanism for conditioning feature transformations (Perez et al., 2018). VCRE-Fib draws on shared learning, conditioning, and aggregation to interpret regional features in their acquisition-view context before combining them with global judgment. Its regions belong to one ultrasound image, and its supervised view and localization outputs also enter the grading computation.

Weak localization and inspectable predictions. BoxSup and BoxInst learn segmentation from boxes without pixel-level masks (Dai et al., 2015; Tian et al., 2021). Here, boxes mark local abnormal patches and may not enclose the entire lesion, so responses beyond them remain candidates for clinical review. Concept bottleneck models expose supervised intermediate concepts (Koh et al., 2020); VCRE-Fib exposes view and localization outputs while retaining a global route. Grad-CAM, Grad-CAM++, and integrated gradients provide post-hoc attribution approaches (Selvaraju et al., 2017; Chattopadhay et al., 2018; Sundararajan et al., 2017). Our trained localization participates in regional aggregation, but this role alone does not establish faithful attribution. Saliency sanity checks motivate evaluating explanations beyond visual plausibility (Adebayo et al., 2018). The present clinician-annotation comparisons support inspection of spatial correspondence.

3 VCRE-Fib: View-Conditioned Regional Evidence

3.1 Task and design overview

Given an ultrasound ROI image x, VCRE-Fib predicts a severity distribution, acquisition view, and weak-localization map. Training uses grading labels y, six-class view labels, and lesion-location annotations; inference requires neither view nor box annotations. We retain the SFibAI score grid $g_k = k/10$, $k = 0, \ldots, 35$, and four-grade thresholds of 0.5, 1.5, and 2.5 (Xu et al., 2026). The design in Figure 1 learns spatial cues, forms regional evidence in its view context, and combines that evidence with global judgment.

3.2 Global judgment and spatial cues

A global path preserves whole-image context so that regional evidence complements an overall prediction. A ResNet-50 encoder (He et al., 2016) produces $h \in \mathbb{R}^{2048}$ and layer-3 features $C \in \mathbb{R}^{1024\times H\times W}$. With 512×512 inputs, $H = W = 32$. The context logits are $b = H_c(h) \in \mathbb{R}^{36}$, view probabilities are $q = \mathrm{softmax}(H_v(h)) \in \mathbb{R}^6$, and weak localization is $a = \sigma(L(C))$. The view and localization predictions are learned through auxiliary supervision and retained as spatial outputs for inspection.

Local boxes identify clinician-marked regions without specifying exhaustive contours. The implemented weak-box objective combines an inside-response term with a constraint on the outside response fraction. It guides spatial learning but supplies no complete lesion-contour target. Whether learned responses extend beyond a local annotation and correspond to broader pathology is therefore an empirical question examined through actual cases.

3.3 Conditioning before regional aggregation

Local features are interpreted in their anatomical context. A 1×1 projection produces $Z = P(C) \in \mathbb{R}^{128\times H\times W}$. Six view adapters T_v, each a $128 \to 128$ pointwise convolution and GELU, share a

$128 \to 64 \to 36$ MLP f:

$$s_{v,u} = f(T_v(Z)_u), \quad v = 1, \ldots, 6, \quad u = 1, \ldots, HW. \tag{1}$$

Conditioning occurs before pooling and thus changes the grading evidence at each location. The vector $s_{v,u}$ is a local logit contribution candidate, not a separately calibrated regional severity score.

3.4 Localization-guided regional grading

Weak localization supplies spatial relevance, while a support head $V = \mathrm{softplus}(S(Z))$ learns aggregation weights through grading. They organize the regional correction:

$$c_{v,u} = \lambda \frac{V_u \, \mathrm{sg}(a_u)}{\max(\sum_{u'} V_{u'}, \epsilon)} s_{v,u}, \qquad D_v = \sum_u c_{v,u}. \tag{2}$$

The reference configuration uses $\lambda = 1$ and $\epsilon = 10^{-8}$, with zero correction when total support is at most ϵ. Normalization by $\sum V$ allows uniformly low localization responses to attenuate the correction. The operator sg stops gradients at the indicated interface.

Regional evidence is added to global context, and view probabilities retain uncertainty across conditional judgments:

$$p_v = \mathrm{softmax}(b + D_v), \qquad p = \sum_v \mathrm{sg}(q_v) p_v, \qquad \hat{y} = \sum_k g_k p_k. \tag{3}$$

The model mixes conditional probabilities. Localization therefore participates in the grading path while remaining an explicit output.

3.5 Supervision and inference outputs

Following shared-representation multitask learning (Caruana, 1997), the training objective is

$$\mathcal{L} = \mathcal{L}_{\mathrm{grade}}(p, y) + 0.1\mathcal{L}_{\mathrm{view}} + 0.1\mathcal{L}_{\mathrm{box}}. \tag{4}$$

View prediction uses cross-entropy; the weak-box loss uses valid local boxes, as specified in Appendix A. The grading interface consumes $\log p$. Gradients are stopped at the grading uses of q and a, while auxiliary losses still update the encoder and grading still trains the context, local-evidence, and support paths. VCRE-Fib uses the same hybrid grading-loss formulation and weights as SFibAI: both compute MSE and the clinical-grade boundary penalty using the real-valued scores $\hat{y} = 0.1 \sum_k k p_k$ and $y = 0.1j$, where j is the target bin. Appendix A gives the common grading-loss formulation.

Inference returns $\hat{y}$, the view $\arg\max_v q_v$, and localization a. Side-by-side original images, clinician annotations, and predictions expose the spatial outputs available for review. Output completeness here means retaining both view and regional information alongside the score; it does not mean complete lesion coverage.

4 Experimental Design

Data and supervision. Our cohort was re-curated from the institutional ultrasound database described by Xu et al. (2026). The prediction target is the existing expert-assigned ultrasound appearance score. We cleaned and organized these scores and the existing clinician-drawn lesion boxes, and recovered six-class view labels from acquisition-plane identifiers. The cohort contains 108,709 ROI images from 6,373 patients across 35 centers. Training, validation, and test comprise 83,722/4,906, 20,880/1,227, and 4,107/240 images/patients, respectively. Patients do not overlap across splits; all four test centers are represented in training. All models use the same grading data and splits, with spatial supervision determined by retained modules. Boxes at the highest recorded score form weak targets, not exhaustive contours. All 4,107 test images have view annotations, and 4,035 meet weak-target evaluation criteria. Appendix A.1 describes annotation provenance; Table S7 lists view names and counts.

Compared models. We compare SFibAI, full VCRE-Fib, and two deletion variants (Table 1). *w/o View* retains weak localization and regional grading but removes the view head, view-specific adapters, posterior mixture, and view loss. *w/o Weak Loc.* retains view-conditioned grading but removes the localization head and weak-box loss, setting $a_u = 1$ in Equation 2. Both retain the support and global paths and are trained independently, rather than fine-tuned from the full model. The contrasts evaluate whole model designs; they do not isolate supervision, capacity, and information routing from one another.

