MULTIMODAL EVIDENCE REASONING FOR INTERPRETABLE BREAST ULTRASOUND VIDEO DIAGNOSIS

Anonymous authors
Paper under double-blind review

ABSTRACT

Deep learning has shown promising performance in breast lesion diagnosis, yet most existing methods are developed based on representative images manually selected by radiologists, limiting their ability to automatically perceive and identify diagnostic information from continuous ultrasound videos acquired during clinical scanning. Current methods are primarily optimized for final diagnostic predictions, providing limited evidence for understanding how their decisions are reached. Furthermore, achieving clinical interpretability in diagnostic models often depends on semantic guidance from radiologist-authored diagnostic reports paired with ultrasound videos, yet such reports are generally unavailable before diagnostic inference. To address these challenges, we propose a multimodal evidence reasoning framework for interpretable breast ultrasound video diagnosis. In clinical scenarios where diagnostic reports are unavailable prior to inference, we show through Fisher information analysis that aligning unpaired clinical text and ultrasound videos within a shared representation space can strengthen visual evidence representation of concepts. To reason over raw ultrasound videos, we further develop an evidence driven fusion strategy equipped with a dual-aware module to identify diagnostically informative frames, aggregate reliable evidence across time, and derive interpretable textual evidence. The resulting visual and textual evidence are subsequently integrated into multimodal explainable diagnostic chains, providing traceable support for the final diagnostic prediction. Extensive experiments on the large-scale BUS-VD breast ultrasound video dataset validate the effectiveness of the proposed framework in both diagnostic performance and interpretability. The source code is provided in the supplementary material.

1 INTRODUCTION

Breast cancer remains the most commonly diagnosed cancer and a leading cause of cancer-related mortality among women worldwide (Bray et al., 2024). Early identification and accurate assessment of breast lesions are essential for timely clinical intervention and improved patient outcomes (Harbeck et al., 2019). Owing to its absence of ionizing radiation and capability for multiplanar imaging, ultrasound is widely used in the clinical diagnosis of breast lesions (Hooley et al., 2013). In clinical practice, radiologists assess breast lesions according to the Breast Imaging Reporting and Data System (BI-RADS)(Mendelson et al., 2013), categorizing them as grade 2–5, which indicate benign, probably benign, suspicious, and highly suggestive of malignancy, respectively. Most breast ultrasound diagnostic models (Shareef et al., 2023; Mo et al., 2023) rely on representative images preselected by radiologists, which limits their direct application in real clinical settings and overlooks diagnostic information from multiple views. In actual clinical diagnosis, ultrasound scans naturally generate videos that capture the morphology of lesions from multiple probe angles and imaging planes, thereby providing more comprehensive diagnostic information than images. Although breast ultrasound videos can capture diagnostic information more comprehensively, there are still challenges in achieving automated diagnoses that align with the clinical reasoning of radiologists.

Compared to images selected by radiologists, raw ultrasound videos often contain frames with reduced image quality or limited diagnostic value due to probe motion and imaging artifacts. Moreover, the same lesion may exhibit different characteristics across different imaging planes, which

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.

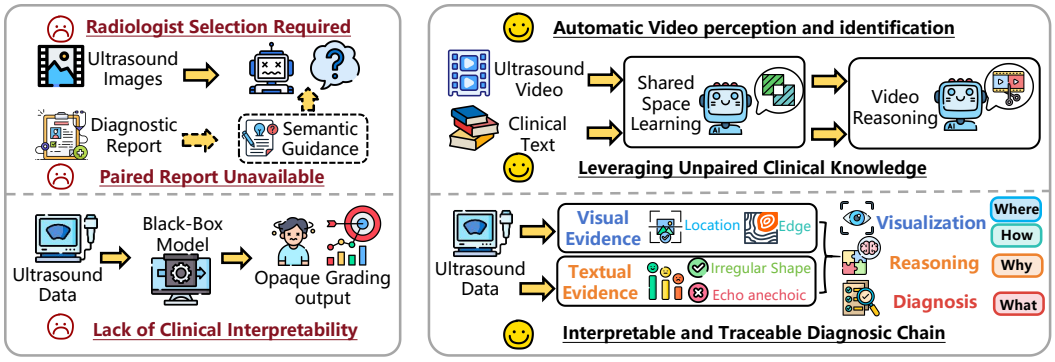


Figure 1: Comparison between existing breast ultrasound diagnostic models and our proposed **MIChain**.

can interfere with diagnostic reasoning based on the video. Therefore, ultrasound video models must not only identify frames with diagnostic value but also integrate evidence of the lesion across frames.

Meanwhile, although recent deep learning methods have demonstrated great potential in the diagnosis of breast lesions(Shareef et al., 2023; Mo et al., 2023; Sun et al., 2023; Jiao et al., 2024; Lin et al., 2025), most methods still rely on black-box models to generate the final prediction. Existing medically interpretable models(Zhang et al., 2017; Wu et al., 2023; Zhang et al., 2023) typically rely on paired clinical reports to provide semantic guidance for learning clinically interpretable features. However, such radiologist-authored reports are typically generated after the diagnostic assessment and are therefore unavailable in advance. To overcome the reliance of interpretable models on unavailable clinical reports, clinical concept knowledge derived from the BI-RADS grading criteria should be introduced, while addressing their unpaired relationship with ultrasound videos.

Furthermore, existing models fail to provide interpretable multimodal evidence to support their diagnostic conclusions. In contrast, radiologists typically select representative images and mark the lesion region. They then characterize the lesion using the core BI-RADS diagnostic dimensions, including shape, orientation, margin, echo pattern, and posterior acoustic features, and integrate these findings to determine the BI-RADS grade and formulate a diagnostic report. Therefore, a comprehensive breast ultrasound diagnosis should include not only a grading prediction but also multimodal results consisting of traceable visual evidence and interpretable textual evidence. The gap between existing models and clinical diagnostic practice limits their clinical reliability.

To address these challenges, we propose **MIChain**, a **M**ultimodal **I**nterpretability **Chain** for breast ultrasound video diagnosis. **MIChain** comprises three key components: **1)** we construct BI-RADS concept text with explicit clinical semantics and project the unpaired concept text and ultrasound videos into a cross-modal shared space, thus enabling the learning of interpretable concept representations without relying on radiologist-authored diagnostic reports. **2)** To reduce the adverse effects of frames with degraded quality or ambiguous diagnostic cues on video reasoning, we propose an evidence-guided video reasoning approach that jointly assesses the sufficiency and directional certainty of frame-wise concept evidence and adaptively aggregates reliable evidence across the video. **3)** The model constructs a multimodal explainable diagnostic chain by integrating visual evidence and textual evidence to support the final diagnosis, bringing the diagnostic process closer to radiologists' clinical reasoning. Our contributions are summarized as follows:

- We propose a cross-modal shared concept representation learning method that projects unpaired clinical concept descriptions and ultrasound videos into a shared semantic space. From a Fisher information perspective, we further demonstrate that unpaired clinical text enhances visual concept representations in the shared space.
- An evidence-driven video reasoning module is developed using dual-aware evidence learning and video evidence aggregation to identify diagnostically informative frames and integrate concept evidence across frames.