Table 1: Compared models and inference outputs. All use image-only inference, the same backbone initialization and grading objective. Deletion variants retain the regional grading path and are trained independently.

Method	View module	Weak-localization module	Outputs
SFibAI	Absent	Absent	Score
w/o View	Removed	Retained	Score + region
w/o Weak Loc.	Retained	Removed	Score + view
VCRE-Fib	Retained	Retained	Score + view + region

Training, selection, and freezing. All four use the same random seed (2026), ResNet-50, ImageNet-1K-V2 initialization, 512×512 inputs, global batch 24, bfloat16 training, AdamW (Loshchilov & Hutter, 2019), and the same real-score hybrid grading objective. All methods, including full VCRE-Fib, use a common 90-epoch cap for the training comparison. The reported trajectories cover epochs 1–90, and validation selects from epochs 21–90 by lower R_{final}, lower image COR, then earlier epoch. The common recipe is specified in Appendix A. Checkpoints, configurations, and the endpoint are frozen before test comparison; test determines the observed ranking.

Evaluation questions. We ask whether each spatially informed variant improves on the baseline, how jointly using the modules changes grading and spatial outputs, and what information the outputs expose. The lower-is-better primary endpoint is

$$R_{\text{final}} = 0.4R_{\text{image}} + 0.4R_{\text{patient-max}} + 0.2R_{\text{center}}. \tag{5}$$

Each component uses the locked clinical objective risk (COR); the center term weights per-center patient-max COR by square-root patient counts. Patient-median is reported separately. Differences describe observed results, without significance testing. Bold table values indicate the best unrounded result per metric, including exact ties.

Spatial outputs and visualization. View metrics include accuracy/macro-F1; weak-target Dice/IoU use threshold 0.5. Four selected test cases spanning four true views illustrate the outputs, with two validation examples in the appendix. Case-selection criteria, patient deduplication, and common display settings are specified in Appendix C.2.

5 Grading and Spatial Prediction Results

5.1 Test set primary results

We tested whether spatial information improves fine-grained grading. On the 4,107-image, 240-patient, four-center test set, full VCRE-Fib achieves the lowest prespecified composite risk ($R_{\text{final}} = 0.186865$), followed by w/o Weak Loc. (0.191308), w/o View (0.193799), and SFibAI (0.201180; Table 2). Their respective absolute differences from SFibAI are -0.014315, -0.009872, and -0.007380, corresponding to reductions of 7.115%, 4.907%, and 3.668%, computed before rounding. All spatial variants outperform SFibAI on this endpoint. Results use validation-selected, frozen checkpoints.

5.2 Module deletions reveal distinct risk profiles

Independently trained deletion variants show distinct advantages. The view-retaining w/o Weak Loc. has lower patient-max COR, whereas the localization-retaining w/o View has lower image

Table 2: Test set primary results: 4,107 images, 240 patients, four centers. All checkpoints were frozen after validation selection. All risks are lower-is-better. Bold denotes the best unrounded value per metric; exact ties are all bold.

Rank	Method	R_{final}	Image COR	Max COR	Median COR	Center COR
1	VCRE-Fib	**0.186865**	**0.177908**	**0.200528**	**0.154344**	**0.177453**
2	w/o Weak Loc.	0.191308	0.186343	0.202422	0.165844	0.179008
3	w/o View	0.193799	0.181518	0.210002	0.158449	0.185957
4	SFibAI	0.201180	0.185938	0.219742	0.162587	0.194539

and patient-median COR (Table 2). The full model leads all three risks and center-balanced COR, improving composite risk by 0.004443 over the better variant, w/o Weak Loc. Gains thus span risk measures. These design-level comparisons do not isolate supervision, capacity, or information use, or establish statistical superiority or module synergy.

5.3 Grading gains extend across images and patients

Specific grading metrics improve alongside composite risk. Relative to SFibAI, full VCRE-Fib reduces image MAE from 0.391313 to 0.377857, increases four-grade accuracy from 64.11% to 67.03%, and reduces severe errors from 9.98% to 9.37% (Table 3). The full model also leads the remaining image metrics: RMSE, TMAE, COR, ± 0.3 and ± 0.5 accuracy, and macro-F1. The observed gains therefore span score error, agreement within tolerances, and four-grade discrimination.

Table 3: Image-level test metrics; 4,107 images per model. Bold denotes the best unrounded value per metric; exact ties are all bold.

Metric	SFibAI	w/o View	w/o Weak Loc.	VCRE-Fib
MAE ↓	0.391313	0.384734	0.392205	**0.377857**
RMSE ↓	0.568137	0.564076	0.567780	**0.553435**
Acc.$\pm$0.3 (%) ↑	56.17	57.51	56.05	**57.95**
Acc.$\pm$0.5 (%) ↑	69.61	70.08	69.47	**70.15**
4-grade acc. (%) ↑	64.11	66.13	64.13	**67.03**
Macro-F1 ↑	0.639702	0.663674	0.643722	**0.671466**
TMAE ↓	0.129177	0.127592	0.129382	**0.123100**
Severe error (%) ↓	9.98	10.03	9.76	**9.37**
COR ↓	0.185938	0.181518	0.186343	**0.177908**

The full model also has the lowest patient-max, patient-median, and center-balanced COR (Table 2). Some thresholded accuracies favor deletion variants; Appendix B.2 reports the complete patient-level metrics and these differences.

5.4 Lower grading risk accompanies spatial-task trade-offs

We next examined the accompanying spatial predictions. Full VCRE-Fib reaches view accuracy of 66.81%, macro-F1 of 0.665861, and CE of 0.916542 on 4,107 test images, with weak-target Dice/IoU of 0.360312/0.235487 on 4,035 eligible images (Table 4). Yet w/o Weak Loc. achieves higher view accuracy and macro-F1 (75.70% and 0.757500), and w/o View achieves higher Dice/IoU (0.430661/0.298825). Thus, the lowest grading risk does not coincide with the best spatial-task scores. Only the full model provides all three outputs, with auxiliary-task trade-offs. Weak-target overlap measures agreement with coarse local annotations, not complete lesion segmentation.

5.5 Joint outputs expose grading–spatial disagreements

Figure 2 illustrates four test cases spanning distinct true views, juxtaposing ultrasound, clinician annotation, and localization with reference/predicted scores and full view names. Rows (a)–(c)

Table 4: Spatial test outputs. N/A indicates an absent head. Weak-target N counts eligible images, not rectangles; Dice/IoU measure agreement with local annotations, not complete lesion segmentation. Bold denotes the best unrounded value per metric; exact ties are all bold.

Metric	SFibAI	w/o View	w/o Weak Loc.	VCRE-Fib
View N	N/A	N/A	4107	4107
View acc. (%) ↑	N/A	N/A	**75.70**	66.81
View macro-F1 ↑	N/A	N/A	**0.757500**	0.665861
View CE ↓	N/A	N/A	**0.690513**	0.916542
Weak-target N	N/A	4035	N/A	4035
Dice@0.5 ↑	N/A	**0.430661**	N/A	0.360312
IoU@0.5 ↑	N/A	**0.298825**	N/A	0.235487

show localization correspondence. In (a), scores agree closely (2.4/2.409), yet the true left subcostal transverse view is classified as subxiphoid sagittal. Close score agreement thus need not imply a correct spatial judgment; the joint outputs make this disagreement visible.

Row (d), with scores 1.4/1.491, shows responses outside local rectangles under the fixed scale and threshold; two validation examples appear separately in the appendix. Responses outside annotations offer locations to inspect but may represent relevant tissue or erroneous activation, establishing neither complete lesion coverage nor clinical correctness. All qualitative case predictions were generated with the frozen checkpoint evaluated quantitatively. They illustrate inspectable information, without assigned clinical interpretations or claims of evaluated clinical-review benefit.

5.6 Computational cost of joint prediction

We quantified the cost of providing grading and both spatial outputs. Matched FP32 measurements on an A100-SXM4-80GB yield median batch-one latencies of 5.620, 6.171, 7.122, and 7.371 ms for SFibAI, w/o View, w/o Weak Loc., and full VCRE-Fib. The full model adds 7.81% parameters, 4.24% counted MACs, and 1.751 ms median latency over SFibAI. Appendix Table S6 reports latency, P95, and memory at batch sizes one and eight.

6 Discussion

Grading risk, view accuracy, and weak-target overlap assess severity prediction, acquisition-view recognition, and agreement with local annotations, respectively. The full model's lowest grading risk and the deletion variants' higher spatial-task scores motivate evaluating the score and each accompanying output together. The module-deletion comparisons evaluate whole designs rather than isolating supervision, capacity, or information use, leaving the mechanisms behind these trade-offs unresolved.

Local rectangular boxes provide positional cues for spatial learning. Comparing regional maps with the original image and clinician boxes makes correspondence to these cues visible and identifies additional regions for inspection. Weak-target overlap measures positional agreement, whereas complete lesion coverage is a separate objective. Responses outside local boxes may reflect relevant tissue or erroneous activation; more complete spatial references and clinician review could distinguish these possibilities.

For clinical review, the joint outputs provide three objects to inspect: the predicted view, regional response, and final score. The illustrated view error despite close score agreement shows how these outputs expose different aspects of a prediction. Displaying them alongside the original image would allow separate checks of anatomical context and local responses. Their practical benefit remains to be evaluated through reader studies of error identification, appropriate assessment revisions, and review time.

Deployment studies should measure preprocessing, inference, and display together on target hardware, beyond the model-only costs reported here. Independent-cohort evaluation should test grading and spatial outputs across different patient populations and acquisition settings.

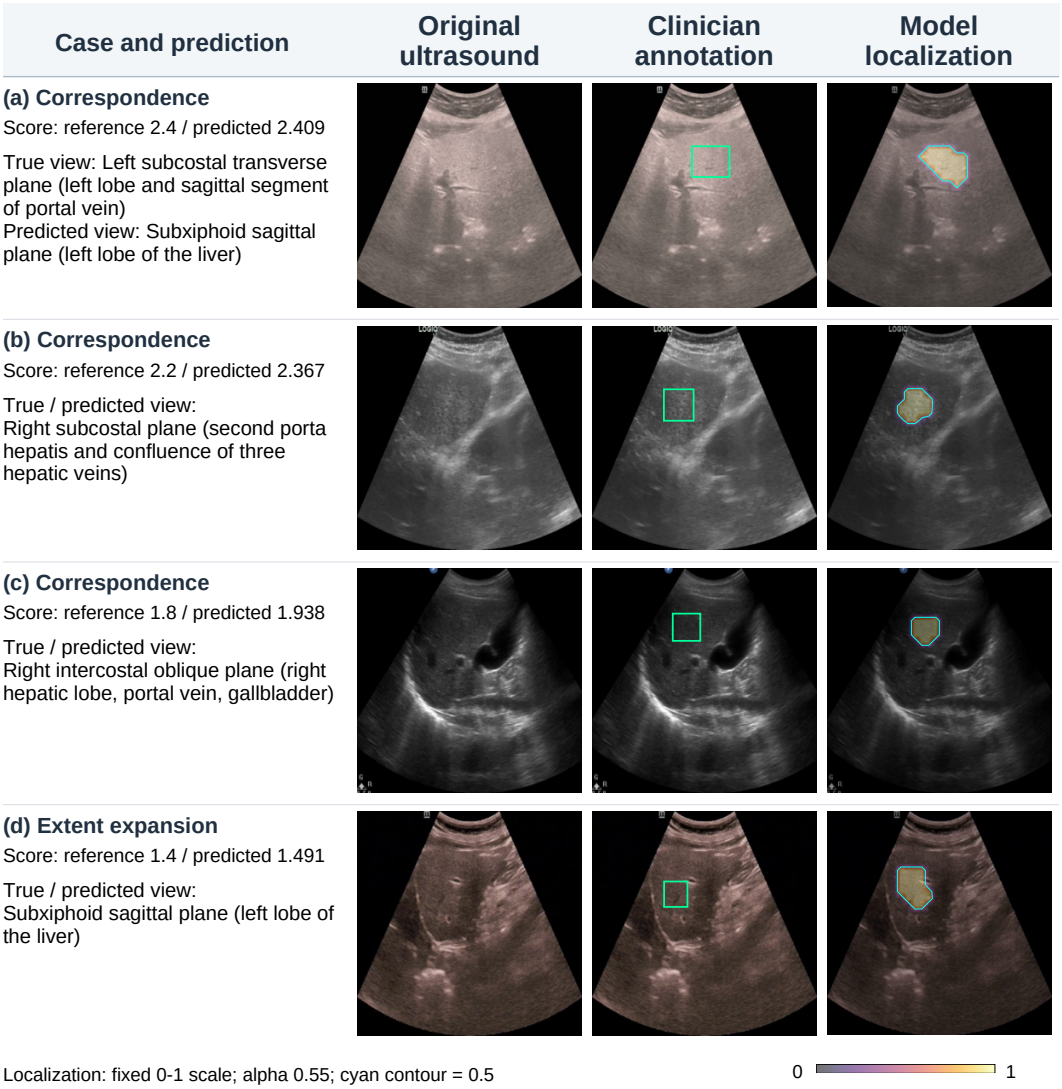

Figure 2: Four selected test cases, with reference and predicted scores and acquisition views at left. Predictions use the frozen checkpoint evaluated quantitatively. Rows (a)–(c) show localization correspondence; (d) shows a response beyond local boxes. Green boxes denote clinician annotations; localization uses a $[0, 1]$ scale, opacity 0.55, and a cyan 0.5 contour. In row (a), the predicted view differs from the reference. These selected examples illustrate spatial outputs without establishing population localization performance or complete lesion extent.