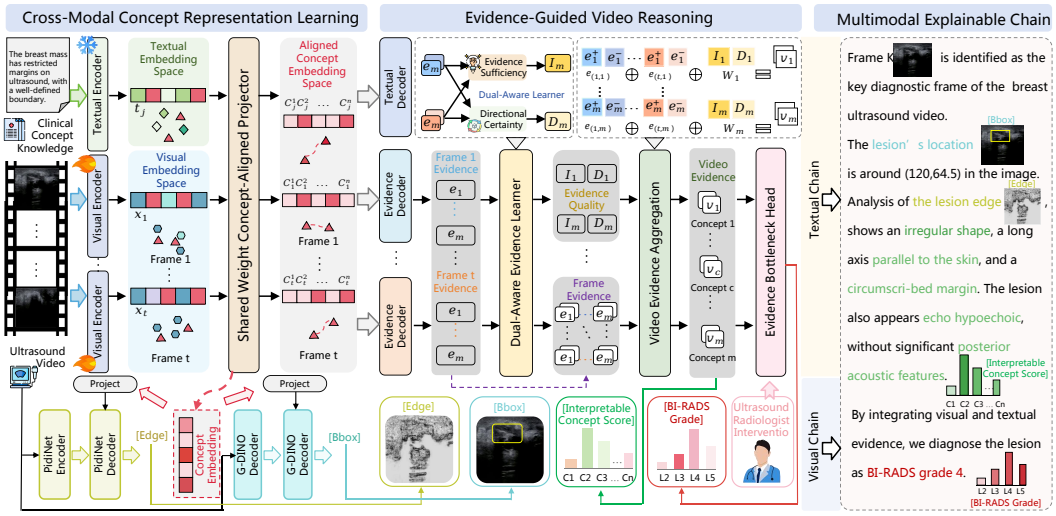


Figure 2: Overview of the proposed **MIChain** framework. **MIChain** first projects unpaired BI-RADS concept descriptions and ultrasound videos into a shared semantic space for interpretable concept representation learning. It then performs evidence-driven video reasoning to identify diagnostically informative frames and aggregate concept evidence across frames. Finally, visual and textual evidence are integrated to construct a multimodal explainable diagnostic chain for BI-RADS grading.

- A multimodal explainable diagnostic chain is constructed by integrating visual and textual evidence, enabling clinically interpretable and verifiable diagnosis.

2 METHOD

In this section, **MIChain** follows a three-stage reasoning pipeline. It begins with cross-modal shared concept representation learning, which introduces unpaired clinical knowledge text to learn concept representations and thereby derive the corresponding visual evidence from ultrasound videos. Building on these, evidence-guided video reasoning extracts textual evidence from the concept representations through dual-aware evidence learning and video evidence aggregation. Finally, the resulting visual and textual evidence are integrated to construct a multimodal explainable diagnostic chain. The overall framework is illustrated in Fig. 2.

2.1 CROSS-MODAL SHARED CONCEPT REPRESENTATION LEARNING

To address the unavailability of paired radiologist-authored diagnostic report, we introduce unpaired clinical concept knowledge as semantic guidance. Specifically, we construct the clinical concept knowledge comprising GPT5 (Singh et al., 2025) generated descriptions of commonly used BI-RADS concepts, which are verified by radiologists, with further details provided in Appendix D.

Given an ultrasound video $\mathcal{X} = \{\mathbf{x}_t\}_{t=1}^{T}$ and the clinical knowledge text $\mathcal{T} = \{\mathbf{t}_j\}_{j=1}^{n}$, where $\mathbf{x}_t$ and $\mathbf{t}_j$ denote the t-th ultrasound frame and the clinical description of the j-th concept, respectively, we employ a DINOv2 visual encoder (Oquab et al., 2024) and a frozen BERT text encoder (Devlin et al., 2019) to extract the corresponding visual embeddings $\mathbf{h}_t^{vis} \in \mathbb{R}^d$ and textual embeddings $\mathbf{h}_j^{txt} \in \mathbb{R}^d$:

$$\mathbf{h}_t^{vis} = f_{vis}(\mathbf{x}_t) = \text{DINOv2}(\mathbf{x}_t), \qquad \mathbf{h}_j^{txt} = f_{txt}(\mathbf{t}_j) = \text{BERT}(\mathbf{t}_j). \tag{1}$$

To enhance visual concept representations with semantic guidance from clinical concept knowledge without requiring paired video–text samples, we introduce a shared-weight concept-aligned projector that maps visual and textual embeddings into a common concept embedding space: $\mathbf{z}_t^{vis} = \mathbf{W}\mathbf{h}_t^{vis} + \mathbf{b}$ and $\mathbf{z}_j^{txt} = \mathbf{W}\mathbf{h}_j^{txt} + \mathbf{b}$, where $\mathbf{W}$ and $\mathbf{b}$ denote the learnable weight matrix and bias shared across the visual and textual modalities. Here, $\mathbf{z}_t^{vis} = \{\mathbf{C}_t^i\}_{i=1}^{n}$ and $\mathbf{z}_j^{txt} = \{\mathbf{C}_j^i\}_{i=1}^{n}$

denote the visual and textual concept representations, respectively, where $\mathbf{C}_t^i$ and $\mathbf{C}_j^i$ are the corresponding visual and textual embeddings of the i-th concept.

To further provide traceable visual evidence, the aligned concept embeddings are projected into the feature dimensions required by the respective decoders and then fed into the decoders of Grounding DINO (Liu et al., 2024) and PiDiNet (Su et al., 2021). These decoder branches generate lesion localization and boundary cues, thereby providing spatially traceable visual evidence. Further implementation details are provided in the Appendix F.

2.2 Why Can Unpaired Clinical Text Enhance Visual Concept Representations?

The cross-modal shared concept representation introduced above uses unpaired clinical concept knowledge to provide semantic guidance for ultrasound videos through a shared concept-aligned projector. However, since the clinical descriptions are constructed independently of individual ultrasound videos, the textual knowledge remains unpaired with the visual observations. This motivates us to theoretically examine why such unpaired textual knowledge can still benefit visual concept representation learning.

> ***Question 1:*** *How can unpaired clinical text provide additional information for visual concept representation learning?*

To answer this question, we analyze the visual and textual embeddings in equation 1 using a linear-Gaussian shared-factor model (Klami et al., 2013). The key idea is to separate the clinical concept information shared across modalities from modality-specific variations. Let $\boldsymbol{\theta} = [\boldsymbol{\theta}_c, \boldsymbol{\theta}_v, \boldsymbol{\theta}_t]^\top$, where $\boldsymbol{\theta}_c$ denotes the shared clinical concept factors, while $\boldsymbol{\theta}_v$ and $\boldsymbol{\theta}_t$ capture visual- and text-specific factors, respectively. We model

$$\mathbf{h}_i^{\text{vis}} = \mathbf{A}_{c,i}\boldsymbol{\theta}_c + \mathbf{A}_{v,i}\boldsymbol{\theta}_v + \boldsymbol{\epsilon}_{v,i}, \qquad \mathbf{h}_j^{\text{txt}} = \mathbf{B}_{c,j}\boldsymbol{\theta}_c + \mathbf{B}_{t,j}\boldsymbol{\theta}_t + \boldsymbol{\epsilon}_{t,j}, \tag{2}$$

where $\boldsymbol{\epsilon}_{v,i}$ and $\boldsymbol{\epsilon}_{t,j}$ denote independent Gaussian noise.

Although the visual and textual observations are unpaired, they both depend on the same shared concept factors $\boldsymbol{\theta}_c$. Consequently, each modality provides an independent constraint on the same latent concept variables. Under conditional independence, their joint likelihood factorizes as

$$p(\mathcal{H}_v, \mathcal{H}_t \mid \boldsymbol{\theta}) = \prod_{i=1}^{N_v} p(\mathbf{h}_i^{\text{vis}} \mid \boldsymbol{\theta}_c, \boldsymbol{\theta}_v) \prod_{j=1}^{N_t} p(\mathbf{h}_j^{\text{txt}} \mid \boldsymbol{\theta}_c, \boldsymbol{\theta}_t), \tag{3}$$

where $\mathcal{H}_v$ and $\mathcal{H}_t$ denote the visual and textual observation sets. Thus, although the two modalities are unpaired, they can still jointly constrain the shared concept factors through $\boldsymbol{\theta}_c$.

However, this joint formulation only establishes that the textual observations can participate in estimating the shared concepts, it does not yet show whether they provide additional information beyond the visual observations. We therefore quantify this contribution using Fisher information.