7 Conclusion

VCRE-Fib combines view-conditioned regional evidence with a global prediction for fine-grained ultrasound grading of *Schistosoma japonicum*-associated liver fibrosis. In the frozen-checkpoint comparison, the full model reduced composite test risk by 7.115% relative to SFibAI while returning score, view, and regional outputs. The full model's grading gains coexisted with higher view accuracy or weak-target overlap in the corresponding deletion variants. This trade-off makes the quality of each spatial output a distinct part of evaluating the joint system, alongside its final grading performance. The joint outputs make anatomical and regional judgments available for inspection alongside the score. Further evaluation should examine their usefulness for clinician review and their performance in independent cohorts.

Ethics Statement

The retrospective use of ultrasound images and associated annotations in this study was covered by institutional ethics approval. Written informed consent was obtained from all participants before enrolment.

Reproducibility Statement

The architecture and evaluation protocol are specified in Sections 3–4. Appendix A details data preparation and the shared training setup; Appendix B reports complete test and resource metrics, Appendix C defines case selection and display, and Appendix D provides checkpoint selection and all four training trajectories.

AI Use Statement

OpenAI Codex assisted with implementation inspection, manuscript drafting, translation, and editing, reference checking, document compilation, and presentation of existing results and case panels. It did not generate experimental observations. The authors are responsible for the scientific content and interpretation.

References

Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. In *Advances in Neural Information Processing Systems*, volume 31, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/294a8ed24b1ad22ec2e7efea049b8737-Abstract.html.

Christian F. Baumgartner, Konstantinos Kamnitsas, Jacqueline Matthew, Tara P. Fletcher, Sandra Smith, Lisa M. Koch, Bernhard Kainz, and Daniel Rueckert. SonoNet: Real-time detection and localisation of fetal standard scan planes in freehand ultrasound. *IEEE Transactions on Medical Imaging*, 36(11):2204–2215, 2017. doi: 10.1109/TMI.2017.2712367. URL https://doi.org/10.1109/TMI.2017.2712367.

Arturo A. Bravo, Sunil G. Sheth, and Sanjiv Chopra. Liver Biopsy. *New England Journal of Medicine*, 344(7):495–500, 2001. doi: 10.1056/NEJM200102153440706. URL https://doi.org/10.1056/NEJM200102153440706.

Elizabeth J. Carlton, Michelle Hsiang, Yi Zhang, Sarah Johnson, Alan Hubbard, and Robert C. Spear. The Impact of Schistosoma japonicum Infection and Treatment on Ultrasound-Detectable Morbidity: A Five-Year Cohort Study in Southwest China. *PLoS Neglected Tropical Diseases*, 4(5):e685, 2010. doi: 10.1371/journal.pntd.0000685. URL https://doi.org/10.1371/journal.pntd.0000685.

Rich Caruana. Multitask learning. *Machine Learning*, 28:41–75, 1997. doi: 10.1023/A:1007379606734. URL https://doi.org/10.1023/A:1007379606734.

Aditya Chattopadhay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian. Grad-CAM++: Generalized Gradient-Based Visual Explanations for Deep Convolutional Networks. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*, pp. 839–847, 2018. doi: 10.1109/WACV.2018.00097. URL https://doi.org/10.1109/WACV.2018.00097.

Daniel G. Colley, Amaya L. Bustinduy, W. Evan Secor, and Charles H. King. Human schistosomiasis. *The Lancet*, 383(9936):2253–2264, 2014. doi: 10.1016/S0140-6736(13)61949-2. URL https://doi.org/10.1016/S0140-6736(13)61949-2.

Jifeng Dai, Kaiming He, and Jian Sun. BoxSup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1635–1643, 2015. URL

https://openaccess.thecvf.com/content_iccv_2015/html/Dai_BoxSup_Exploiting_Bounding_ICCV_2015_paper.html.

Zhaoyu Guo, Miaomiao Zhao, Zhenhua Liu, Jinxin Zheng, Yanfeng Gong, Lulu Huang, Jingbo Xue, Xiaonong Zhou, and Shizhu Li. Feasibility of ultrasound radiomics based models for classification of liver fibrosis due to Schistosoma japonicum infection. *PLOS Neglected Tropical Diseases*, 18(6): e0012235, 2024. doi: 10.1371/journal.pntd.0012235. URL https://doi.org/10.1371/journal.pntd.0012235.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016. URL https://openaccess.thecvf.com/content_cvpr_2016/html/He_Deep_Residual_Learning_CVPR_2016_paper.html.

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

Syed Ashar Javed, Dinkar Juyal, Harshith Padigela, Amaro Taylor-Weiner, Limin Yu, and Aaditya Prakash. Additive MIL: Intrinsically interpretable multiple instance learning for pathology. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/82764461a05e933cc2fd9d312e107d12-Abstract-Conference.html.

Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 5338–5348, 2020. URL https://proceedings.mlr.press/v119/koh20a.html.

Jeong Hyun Lee, Ijin Joo, Tae Wook Kang, Yong Han Paik, Dong Hyun Sinn, Sang Yun Ha, Kyunga Kim, Choonghwan Choi, Gunwoo Lee, Jonghyon Yi, and Won-Chul Bang. Deep learning with ultrasonography: automated classification of liver fibrosis using a deep convolutional neural network. *European Radiology*, 30(2):1264–1273, 2020. doi: 10.1007/s00330-019-06407-1. URL https://doi.org/10.1007/s00330-019-06407-1.

Wei Li, Yang Huang, Bo-Wen Zhuang, Guang-Jian Liu, Hang-Tong Hu, Xin Li, Jin-Yu Liang, Zhu Wang, Xiao-Wen Huang, Chu-Qing Zhang, Si-Min Ruan, Xiao-Yan Xie, Ming Kuang, Ming-De Lu, Li-Da Chen, and Wei Wang. Multiparametric ultrasomics of significant liver fibrosis: A machine learning-based analysis. *European Radiology*, 29(3):1496–1506, 2019. doi: 10.1007/s00330-018-5680-z. URL https://doi.org/10.1007/s00330-018-5680-z.

Nathan C Lo, Fernando Schemelzer Moraes Bezerra, Daniel G Colley, Fiona M Fleming, Mamoun Homeida, Narcis Kabatereine, Fatma M Kabole, Charles H King, Margaret A Mafe, Nicholas Midzi, Francisca Mutapi, Joseph R Mwanga, Reda M R Ramzy, Fadjar Satrija, J Russell Stothard, Mamadou Souncalo Traoré, Joanne P Webster, Jürg Utzinger, Xiao-Nong Zhou, Anthony Danso-Appiah, Paolo Eusebi, Eric S Loker, Charles O Obonyo, Reginald Quansah, Song Liang, Michel Vaillant, M Hassan Murad, Paul Hagan, and Amadou Garba. Review of 2022 WHO guidelines on the control and elimination of schistosomiasis. *The Lancet Infectious Diseases*, 22(11):e327–e335, 2022. doi: 10.1016/S1473-3099(22)00221-3. URL https://doi.org/10.1016/S1473-3099(22)00221-3.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Ming Y. Lu, Drew F. K. 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*, 5:555–570, 2021. doi: 10.1038/s41551-020-00682-w. URL https://doi.org/10.1038/s41551-020-00682-w.