Let $\mathcal{I}_v(\boldsymbol{\theta})$ and $\mathcal{I}_t(\boldsymbol{\theta})$ denote the Fisher information matrices induced by the visual and textual observations, respectively. Since our analysis focuses on the shared clinical concept factors $\boldsymbol{\theta}_c$, we consider the corresponding shared-factor Fisher information blocks. Under the conditional independence of the visual and textual observations given the latent factors, their information contributions are additive:

$$\begin{aligned} \mathcal{J}_v &:= (\mathcal{I}_v(\boldsymbol{\theta}))_{\theta_c,\theta_c} = \sum_{i=1}^{N_v} \mathbf{A}_{c,i}^\top \mathbf{A}_{c,i}, \qquad \mathcal{J}_t := (\mathcal{I}_t(\boldsymbol{\theta}))_{\theta_c,\theta_c} = \sum_{j=1}^{N_t} \mathbf{B}_{c,j}^\top \mathbf{B}_{c,j}, \\ \mathcal{J}_{v,t} &:= (\mathcal{I}_v(\boldsymbol{\theta}) + \mathcal{I}_t(\boldsymbol{\theta}))_{\theta_c,\theta_c} = \mathcal{J}_v + \mathcal{J}_t = \sum_{i=1}^{N_v} \mathbf{A}_{c,i}^\top \mathbf{A}_{c,i} + \sum_{j=1}^{N_t} \mathbf{B}_{c,j}^\top \mathbf{B}_{c,j}. \end{aligned} \tag{4}$$

Here, $\mathcal{J}_v$ and $\mathcal{J}_t$ characterize the Fisher information contributed by the visual and textual observations, respectively, for estimating the shared concept factors, while $\mathcal{J}_{v,t}$ denotes their joint contribution. The corresponding joint least-squares formulation and the complete block-wise Fisher information derivations are provided in Appendix E.2.

Theorem 1 (Variance Reduction from Unpaired Clinical Text). *Let $\widehat{\boldsymbol{\theta}}_c^{v}$ and $\widehat{\boldsymbol{\theta}}_c^{v,t}$ denote the least-squares estimators of the shared concept factors obtained from visual observations alone and jointly from visual observations and unpaired clinical text, respectively. Under the standard regularity and identifiability conditions, if at least one textual design matrix $\mathbf{B}_{c,j}$ has full column rank, then $\mathcal{J}_t \succ \mathbf{0}$, and*

$$\mathcal{J}_{v,t} = \left(\mathcal{I}_v(\boldsymbol{\theta}) + \mathcal{I}_t(\boldsymbol{\theta})\right)_{\theta_c,\theta_c} \succ \left(\mathcal{I}_v(\boldsymbol{\theta})\right)_{\theta_c,\theta_c} = \mathcal{J}_v, \tag{5}$$

Consequently, the covariance of the shared concept factor estimator satisfies the strict Loewner ordering

$$\mathrm{Var}\left(\widehat{\boldsymbol{\theta}}_c^{v,t}\right) \prec \mathrm{Var}\left(\widehat{\boldsymbol{\theta}}_c^{v}\right). \tag{6}$$

The complete proof is provided in Appendix E.4, where Lemma 1 establishes the inverse-order relation used to derive the reduction in estimator covariance. Theorem 1 therefore establishes a global uncertainty reduction when the unpaired clinical text provides non-degenerate information throughout the shared concept space. In practice, however, clinical descriptions may characterize only a subset of diagnostic concepts, motivating us to examine whether such partial textual information can still be beneficial.

> ***Question 2:*** *Can unpaired clinical text still provide additional information when it covers only part of the shared concept space?*

The global condition in Theorem 1, $\mathcal{J}_t \succ \mathbf{0}$, requires $\mathbf{r}^\top \mathcal{J}_t \mathbf{r} > 0$, for every $\mathbf{r} \neq \mathbf{0}$ meaning that the textual observations provide information along every direction of the shared concept space. We now relax this global condition and consider whether textual information along only a particular concept direction can still be beneficial.

Theorem 2 (Directional Information Gain from Unpaired Clinical Text). *Let $\mathbf{r} \in \mathbb{R}^{d_c}$ be a nonzero direction in the shared concept space. If there exists at least one textual observation $j \in \{1, \ldots, N_t\}$ such that $\mathbf{B}_{c,j}\mathbf{r} \neq \mathbf{0}$, equivalently, $\mathbf{r}^\top \mathcal{J}_t \mathbf{r} > 0$ then*

$$\mathbf{r}^\top \mathcal{J}_{v,t} \mathbf{r} = \mathbf{r}^\top \left[\mathcal{I}_v(\boldsymbol{\theta}) + \mathcal{I}_t(\boldsymbol{\theta})\right]_{\theta_c,\theta_c} \mathbf{r} > \mathbf{r}^\top \left[\mathcal{I}_v(\boldsymbol{\theta})\right]_{\theta_c,\theta_c} \mathbf{r} = \mathbf{r}^\top \mathcal{J}_v \mathbf{r}. \tag{7}$$

Furthermore, if $\mathbf{r} \notin \mathrm{range}(\mathcal{J}_v)$ and becomes identifiable under the joint observations, its estimation uncertainty is reduced from unbounded under visual-only observations to finite under the joint observations.

The complete proof of theorem 2 is provided in Appendix E.5. The Appendix further considers directions that are already identifiable from visual observations and uses Lemma 2 to characterize the conditions under which their estimation uncertainty is reduced.

Together, Theorems 1 and 2 show that unpaired clinical text can provide complementary information for shared concept estimation at both global and directional levels, without requiring paired video–text observations.

2.3 Evidence-Guided Video Reasoning

To reduce the adverse effects of frames with degraded image quality or ambiguous diagnostic cues on video reasoning, we propose an evidence-guided video reasoning approach that evaluates the reliability of frame concept evidence and adaptively aggregates reliable evidence across the video.

Dual-Aware Evidence Learning For each frame t and concept c, the concept embedding $\mathbf{C}_t^c$ is mapped by a concept-specific evidence decoder $\mathcal{D}_c^e$ to positive and negative evidence. To assess the reliability of concept evidence, we consider two complementary aspects: whether sufficient evidence is available and whether the evidence provides a clear direction toward one of the two concept states.

Following evidential deep learning based on subjective logic (Sensoy et al., 2018), we set $\alpha_{t,c}^{+} = e_{t,c}^{+} + 1$ and $\alpha_{t,c}^{-} = e_{t,c}^{-} + 1$ as the corresponding Beta parameters. The evidence vector $\mathbf{e}_{t,c}$, total

evidence strength $\Phi_{t,c}$, and evidential uncertainty $U_{t,c}$ are given by

$$\begin{cases} \mathbf{e}_{t,c} = \text{Softplus}\left(\mathcal{D}_c^e(\mathbf{C}_t^c)\right) = \log\left(1+\exp\left(\mathcal{D}_c^e(\mathbf{C}_t^c)\right)\right) = \left[e_{t,c}^+, e_{t,c}^-\right], \\ \Phi_{t,c} = \alpha_{t,c}^+ + \alpha_{t,c}^- = e_{t,c}^+ + e_{t,c}^- + 2, \\ U_{t,c} = \dfrac{2}{\Phi_{t,c}} = \dfrac{2}{e_{t,c}^+ + e_{t,c}^- + 2}. \end{cases} \tag{8}$$

Here, the logarithm and exponential are applied element-wise, and $e_{t,c}^+$ and $e_{t,c}^-$ denote nonnegative evidence supporting the presence and absence of concept c, respectively. A larger $U_{t,c}$ indicates weaker evidence for assessing concept c, and the corresponding evidential sufficiency is defined as $I_{t,c} = 1 - U_{t,c}$. Evidential sufficiency captures whether sufficient evidence is available, but does not characterize whether the positive and negative evidence provide a clear concept-state direction. We therefore further consider directional certainty $D_{t,c}$ and jointly integrate these two complementary aspects into a dual-aware evidence reliability score:

$$R_{t,c} = I_{t,c} D_{t,c} = \underbrace{\frac{e_{t,c}^+ + e_{t,c}^-}{e_{t,c}^+ + e_{t,c}^- + 2}}_{\text{Evidential Sufficiency}} \cdot \underbrace{\frac{\left|e_{t,c}^+ - e_{t,c}^-\right|}{e_{t,c}^+ + e_{t,c}^-}}_{\text{Directional Certainty}} = \frac{\left|e_{t,c}^+ - e_{t,c}^-\right|}{e_{t,c}^+ + e_{t,c}^- + 2}. \tag{9}$$

where $R_{t,c}$ measures the diagnostic reliability of frame t for assessing concept c in the ultrasound video.

Video Evidence Aggregation Based on the dual-aware diagnostic reliability $R_{t,c}$, we further aggregate reliable concept evidence from different frames:

$$w_{t,c} = \frac{\exp(R_{t,c}/\eta)}{\sum_{t'=1}^{T} \exp(R_{t',c}/\eta)}, \qquad \mathbf{v}_c = \sum_{t=1}^{T} w_{t,c} \mathbf{e}_{t,c} = [v_c^+, v_c^-], \tag{10}$$

where $\eta > 0$ controls the concentration of frame evidence in ultrasound video reasoning, with smaller values emphasizing a few highly reliable frames and larger values incorporating evidence from a broader range of frames.

Following the evidential framework introduced above, the aggregated evidence v_c defines the video Dirichlet parameters $\alpha_c^+ = v_c^+ + 1$ and $\alpha_c^- = v_c^- + 1$, with the corresponding strength $S_c = \alpha_c^+ + \alpha_c^-$. We supervise the video evidence using the uncertain cross-entropy loss (Charpentier et al., 2020):

$$\mathcal{L}_{\text{concept}} = \sum_{c \in \mathcal{C}} \left[\psi(S_c) - \psi\left(y_c \alpha_c^+ + (1-y_c)\alpha_c^-\right)\right], \tag{11}$$

where $y_c \in \{0,1\}$ is the ground-truth label of concept c and $\psi(\cdot)$ denotes the digamma function.

Evidence-Guided Grading and Optimization For video grading, the aggregated evidence is converted into the presence probability of concept c as $p_c = \alpha_c^+ / S_c$. Each p_c is associated with its corresponding clinical concept description to provide interpretable textual evidence. The concept probability further gates the positive and negative evidence to form an evidential concept bottleneck:

$$\mathbf{g} = \text{Concat}_{c=1}^{C}\left[p_c v_c^+, (1-p_c) v_c^-\right]. \tag{12}$$

This explicit gating enables concept intervention by replacing p_c with a corrected value $\widetilde{p}_c \in [0,1]$. The resulting bottleneck $\mathbf{g}$ is fed into a linear grading head to predict the grade, supervised by the focal loss $\mathcal{L}_{\text{grade}} = -\sum_{k=1}^{K} y_k^g (1-\widehat{p}_k^g)^{\gamma_g} \log \widehat{p}_k^g$, where γ_g controls the emphasis on difficult samples. Meanwhile, $\mathcal{L}_{\text{text}}$ supervises the textual branch by predicting the grade from each aligned textual concept representation $\mathbf{z}_j^{\text{txt}}$ using its associated label.

The overall loss of **MIChain**, denoted by $\mathcal{L}$, combines the grading loss $\mathcal{L}_{\text{grade}}$, concept loss $\mathcal{L}_{\text{concept}}$, and textual supervision loss $\mathcal{L}_{\text{text}}$, and is formulated as:

$$\mathcal{L} = \underbrace{-\sum_{k=1}^{K} y_k^g (1-\widehat{p}_k^g)^{\gamma_g} \log \widehat{p}_k^g}_{\mathcal{L}_{\text{grade}}} + \lambda_c \underbrace{\sum_{c \in \mathcal{C}} \left[\psi(S_c) - \psi\left(y_c \alpha_c^+ + (1-y_c)\alpha_c^-\right)\right]}_{\mathcal{L}_{\text{concept}}} + \mathcal{L}_{\text{text}}, \tag{13}$$

where γ_g is the focusing parameter of the grading loss and λ_c controls the contribution of concept supervision.

Table 1: Comparison of BI-RADS grading and benign/malignant classification performance (%), together with inference time (ms). Best results are in bold, second-best results are underlined. Colored arrows indicate changes relative to Video Swin.

Type	Method	BI-RADS Grading				Benign/Malignant Classification				Time (ms)↓
		Acc↑	Spe↑	Kappa↑	F1↑	Acc↑	Spe↑	Kappa↑	F1↑	
Non-interpretable	Video Swin (CVPR'22)	84.33	89.63	71.77	76.70	85.71	**96.25**	58.69	79.11	41.21
	UniFormerV2 (ICCV'23)	82.95 ↓1.38	90.73 ↑1.10	68.93 ↓2.84	76.35 ↓0.35	85.25 ↓0.46	92.50 ↓3.75	60.13 ↑1.44	80.03 ↑0.92	102.25 ↑61.04
	VideoMAEv2 (CVPR'23)	84.33 0.00	90.84 ↑1.21	69.97 ↓1.80	70.93 ↓5.77	86.18 ↑0.47	93.13 ↓3.12	62.62 ↑3.93	81.28 ↑2.17	32.07 ↓9.14
	InternVideo2 (ECCV'24)	80.65 ↓3.68	87.68 ↓1.95	63.69 ↓8.08	56.37 ↓20.33	85.25 ↓0.46	**96.25** 0.00	57.08 ↓1.61	78.27 ↓0.84	97.07 ↑55.86
	VideoMamba (ECCV'24)	82.03 ↓2.30	88.10 ↓1.53	64.89 ↓6.88	72.17 ↓4.53	84.79 ↓0.92	<u>95.63</u> ↓0.62	56.02 ↓2.67	77.76 ↓1.35	28.78 ↓12.43
	DejaVid (CVPR'25)	85.25 ↑0.92	91.45 ↑1.82	71.21 ↓0.56	<u>79.24</u> ↑2.54	<u>87.56</u> ↑1.85	95.00 ↓1.25	65.75 ↑7.06	82.81 ↑3.70	96.71 ↑55.50
Interpretable	CBM (ICML'20)	82.02 ↓2.31	90.02 ↑0.39	59.85 ↓11.92	57.59 ↓19.11	85.25 ↓0.46	94.37 ↓1.88	58.66 ↓0.03	79.21 ↑0.10	**10.81** ↓30.40
	CEM (NeurIPS'22)	82.03 ↓2.30	90.40 ↑0.77	71.51 ↓0.26	75.73 ↓0.97	85.25 ↓0.46	91.25 ↓5.00	61.05 ↑2.36	80.52 ↑1.41	16.84 ↓24.37
	ProbCBM (ICML'23)	80.18 ↓4.15	87.09 ↓2.54	61.82 ↓9.95	55.45 ↓21.25	84.79 ↓0.92	**96.25** 0.00	55.46 ↓3.23	77.42 ↓1.69	<u>10.91</u> ↓30.30
	ECBM (ICLR'24)	83.87 ↓0.46	91.33 ↑1.70	71.97 ↑0.20	76.85 ↑0.15	86.64 ↑0.93	94.38 ↓1.87	63.21 ↑4.52	81.54 ↑2.43	11.41 ↓29.80
	evi-CEM (MICCAI'24)	<u>85.25</u> ↑0.92	91.53 ↑1.90	72.13 ↑0.36	77.03 ↑0.33	<u>87.56</u> ↑1.85	90.00 ↓6.25	<u>68.76</u> ↑10.07	<u>84.37</u> ↑5.26	11.13 ↓30.08
	SSCBM (ICCV'25)	83.87 ↓0.46	<u>91.70</u> ↑2.07	<u>74.01</u> ↑2.24	79.00 ↑2.30	86.64 ↑0.93	92.50 ↓3.75	64.50 ↑5.81	82.24 ↑3.13	11.34 ↓29.87
	MIChain (Ours)	**87.10** ↑2.77	**91.77** ↑2.14	**77.46** ↑5.69	**82.38** ↑5.68	**88.94** ↑3.23	**96.25** 0.00	**69.37** ↑10.68	**84.62** ↑5.51	47.61 ↑6.40

Table 2: Comparison of concept prediction performance in terms of AUC, AUPR, accuracy, F1-score, recall, and ranking loss. Best results are in bold, second-best results are underlined. (%)