Donald P. McManus, Darren J. Gray, Yuesheng Li, Zheng Feng, Gail M. Williams, Donald Stewart, Jose Rey-Ladino, and Allen G. Ross. Schistosomiasis in the People's Republic of China: the Era of the Three Gorges Dam. *Clinical Microbiology Reviews*, 23(2):442–466, 2010. doi: 10.1128/CMR.00044-09. URL https://doi.org/10.1128/CMR.00044-09.

Donald P. McManus, David W. Dunne, Moussa Sacko, Jürg Utzinger, Birgitte J. Vennervald, and Xiao-Nong Zhou. Schistosomiasis. *Nature Reviews Disease Primers*, 4:13, 2018. doi: 10.1038/s41572-018-0013-8. URL https://doi.org/10.1038/s41572-018-0013-8.

Keyur Patel and Giada Sebastiani. Limitations of non-invasive tests for assessment of liver fibrosis. *JHEP Reports*, 2(2):100067, 2020. doi: 10.1016/j.jhepr.2020.100067. URL https://doi.org/10.1016/j.jhepr.2020.100067.

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. URL https://ojs.aaai.org/index.php/AAAI/article/view/11671.

Joachim Richter, Andreas Neumayr, Amadou Garba-Djirmay, Hiroshi Ohmae, Ralph Aniceto, Xiao-Nong Zhou, Jing Xu, Zhaoyu Guo, An Ning, Edward Mberu Kamau, Francesca Tamarozzi, Hannah Wei Wu, Charles King, Birgitte Jyding Vennervald, Goylette F. Chami, Jürg Utzinger, and Christoph Hatz. The Basel ultrasonography protocol for assessing hepatosplenic pathologies in Asian schistosomiasis: report of a WHO expert meeting. *Infectious Diseases of Poverty*, 14(1):83, 2025. doi: 10.1186/s40249-025-01349-x. URL https://doi.org/10.1186/s40249-025-01349-x.

Vivek Kumar Sah, Liang Wang, Xiangde Min, Rajiv Rizal, Zhaoyan Feng, Zan Ke, Ming Deng, Liang Li, and Hongjun Li. Human schistosomiasis: A diagnostic imaging focused review of a neglected disease. *Radiology of Infectious Diseases*, 2(3):150–157, 2015. doi: 10.1016/j.jrid.2015.11.007. URL https://doi.org/10.1016/j.jrid.2015.11.007.

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. URL https://openaccess.thecvf.com/content_iccv_2017/html/Selvaraju_Grad-CAM_Visual_Explanations_ICCV_2017_paper.html.

Zhuchen Shao, Hao Bian, Yang Chen, Yifeng Wang, Jian Zhang, Xiangyang Ji, and Yongbing Zhang. TransMIL: Transformer based correlated multiple instance learning for whole slide image classification. In *Advances in Neural Information Processing Systems*, volume 34, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/10c272d06794d3e5785d5e7c5356e9ff-Abstract.html.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 3319–3328, 2017. URL https://proceedings.mlr.press/v70/sundararajan17a.html.

Zhi Tian, Chunhua Shen, Xinlong Wang, and Hao Chen. BoxInst: High-performance instance segmentation with box annotations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5443–5452, 2021. URL https://openaccess.thecvf.com/content/CVPR2021/html/Tian_BoxInst_High-Performance_Instance_Segmentation_With_Box_Annotations_CVPR_2021_paper.html.

Kosala G. A. D. Weerakoon, Geoffrey N. Gobert, Pengfei Cai, and Donald P. McManus. Advances in the Diagnosis of Human Schistosomiasis. *Clinical Microbiology Reviews*, 28(4):939–967, 2015. doi: 10.1128/CMR.00137-14. URL https://doi.org/10.1128/CMR.00137-14.

Ziyang Xu, Jianfeng Zhang, Tingting Wu, Haiyong Hua, Kun Yang, and Tieyong Zeng. Deep learning for precision grading of Schistosoma japonicum-induced liver fibrosis in ultrasound images. *Nature Communications*, 17(1):9998, 2026. doi: 10.1038/s41467-026-76287-9. URL https://doi.org/10.1038/s41467-026-76287-9.

Wen-Chun Yeh, Sheng-Wen Huang, and Pai-Chi Li. Liver fibrosis grade classification with B-mode ultrasound. *Ultrasound in Medicine & Biology*, 29(9):1229–1235, 2003. doi: 10.1016/S0301-5629(03)01010-X. URL https://doi.org/10.1016/S0301-5629(03)01010-X.

APPENDIX

A COHORT, ANNOTATIONS, AND IMPLEMENTATION DETAILS

A.1 COHORT PROVENANCE AND ANNOTATION SCOPE

Reference scores and cohort curation. The cohort was re-curated from the institutional ultrasound database described by Xu et al. (2026). We retained its expert-assigned ultrasound appearance scores, ranging from 0.0 to 3.5 in increments of 0.1. The published annotation protocol refines four clinical grades into this score scale and describes standardized instruction, centralized review, and expert review of grading disagreements. The present study cleaned and organized the existing images and annotations without regrading them. Its cohort and splits differ from those of the earlier study; the SFibAI baseline was trained and evaluated on the present splits.

Acquisition views and lesion-location annotations. Six-class view labels were recovered from the plane identifiers recorded during image acquisition and organized during data cleaning. Lesion-location supervision likewise uses the existing clinician-drawn rectangular boxes described by Xu et al. (2026), after cleaning and organization. These boxes mark image regions requiring attention and carry the original grading annotations. View labels therefore describe the recorded acquisition plane, whereas boxes identify local regions within the image. Neither annotation type is required at inference.

Coverage and weak-target scope. The curated records contain view and lesion-location annotations for every image, although not every record supplies an eligible nonempty weak target. The box objective uses valid nonempty boxes with a positive target score, as specified in Appendix A. Weak-localization evaluation includes 4,035 eligible test images; view evaluation includes all 4,107 test images. Local boxes do not delineate complete lesion contours, so responses beyond them require separate assessment of spatial relevance.

Table S1: Composition of the curated cohort used in this study. Centers overlap between splits; patients do not.

	Total	Train	Validation	Test
Images	108,709	83,722	20,880	4,107
Patients	6,373	4,906	1,227	240
Centers	35	35	33	4

A.2 SFIBAI BASELINE AND TRAINING IMPLEMENTATION

Identifying the SFibAI baseline. The method source is Xu et al. (2026). The baseline follows the authors' public implementation. The baseline uses ResNet-50, global average pooling, a linear $2048 \to 36$ head, softmax, and an expected score on 0.0–3.5. View and localization prediction heads are absent. Upstream permits an explicit initialization checkpoint and otherwise constructs a randomly initialized backbone; our common ImageNet initialization is an adaptation for this comparison, not an upstream default.