Method	AUC↑	AUPR↑	Acc↑	F1↑	Recall↑	RL↓
CBM (Koh et al., 2020)	69.17	46.84	81.88	48.71	60.11	11.74
CEM (Zarlenga et al., 2022)	72.88	52.80	90.17	47.57	48.18	7.74
ProbCBM (Kim et al., 2023)	70.78	48.53	86.07	47.99	<u>52.28</u>	10.50
ECBM (Xu et al., 2024)	73.81	<u>53.98</u>	<u>90.73</u>	46.54	47.72	<u>7.40</u>
evi-CEM (Gao et al., 2024)	<u>74.60</u>	52.26	89.40	<u>49.54</u>	50.33	7.42
SSCBM (Hu et al., 2025)	70.62	52.23	89.81	44.58	45.99	8.61
MIChain (Ours)	**81.69**	**60.92**	**91.98**	**60.92**	**62.98**	**6.88**

Table 3: Comparison of lesion detection performance in terms of Precision, Recall, AP, AP_{50}, and AP_{75}. Best results are in bold, second-best results are underlined. (%)

Method	Precision↑	Recall↑	AP↑	AP_{50}↑	AP_{75}↑
DFF (Zhu et al., 2017b)	55.3	86.6	30.0	58.6	29.3
FGFA (Zhu et al., 2017a)	58.4	85.3	30.8	61.1	27.5
RDN (Deng et al., 2019)	61.7	**88.7**	30.8	64.0	20.0
MEGA (Chen et al., 2020)	58.3	85.1	33.1	60.5	30.1
CVA-Net (Lin et al., 2022)	79.9	80.1	47.9	80.4	<u>51.9</u>
YOLOV (Shi et al., 2023)	<u>80.9</u>	<u>87.3</u>	<u>48.6</u>	<u>80.7</u>	51.3
STNet (Qin et al., 2023)	78.8	83.2	46.3	79.2	49.6
MIChain (Ours)	**82.2**	86.9	**50.1**	**81.4**	**53.2**

2.4 Multimodal Explainable Diagnostic Chain

To support clinically interpretable diagnosis, **MIChain** integrates the previously derived visual and textual evidence into a multimodal explainable diagnostic chain. Specifically, it first incorporates lesion localization and margin information to provide traceable visual evidence. Building on this visual evidence, the concept scores are further associated with corresponding clinical descriptions to provide interpretable textual evidence. Finally, the integrated multimodal evidence is used to determine the BI-RADS grade, forming a traceable reasoning process from visual findings and clinical concepts to the final diagnosis.

3 Experiments

3.1 Experimental Setups

Datasets Existing breast ultrasound video datasets are limited in scale and generally lack concept annotations required for interpretable diagnosis. To address these limitations, we conduct experiments on a large-scale private breast ultrasound video dataset, referred to as BUS-VD, which contains 1,085 high-resolution videos acquired using multiple ultrasound devices and by different operators. It is divided into training, validation, and test sets at a ratio of 6:2:2. Five attending physicians independently annotated each video, and all annotations were confirmed by a senior physician. The concept annotations cover 14 clinically relevant diagnostic concepts derived from the ACR BI-RADS. For text annotation, the physicians used the BI-RADS to grade each concept text. Further details of the dataset are provided in Appendix B.

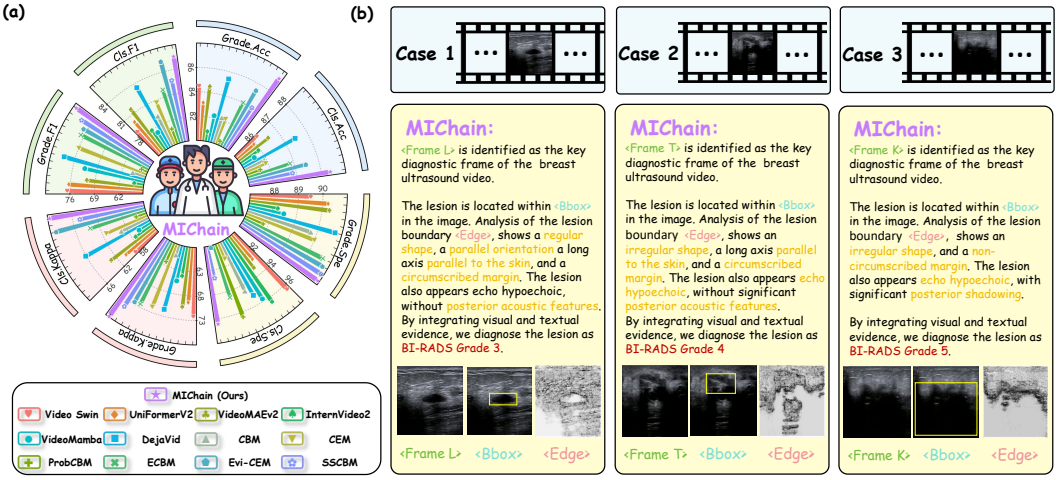


Figure 3: Quantitative comparison and qualitative visualization of MIChain. (a) Performance comparison with existing methods on BI-RADS grading and concept prediction. (b) Representative cases illustrating the multimodal explainable diagnostic chain of MIChain, including visual evidence, textual evidence, and final BI-RADS grade.

Evaluation Metrics and Compared Methods We evaluate breast ultrasound video diagnosis on BI-RADS grading, benign/malignant classification, and lesion concept classification. For the first two tasks, we report accuracy (Acc), specificity (Spe), Kappa, and macro-F1 (F1), while concept classification is evaluated using AUC, AUPR, Acc, F1, Recall, and ranking loss (RL). The compared methods include non-interpretable video models, Video Swin Transformer (Liu et al., 2022), UniformerV2 (Li et al., 2022), VideoMAEv2 (Wang et al., 2023), InternVideo2 (Wang et al., 2024), VideoMamba (Li et al., 2024), and DejaVid (Ho & Madden, 2025), as well as interpretable methods, CBM (Koh et al., 2020), CEM (Zarlenga et al., 2022), ProbCBM (Kim et al., 2023), ECBM (Xu et al., 2024), evi-CEM (Gao et al., 2024), and SSCBM (Hu et al., 2025). For lesion localization, we further compare **MIChain** with representative video object detection methods, DFF (Zhu et al., 2017b), FGFA (Zhu et al., 2017a), RDN (Deng et al., 2019), MEGA (Chen et al., 2020), CVA-Net (Lin et al., 2022), YOLOV (Shi et al., 2023), and STNet (Qin et al., 2023), using Precision, Recall, AP, AP_{50}, and AP_{75}.

Implementation Details All experiments are conducted on an NVIDIA A100 GPU with 80 GB of memory. For each ultrasound video, we sample 16 frames during training and 32 frames during validation, using a frame interval of 5. All sampled frames are resized and cropped to 224 × 224. All models are trained for 10,000 iterations with a batch size of 8. The AdamW optimizer is used with an initial learning rate of 5×10^{-4}, which is dynamically adjusted using a cosine annealing scheduler.

3.2 Comparison with Advanced Methods

Analysis of Diagnostic Performance. As shown in Table 1 and Fig. 3(a), **MIChain** achieves the best overall performance on both BI-RADS grading and benign/malignant classification. For BI-RADS grading, it improves Acc, Kappa, and F1 by 1.85%, 3.45%, and 3.14%, respectively, over the best competing results. It also achieves the highest Acc, Kappa, and F1 for benign/malignant classification, ranking first or tying for first across all eight diagnostic metrics. These results demonstrate the accuracy and reliability of **MIChain** for breast lesion diagnosis.

Analysis of Interpretable Performance. We further evaluate the interpretability of **MIChain** from both textual and visual perspectives. As shown in Table 2, it achieves the best performance across all concept prediction metrics, improving AUC, AUPR, F1, and Recall by 7.09%, 6.94%, 11.38%, and 2.87%, respectively, while reducing RL to 6.88%. This demonstrates its ability to provide reliable concept textual evidence. Meanwhile, as shown in Table 3, MIChain achieves the

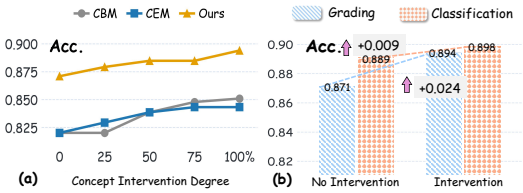


Figure 4: Results of the concept intervention experiments. (a) Comparison of diagnostic accuracy under different concept intervention degrees. (b) Performance comparison before and after concept intervention.

Table 4: Ablation studies of the key components in our model. × and ✓ denote the absence and presence of each component, respectively. All results are reported in percentage (%).

Unpaired Text	Shared Proj.	Evidence Reason.	Evidence Aggr.	Grading		Concept Pred.	
				Acc	Kappa	AUPR	F1
×	×	×	×	81.11	70.73	57.49	53.95
×	✓	×	×	83.41	68.73	58.98	54.86
✓	✓	×	×	84.79	72.22	59.32	55.23
✓	✓	✓	×	84.33	72.45	56.41	59.24
×	×	✓	✓	84.79	74.46	57.25	57.28
×	✓	✓	✓	86.18	75.12	59.56	58.96
✓	✓	✓	✓	**87.10**	**77.46**	**60.92**	**60.92**

best Precision, AP, AP_{50}, and AP_{75} of 82.2%, 50.1%, 81.4%, and 53.2% respectively. Overall, these results demonstrate that **MIChain** provides reliable textual evidence and visual localization for multimodal interpretable diagnosis.

Analysis of Test-time Intervention Capability As a concept-based diagnostic model, **MIChain** supports test-time intervention. We evaluate this capability by progressively increasing the proportion of corrected concepts and comparing MIChain with existing concept-based models. As shown in Fig. 4, MIChain consistently improves with increasing intervention degree and outperforms CBM and CEM throughout. As shown in Fig. 4(b), concept intervention improves the grading and classification accuracy by 2.4% and 0.9%, respectively. This verifies the effectiveness of the proposed intervention method.

3.3 Ablation and Hyperparameter Analysis

Effectiveness of Cross-Modal Shared Concept Representation Learning As shown in Table 4, unpaired text consistently benefits both grading and concept prediction. Adding it to the shared projector improves grading Acc and Kappa by 1.38% and 3.49%, while under evidence-guided video reasoning, Acc, Kappa, AUPR, and F1 increase by 0.92%, 2.34%, 1.36%, and 1.96%. Although the shared projector alone improves Acc, AUPR, and F1 over the baseline, it reduces Kappa by 2.00%, incorporating unpaired text instead yields consistent gains across metrics, confirming the complementarity of textual guidance and shared concept alignment.

Effectiveness of Evidence-Guided Video Reasoning As shown in Table 4, the complete evidence-guided video reasoning module improves grading Acc and Kappa by 2.31% and 5.24%, and concept AUPR and F1 by 1.60% and 5.69%, respectively. Evidence learning alone raises concept F1 from 55.23% to 59.24% but provides limited gains on other metrics. With video evidence aggregation, Acc, Kappa, AUPR, and F1 further increase by 2.77%, 5.01%, 4.51%, and 1.68%, demonstrating their complementary roles in video reasoning.

Additional hyperparameter analyses are provided in Appendix C.2.

4 Conclusion

In this work, we present **MIChain**, a multimodal interpretability chain for video reasoning in breast lesion diagnosis. By projecting unpaired clinical text and ultrasound videos into a shared semantic space, **MIChain** learns clinically meaningful concept representations without relying on paired diagnostic reports. To reason over ultrasound videos, we further introduce evidence-guided video reasoning that evaluates frame evidence reliability and adaptively aggregates diagnostic evidence across frames. Furthermore, visual and textual evidence are integrated into a multimodal explainable diagnostic chain, enabling interpretable BI-RADS grading. Extensive experiments on the BUS-VD dataset demonstrate the effectiveness and capability of our method.

AI USE STATEMENT

We used generative AI tools for language polishing of the manuscript, including improving grammar, clarity, and readability. This use was limited to editing the presentation of the paper. The authors take full responsibility for the final content of this work, including all text, claims, and conclusions.

REFERENCES

Katarzyna Borys, Yasmin Alyssa Schmitt, Meike Nauta, Christin Seifert, Nicole Krämer, Christoph M Friedrich, and Felix Nensa. Explainable ai in medical imaging: An overview for clinical practitioners–saliency-based xai approaches. *European Journal of Radiology*, 162:110787, 2023.

Freddie Bray, Mathieu Laversanne, Hyuna Sung, Jacques Ferlay, Rebecca L Siegel, Isabelle Soerjomataram, and Ahmedin Jemal. Global cancer statistics 2022: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians*, 74(3):229–263, 2024.

Arianna Bunnell, Yannik Glaser, Dustin Valdez, Thomas Wolfgruber, Aleen Altamirano, Carol Zamora González, Brenda Y Hernandez, Peter Sadowski, and John A Shepherd. Learning a clinically-relevant concept bottleneck for lesion detection in breast ultrasound. In *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 650–659, 2024.

Bertrand Charpentier, Daniel Zügner, and Stephan Günnemann. Posterior network: Uncertainty estimation without ood samples via density-based pseudo-counts. *Advances in Neural Information Processing Systems*, 33:1356–1367, 2020.

Hugh Chen, Ian C Covert, Scott M Lundberg, and Su-In Lee. Algorithms to estimate shapley value feature attributions. *Nature Machine Intelligence*, 5(6):590–601, 2023.

Yihong Chen, Yue Cao, Han Hu, and Liwei Wang. Memory enhanced global-local aggregation for video object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10337–10346, 2020.

Jiajun Deng, Yingwei Pan, Ting Yao, Wengang Zhou, Houqiang Li, and Tao Mei. Relation distillation networks for video object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7023–7032, 2019.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the North American Chapter of the Association for Computational Linguistics*, volume 1, pp. 4171–4186, 2019.

Yibo Gao, Zheyao Gao, Xin Gao, Yuanye Liu, Bomin Wang, and Xiahai Zhuang. Evidential concept embedding models: Towards reliable concept explanations for skin disease diagnosis. In *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 308–317, 2024.

Yibo Gao, Hangqi Zhou, Zheyao Gao, Bomin Wang, Shangqi Gao, Sihan Wang, and Xiahai Zhuang. Learning concept-driven logical rules for interpretable and generalizable medical image classification. In *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 291–300, 2025.

Nadia Harbeck, Frédérique Penault-Llorca, Javier Cortes, Michael Gnant, Nehmat Houssami, Philip Poortmans, Kathryn Ruddy, Janice Tsang, and Fatima Cardoso. Breast cancer. *Nature Reviews. Disease primers*, 5(1):66–66, 2019.

Darryl Ho and Samuel Madden. Dejavid: Encoder-agnostic learned temporal matching for video classification. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 24023–24032, 2025.

Regina J Hooley, Leslie M Scoutt, and Liane E Philpotts. Breast ultrasonography: state of the art. *Radiology*, 268(3):642–659, 2013.

Lijie Hu, Tianhao Huang, Huanyi Xie, Xilin Gong, Chenyang Ren, Zhengyu Hu, Lu Yu, Ping Ma, and Di Wang. Semi-supervised concept bottleneck models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2110–2119, 2025.

Ruobing Huang, Qilong Ying, Zehui Lin, Zijie Zheng, Long Tan, Guoxue Tang, Qi Zhang, Man Luo, Xiuwen Yi, Pan Liu, et al. Extracting keyframes of breast ultrasound video using deep reinforcement learning. *Medical Image Analysis*, 80:102490, 2022.

Jing Jiao, Jin Zhou, Xiaokang Li, Menghua Xia, Yi Huang, Lihong Huang, Na Wang, Xiaofan Zhang, Shichong Zhou, Yuanyuan Wang, et al. Usfm: A universal ultrasound foundation model generalized to tasks and organs towards label efficient image analysis. *Medical Image Analysis*, 96:103202, 2024.