Grading objectives. VCRE-Fib and SFibAI use the same fibrosis-score scale and hybrid-loss weights $(\alpha, \beta, \gamma) = (1, 0.02, 0.02)$. For target bin $j \in \{0, \ldots, 35\}$ and predicted probabilities p_k, both convert bin indices to real-valued clinical scores before computing MSE and the clinical-grade

boundary penalty:

$$y = 0.1j, \qquad \hat{y} = 0.1 \sum_{k=0}^{35} k p_k,$$

$$L_{\text{grade}} = L_{\text{KL,local}} + 0.02(\hat{y} - y)^2 + 0.02|\hat{y} - y|\mathbf{1}(\hat{c} \neq c).$$

Here c and $\hat{c}$ are obtained from y and $\hat{y}$ using thresholds 0.5, 1.5, and 2.5, with a score exactly at a threshold assigned to the higher grade. The displayed per-image terms are averaged over the batch. Both MSE and the absolute-error boundary term are evaluated on the 0.0–3.5 score scale.

The local-KL term uses a normalized Gaussian soft target over 36 bins with standard deviation one bin and a local window centered on the target bin. Both methods use this hybrid grading-loss formulation with aligned score units and weights.

Training protocol. All four models use the same random seed (2026), ImageNet-1K-V2 weights, batch size 24, bfloat16 precision, AdamW learning rate 10^{-4} and weight decay 10^{-4}, and StepLR factor 0.6 every 15 epochs. They share ROI preprocessing, stretch resize, right-angle rotation/saturation augmentation, patient-disjoint splits, and the grading objective above. This standardizes the public SFibAI release's different crop/augmentation and initialization defaults. Training trajectories and checkpoint comparison use the same 90-epoch cap for SFibAI, both deletion variants, and full VCRE-Fib. Validation selects from epochs 21–90 by R_{final}, then image COR and earlier epoch, replacing the upstream tolerance-accuracy selector. The selected epochs are 65, 80, 39, and 47, respectively. This comparison on the curated cohort is not a reproduction of the original SFibAI cohort experiment.

Additional VCRE-Fib supervision. Six-class view cross-entropy has weight 0.1. The weak-box loss combines inside log-sum-exp pooling (temperature 10) and the outside response fraction with unit inner weights and overall weight 0.1. Only valid nonempty boxes with target bin $j > 0$ participate; exact score 0.0 is excluded, not the entire F0 grade. Weak targets are transformed in the ROI frame. The support map is learned from grading. Auxiliary losses update the shared encoder, while q and a are detached only at their grading interfaces. The w/o View variant removes the view head, six adapters, posterior mixture, and view CE while retaining unconditioned regional evidence and weak localization. The w/o Weak Loc. variant removes the localization head and box loss, using neutral $a_u = 1$ with the view and support paths retained. Each deletion variant starts independently from the prescribed initialization.

B Complete test and resource reporting

B.1 Test set primary results

The frozen, validation-selected checkpoints are evaluated on 4,107 images, 240 patients, and four centers. The order is VCRE-Fib, w/o Weak Loc., w/o View, and SFibAI (Table S2), with lower R_{final} indicating better performance. Main Tables 2–4 report the primary, image, and spatial outcomes; Appendix B.2 gives the detailed patient-level metrics.

Table S2: Observed test differences relative to SFibAI. Differences are method minus baseline; relative differences divide by the baseline, using unrounded values. Bold denotes the best unrounded value per metric; exact ties are all bold.

Method	Rank	$R_{\text{final}} \downarrow$	Δ absolute $\downarrow$	Δ relative (%) $\downarrow$
VCRE-Fib	1	**0.186865**	**-0.014315**	**-7.115**
w/o Weak Loc.	2	0.191308	-0.009872	-4.907
w/o View	3	0.193799	-0.007380	-3.668
SFibAI	4	0.201180	0.000000	0.000

Metric definitions. Let $e = |\hat{y} - y|$ and let d be the absolute difference in four-level grades with thresholds 0.5, 1.5, and 2.5. The locked per-item risk is

$$\text{COR} = 0.35e/3.5 + 0.15\mathbf{1}(e > 0.3) + 0.15\mathbf{1}(e > 0.5) + 0.25d/3 + 0.10\mathbf{1}(d \geq 2).$$

Patient-max and patient-median aggregate reference and predicted scores separately. Center-balanced COR uses square-root patient-count weights. TMAE is $\text{mean}\max(e - 0.5, 0)$ and severe-error rate is $\text{mean}\,\mathbf{1}(e > 1.0)$, distinct from the two-grade error term inside COR.

B.2 Detailed patient-level grading

Under patient-max and patient-median aggregation, the full model retains the lowest COR (0.200528 and 0.154344), alongside the lowest median MAE (0.339273; Table S3) and center-balanced COR (0.177453; main-text Table 2). Some thresholded accuracies still favor variants: w/o Weak Loc. has higher patient-max ± 0.5 accuracy (67.50%), and w/o View has higher median four-grade accuracy (72.08%). The full model's advantage thus spans overall risk and multiple grading measures.

Table S3: Patient-level test metrics, 240 patients per row. Reference and prediction are aggregated separately; accuracies and severe errors are percentages. Bold compares models within each aggregation.

Method	Level	MAE ↓	Acc. ±.3 ↑	Acc. ±.5 ↑	4-grade ↑	TMAE ↓	Severe ↓	COR ↓
SFibAI	Max	0.472421	50.42	63.75	61.25	0.182625	16.25	0.219742
w/o View	Max	0.445159	**53.75**	66.25	63.75	0.171660	15.42	0.210002
w/o Weak Loc.	Max	0.427688	53.33	**67.50**	63.75	0.149615	12.50	0.202422
VCRE-Fib	Max	**0.420558**	**53.75**	65.42	**65.42**	**0.143967**	**11.67**	**0.200528**
SFibAI	Median	0.348790	62.08	72.08	67.08	0.102808	7.92	0.162587
w/o View	Median	0.346292	61.25	73.75	**72.08**	0.108306	7.92	0.158449
w/o Weak Loc.	Median	0.361912	59.58	72.92	69.58	0.113333	9.17	0.165844
VCRE-Fib	Median	**0.339273**	**62.92**	**74.58**	71.67	**0.102352**	**7.50**	**0.154344**

B.3 Additional grading metrics and center-level results

Per-center estimates are descriptive; one center has two patients. Test/val exports retain per-grade support/recall and all COR terms. Four-grade probabilities sum the 36 bins; image AUPRC uses average precision and ECE uses ten bins. Class-wise AUROC/AUPRC are reported as N/A when either positive or negative examples are absent. Patient probabilities are not aggregated.

Table S4: Additional test grading and image-only probability metrics. Bold denotes the best unrounded value per metric; exact ties are all bold.

Metric	SFibAI	w/o View	w/o Weak Loc.	VCRE-Fib
Patient Max RMSE ↓	0.665765	0.642874	0.611087	**0.596914**
Patient Max macro-F1 ↑	0.606287	0.635779	0.636286	**0.655596**
Patient Median RMSE ↓	0.508183	0.519102	0.528079	**0.505147**
Patient Median macro-F1 ↑	0.675454	**0.726897**	0.702926	0.721044
Image macro-AUROC ↑	0.879791	0.886155	0.879378	**0.891698**
Image macro-AUPRC ↑	0.699348	0.719365	0.699530	**0.729264**
Image multiclass Brier ↓	0.526870	0.499022	0.515159	**0.474036**
Image ECE ↓	0.206199	0.181620	0.177804	**0.164196**

Table S5: Per-center test patient-max COR (lower is better). Bold compares methods within a center; small-center estimates remain descriptive.

Center	Images	Patients	SFibAI	w/o View	w/o Weak Loc.	VCRE-Fib
center_05	25	2	0.005361	0.006614	0.005765	**0.005329**
center_07	1948	115	0.201563	0.192100	0.192990	**0.183879**
center_15	1599	93	0.308194	0.295332	**0.274507**	0.281672
center_17	535	30	0.029521	0.027661	0.028223	**0.025816**

B.4 Resource measurements

Table S6: Matched resource measurements; latency is milliseconds per batch and memory is peak allocated MiB. All quantities are lower-is-better; bold compares models within each batch size.

Method	Batch	Parameters	GMACs/image	Median ms	P95 ms	Peak MiB
SFibAI	1	**23581796**	**21.353**	**5.620**	**5.626**	**154.318**
w/o View	1	25311562	22.103	6.171	6.188	160.919
w/o Weak Loc.	1	24832975	21.654	7.122	7.165	159.093
VCRE-Fib	1	25422928	22.258	7.371	7.397	161.344
SFibAI	8	**23581796**	**21.353**	**30.088**	**30.119**	**539.318**
w/o View	8	25311562	22.103	31.067	31.133	545.919
w/o Weak Loc.	8	24832975	21.654	32.338	32.442	544.093
VCRE-Fib	8	25422928	22.258	32.800	32.979	546.344

All four frozen models were measured on the same NVIDIA A100-SXM4-80GB with PyTorch 2.11.0+cu128, FP32, TF32 disabled, and 512×512 inputs. Fifty warm-up iterations preceded 200 synchronized wall-time measurements. Latency is per batch, excludes loading/transfer, and covers the model forward with inputs already on the GPU. Peak allocated memory is reported in MiB. Counted Conv2d/Linear operations exclude elementwise, pooling, and normalization; FLOPs are twice the reported MACs.

C Spatial predictions and qualitative examples

C.1 View names and weak-target summaries

Table S7: Six acquisition views and test recall (percent). Bold compares the two view-capable models within each class. Main-paper cases cover four true views.

View	Anatomical/acquisition view	N	w/o Weak Loc.	VCRE-Fib
1	Subxiphoid sagittal plane (left lobe of the liver)	694	**93.66**	84.73
2	Left subcostal transverse plane (left lobe and sagittal segment of portal vein)	699	**74.82**	65.09
3	Right subcostal plane (second porta hepatis and confluence of three hepatic veins)	689	**59.51**	54.57
4	Right subcostal plane (right hepatic lobe, right hepatic vein, diaphragm)	671	**58.72**	52.61
5	Right intercostal oblique plane (right hepatic lobe, portal vein, gallbladder)	682	**81.52**	66.86
6	Right intercostal oblique plane (right hepatic lobe and right kidney)	672	**85.71**	76.79

Additional weak-target summaries. On test, full VCRE-Fib has mean inside attention 0.570815, outside response fraction 0.666595, and inside/outside mean-response ratio 359,012.493654; w/o

View has 0.499237, 0.550558, and 1,678,411.504391, respectively. The outside response fraction is the summed attention outside the target boxes divided by total attention. The last ratio clips its denominator at 10^{-8}, so full-field boxes can yield very large values. These are descriptive response summaries, not localization-accuracy measures. Table S7 compares class recall for the two view-capable models; evaluation artifacts also provide prediction frequencies and confusion matrices.

C.2 Case selection and display protocol

View metrics include accuracy/macro-F1; weak-target Dice/IoU use threshold 0.5. From pools of 12 test and four validation cases, four full-model test cases spanning four true views were selected for the main figure and two validation cases for the appendix. Correspondence candidates were prioritized by weak-box Dice within view. Expansion candidates required an outside response fraction of 0.5–0.9 (summed attention outside the target boxes divided by total attention), Dice above 0.1, a thresholded response intersecting the boxes, and at least 10% of its area outside all displayed boxes. Patients were deduplicated within each pool. All panels use the same processed ROI, fixed $[0, 1]$ localization scale, opacity 0.55, and cyan 0.5 contour. These are selected illustrations; clinical descriptions of the extended regions are not assigned.

The main paper presents four selected test cases. Figure S1 shows two selected validation examples with responses extending beyond local boxes. These examples are separate from the test ranking and spatial metrics; the first also shows a disagreement between the reference and predicted views.