Eunji Kim, Dahuin Jung, Sangha Park, Siwon Kim, and Sungroh Yoon. Probabilistic concept bottleneck models, 2023.

Arto Klami, Seppo Virtanen, and Samuel Kaski. Bayesian canonical correlation analysis. *Journal of Machine Learning Research*, 14(4), 2013.

Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In *Proceedings of the International Conference on Machine Learning*, pp. 5338–5348, 2020.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Limin Wang, and Yu Qiao. Uniformerv2: Spatiotemporal learning by arming image vits with video uniformer, 2022. URL https://arxiv.org/abs/2211.09552.

Kunchang Li, Xinhao Li, Yi Wang, Yinan He, Yali Wang, Limin Wang, and Yu Qiao. Videomamba: State space model for efficient video understanding. In *Proceedings of the European Conference on Computer Vision*, pp. 237–255, 2024.

Zehui Lin, Shuo Li, Shanshan Wang, Zhifan Gao, Yue Sun, Chan-Tong Lam, Xindi Hu, Xin Yang, Dong Ni, and Tao Tan. An orchestration learning framework for ultrasound imaging: Prompt-guided hyper-perception and attention-matching downstream synchronization. *Medical Image Analysis*, 104:103639, 2025.

Zhi Lin, Junhao Lin, Lei Zhu, Huazhu Fu, Jing Qin, and Liansheng Wang. A new dataset and a baseline model for breast lesion detection in ultrasound videos. In *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 614–623, 2022.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *Proceedings of the European Conference on Computer Vision*, pp. 38–55, 2024.

Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3202–3211, 2022.

Ellen B Mendelson, Marcela Böhm-Vélez, Wendie A Berg, GJ Whitman, MI Feldman, H Madjar, et al. Acr bi-rads® ultrasound. *ACR BI-RADS® Atlas, Breast Imaging Reporting and Data System*, 2013, 2013.

Yuhao Mo, Chu Han, Yu Liu, Min Liu, Zhenwei Shi, Jiatai Lin, Bingchao Zhao, Chunwang Huang, Bingjiang Qiu, Yanfen Cui, et al. Hover-trans: Anatomy-aware hover-transformer for roi-free breast cancer diagnosis in ultrasound images. *IEEE Transactions on Medical Imaging*, 42(6): 1696–1706, 2023.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2024. URL https://arxiv.org/abs/2304.07193.

Chao Qin, Jiale Cao, Huazhu Fu, Rao Muhammad Anwer, and Fahad Shahbaz Khan. A spatial-temporal deformable attention based framework for breast lesion detection in videos. In *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 479–488, 2023.

Adriel Saporta, Xiaotong Gui, Ashwin Agrawal, Anuj Pareek, Steven QH Truong, Chanh DT Nguyen, Van-Doan Ngo, Jayne Seekins, Francis G Blankenberg, Andrew Y Ng, et al. Benchmarking saliency methods for chest x-ray interpretation. *Nature Machine Intelligence*, 4(10): 867–878, 2022.

Murat Sensoy, Lance Kaplan, and Melih Kandemir. Evidential deep learning to quantify classification uncertainty. *Advances in Neural Information Processing Systems*, 31, 2018.

Bryar Shareef, Min Xian, Aleksandar Vakanski, and Haotian Wang. Breast ultrasound tumor classification using a hybrid multitask cnn-transformer network. In *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 344–353, 2023.

Yuheng Shi, Naiyan Wang, and Xiaojie Guo. Yolov: Making still image object detectors great at video object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 2254–2262, 2023.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, et al. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*, 2025.

Zhuo Su, Wenzhe Liu, Zitong Yu, Dewen Hu, Qing Liao, Qi Tian, Matti Pietikäinen, and Li Liu. Pixel difference networks for efficient edge detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5117–5127, 2021.

Anlan Sun, Zhao Zhang, Meng Lei, Yuting Dai, Dong Wang, and Liwei Wang. Boosting breast ultrasound video classification by the guidance of keyframe feature centers. In *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 441–451, 2023.

Limin Wang, Bingkun Huang, Zhiyu Zhao, Zhan Tong, Yinan He, Yi Wang, Yali Wang, and Yu Qiao. Videomae v2: Scaling video masked autoencoders with dual masking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14549–14560, 2023.

Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Guo Chen, Baoqi Pei, Rongkun Zheng, Zun Wang, Yansong Shi, et al. Internvideo2: Scaling foundation models for multimodal video understanding. In *Proceedings of the European Conference on Computer Vision*, pp. 396–416, 2024.

Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. Medklip: Medical knowledge enhanced language-image pre-training for x-ray diagnosis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 21372–21383, 2023.

Xinyue Xu, Yi Qin, Lu Mi, Hao Wang, and Xiaomeng Li. Energy-based concept bottleneck models: Unifying prediction, concept intervention, and probabilistic interpretations. *arXiv preprint arXiv:2401.14142*, 2024.

Mateo Espinosa Zarlenga, Pietro Barbiero, Gabriele Ciravegna, Giuseppe Marra, Francesco Giannini, Michelangelo Diligenti, Zohreh Shams, Frederic Precioso, Stefano Melacci, Adrian Weller, Pietro Lio, and Mateja Jamnik. Concept embedding models: Beyond the accuracy-explainability trade-off, 2022. URL https://arxiv.org/abs/2209.09056.

Xiaoman Zhang, Chaoyi Wu, Ya Zhang, Weidi Xie, and Yanfeng Wang. Knowledge-enhanced visual-language pre-training on chest radiology images. *Nature Communications*, 14(1):4542, 2023.

Zizhao Zhang, Yuanpu Xie, Fuyong Xing, Mason McGough, and Lin Yang. Mdnet: A semantically and visually interpretable medical image diagnosis network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6428–6436, 2017.

Xizhou Zhu, Yujie Wang, Jifeng Dai, Lu Yuan, and Yichen Wei. Flow-guided feature aggregation for video object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 408–417, 2017a.

Xizhou Zhu, Yuwen Xiong, Jifeng Dai, Lu Yuan, and Yichen Wei. Deep feature flow for video recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2349–2358, 2017b.

A Related Works

A.1 Breast Ultrasound Diagnosis

Deep learning has significantly advanced breast lesion diagnosis from ultrasound images. Specifically, Shareef et al. (2023) combine CNNs and Swin Transformers to perform breast lesion classification, while Mo et al. (2023) incorporate anatomical relationships among breast tissue layers to diagnose breast cancer. More recently, Jiao et al. (2024) proposed USFM to learn transferable ultrasound representations for label-efficient breast lesion analysis. Lin et al. (2025) introduced PerceptGuide, which employs prompt-guided perception and attention matching for ultrasound classification. Several studies have extended breast ultrasound analysis to the realm of video. For example, Huang et al. (2022) proposed a deep reinforcement learning framework for extracting valuable key frames from breast ultrasound videos. Sun et al. (2023) developed KGA-Net, a model that uses feature centers extracted from static key frames to guide frame attention and aggregate video features for classification. However, existing methods still lack interpretability, which limits their application in clinical practice.

A.2 Explainability in Medical Imaging

Interpretability is essential for reliable medical image diagnosis. Earlier studies primarily relied on post-hoc interpretability methods, which make predictions and subsequently explain the resulting decisions. Typical approaches include saliency maps (Saporta et al., 2022; Borys et al., 2023) and Shapley value-based feature attribution (Chen et al., 2023). However, these post-hoc attribution methods mainly provide local explanations and offer limited support for clinically reasoning. Consequently, recent research has shifted toward intrinsically interpretable approaches that incorporate clinically meaningful concepts or reasoning structures directly into the prediction process, including uncertainty-aware concept explanations for skin disease diagnosis (Gao et al., 2024), diagnostic rule learning from visual concepts (Gao et al., 2025), and BI-RADS-based concept bottlenecks for breast ultrasound diagnosis (Bunnell et al., 2024). However, these intrinsically interpretable methods remain limited to unimodal explanations and fail to jointly provide interpretable visual and textual evidence.