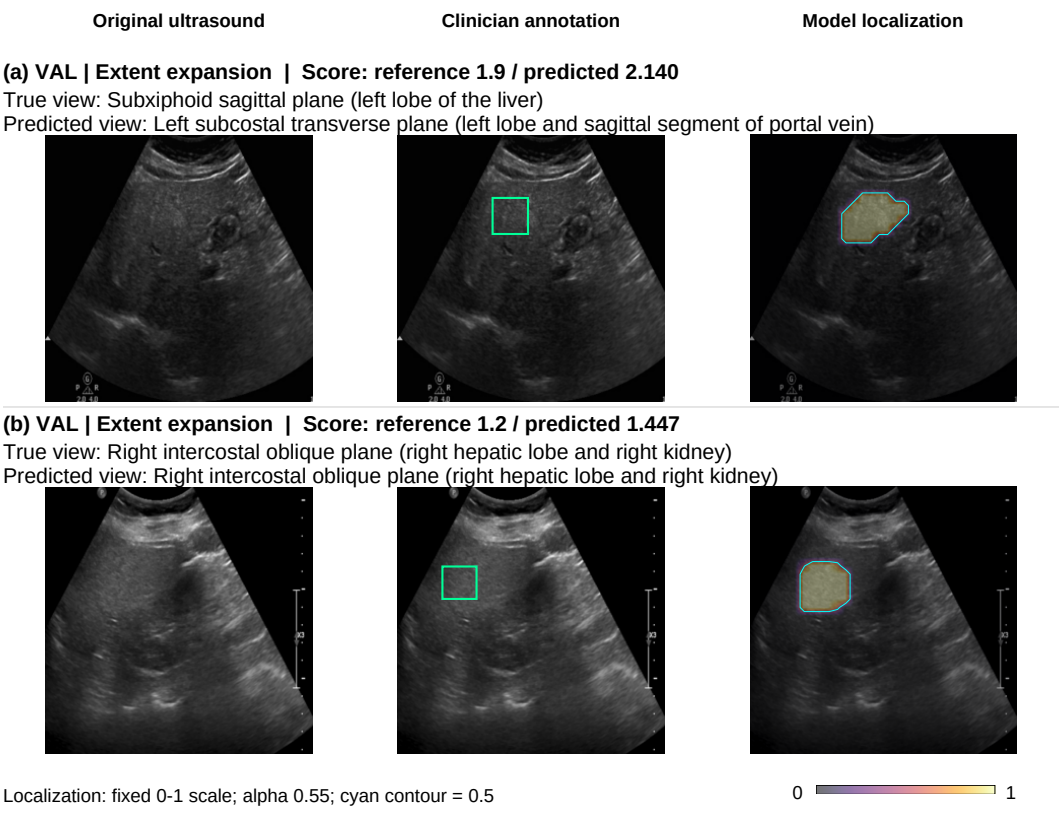

Figure S1: Selected validation examples with responses beyond local boxes. Predictions use the frozen checkpoint evaluated quantitatively. The three columns show the original ROI, clinician annotation, and model localization; labels report reference and predicted scores and acquisition views. In the first example, the predicted view differs from the reference. Responses beyond local boxes identify locations for inspection, without establishing complete lesion extent.

D VALIDATION-SELECTED CHECKPOINTS AND TRAINING DIAGNOSTICS

For SFibAI, w/o View, w/o Weak Loc., and full VCRE-Fib, validation selects epochs 65, 80, 39, and 47, with R_{final} of 0.172661, 0.157844, 0.159315, and 0.156067. Validation contains 20,880 images, 1,227 patients, and 33 centers. Appendix Table S8 records the four frozen checkpoints; Figure S2 shows their trajectories under the common 90-epoch cap. Test risk determines ranking; validation documents selection. The four models use four GPUs with SyncBatchNorm and global batch 24

Table S8: Validation-selected frozen checkpoints. Validation: 20,880 images, 1,227 patients, 33 centers. All four methods, including VCRE-Fib, use a 90-epoch cap for the training comparison; checkpoints are selected from epochs 21–90. Bold denotes the best unrounded value per metric; exact ties are all bold.

Method	Epoch cap	Selected epoch	Val R_{final} ↓	Val image COR ↓
SFibAI	90	65	0.172661	0.142376
w/o View	90	80	0.157844	0.134258
w/o Weak Loc.	90	39	0.159315	**0.134234**
VCRE-Fib	90	47	**0.156067**	0.134735

under the common training settings in Appendix A. Validation selects from epochs 21–90 by R_{final}, image COR, then earlier epoch; selected checkpoints are frozen before test evaluation. Figure S2 shows all 90 recorded epochs without smoothing. Total losses contain different auxiliary terms and serve as within-model diagnostics. Figure S3 shows the full model's loss components.

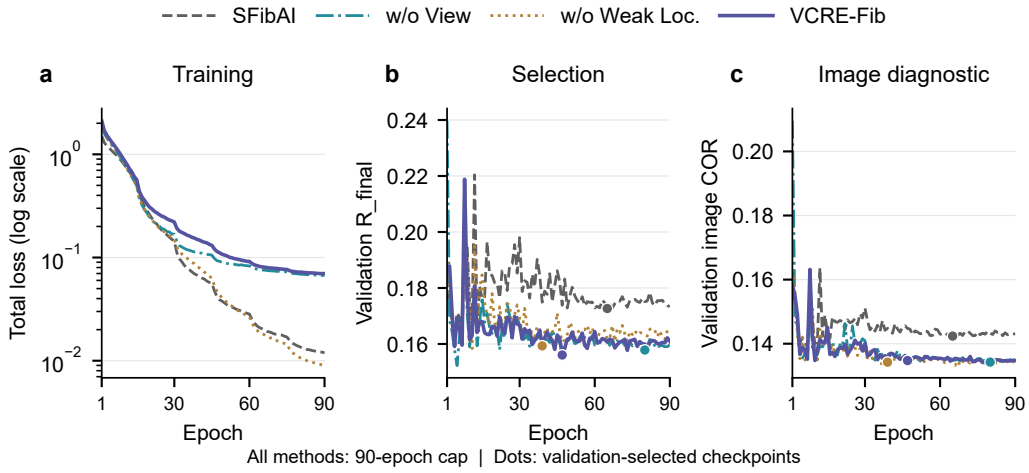

Figure S2: Training and validation trajectories under a common 90-epoch cap. (a) Total training loss on a logarithmic axis, including each model's auxiliary terms. (b) Validation R_{final}. (c) Validation image COR. Curves show unsmoothed epoch-level observations; dots mark the validation-selected checkpoints in Table S8.

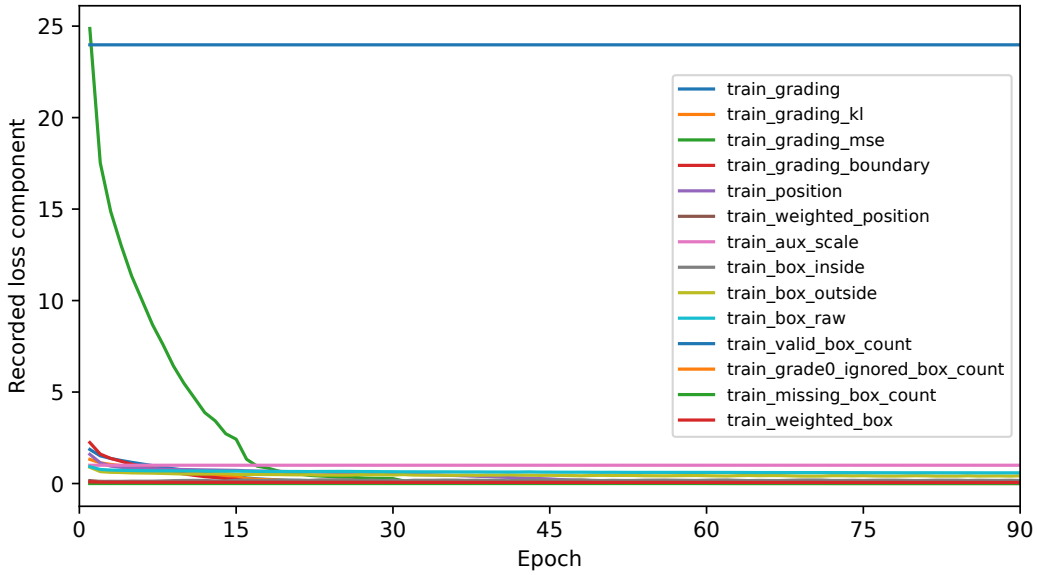

Figure S3: Full VCRE-Fib loss components under the same 90-epoch cap, shown as training diagnostics rather than estimates of module effects.