B Dataset and Annotation Details

B.1 Dataset Composition

BUS-VD contains 1,085 high-resolution breast ultrasound videos collected using multiple ultrasound devices and by different operators. Each case contributes one diagnostic ultrasound video, resulting in a one-to-one correspondence between cases and videos. The dataset is split at the case level into training, validation, and test subsets, ensuring that the same case does not appear in different subsets.

The dataset includes four BI-RADS grades, comprising 114 Grade 2, 701 Grade 3, 250 Grade 4, and 20 Grade 5 cases. Their distribution is shown in Fig. 6.

B.2 Clinical Concept Annotation

To provide clinically interpretable concept supervision, each case is annotated with 14 BI-RADS ultrasound concepts. These concepts cover the major ultrasound features used in breast lesion assess-

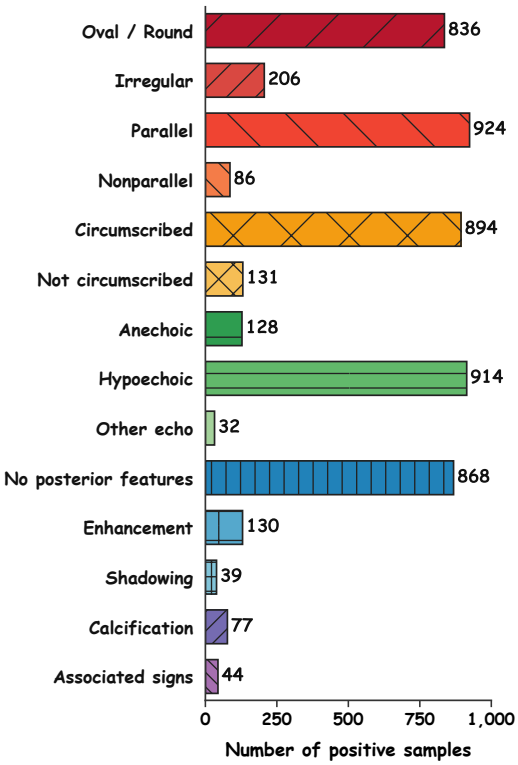


Figure 5: Number of concept-positive cases for each of the 14 clinical concepts in BUS-VD. Each case may be associated with multiple concepts.

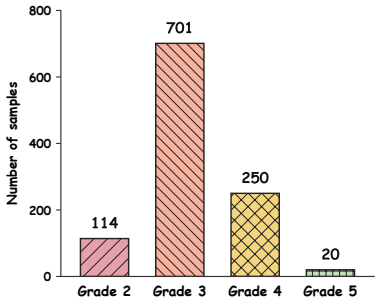


Figure 6: Distribution of BUS-VD cases across the four BI-RADS grades.

Table 5: Clinical concepts annotated in BUS-VD.

Feature group	Clinical concept
Shape	Oval/Round Irregular
Orientation	Parallel Non-parallel
Margin	Circumscribed Non-circumscribed
Echo pattern	Anechoic Hypoechoic Other Echo Pattern
Posterior features	No Posterior Features Posterior Enhancement Posterior Shadowing
Additional features	Calcification Associated Signs

ment, including shape, orientation, margin, echo pattern, posterior acoustic features, calcification, and associated signs. The complete concept set is summarized in Table 5.

The concept annotations are multi-label, since multiple ultrasound features may coexist within the same lesion. Five attending radiologists independently annotated the concept labels, followed by final confirmation by a senior radiologist. The number of positive cases for each concept is shown in Fig. 5. Since each case may contain multiple concepts, the total number of concept-positive annotations exceeds the number of cases in BUS-VD.

Fig. 7 further presents the case-level correspondence between the clinical concepts and BI-RADS grades. The width of each flow represents the number of cases containing the corresponding concept within a given grade.

B.3 Visual Evidence Annotation

To provide spatial supervision for visual evidence generation, BUS-VD includes lesion bounding boxes and lesion masks. The lesion bounding boxes are annotated and verified by professional radiologists, and are stored using a COCO-style JSON representation adapted to ultrasound videos. Each annotation record identifies the video, frame, lesion category, and pixel-level bounding-box coordinates. The fields are summarized in Table 6.

For lesion boundary supervision, the expert-annotated lesion mask is used to obtain the lesion contour and determine the local region around the lesion. Local image edges are then extracted from the corresponding ultrasound frame within this region and combined with the mask-derived contour to

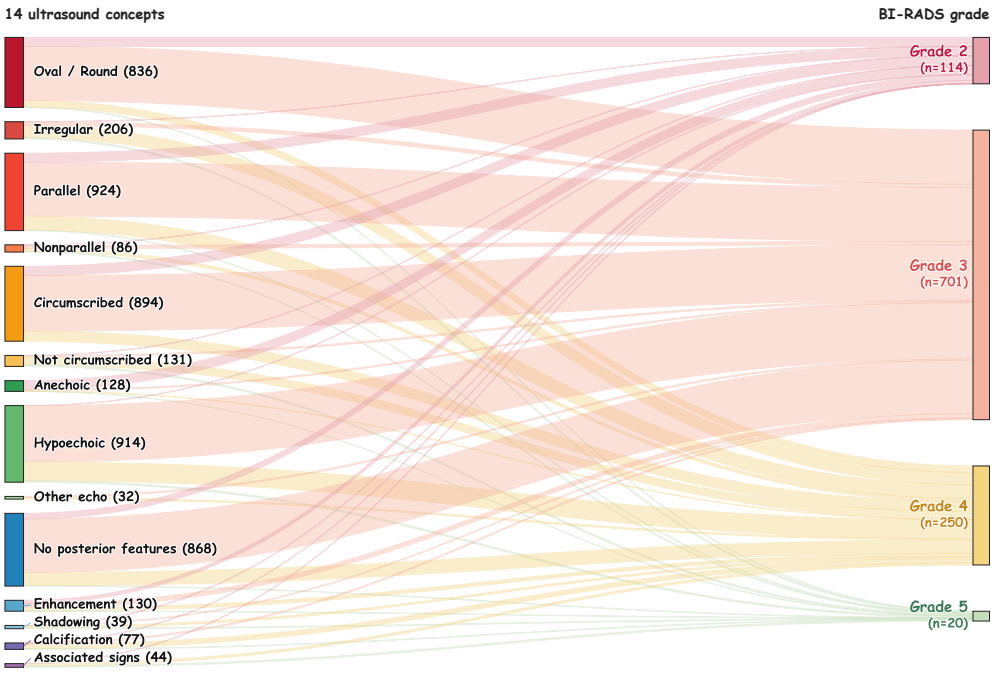


Figure 7: Distribution of the 14 clinical concepts across BI-RADS grades in BUS-VD. The left side shows the clinical concepts and their numbers of positive cases, while the right side shows the number of cases in each BI-RADS grade. The flows indicate their case-level correspondence.

Table 6: Schema of the frame-level lesion localization annotations.

Field	Description
`id`	Unique annotation identifier.
`video_id`	Numerical identifier of the ultrasound video.
`video_name`	Unique name of the ultrasound video.
`frame_id`	Identifier of the annotated frame.
`bbox`	Pixel-level coordinates of the lesion bounding box.
`category`	Semantic category of the annotated object.
`category_id`	Numerical identifier of the object category.

construct the final boundary target. This procedure preserves the expert-defined lesion extent while incorporating local boundary structures from the ultrasound image.

C Additional Experiments Results

C.1 Additional Qualitative Results

To further illustrate the interpretability of **MIChain**, Fig. 8 presents four representative cases covering BI-RADS Grades 2–5. For each case, **MIChain** first identifies a key diagnostic frame from the ultrasound video and localizes the lesion using a bounding box. The corresponding lesion boundary is then provided as additional visual evidence. Based on the selected frame, lesion location, and boundary information, **MIChain** generates concept-level textual evidence describing the relevant ultrasound features and integrates the visual and textual evidence to produce the final BI-RADS grade.

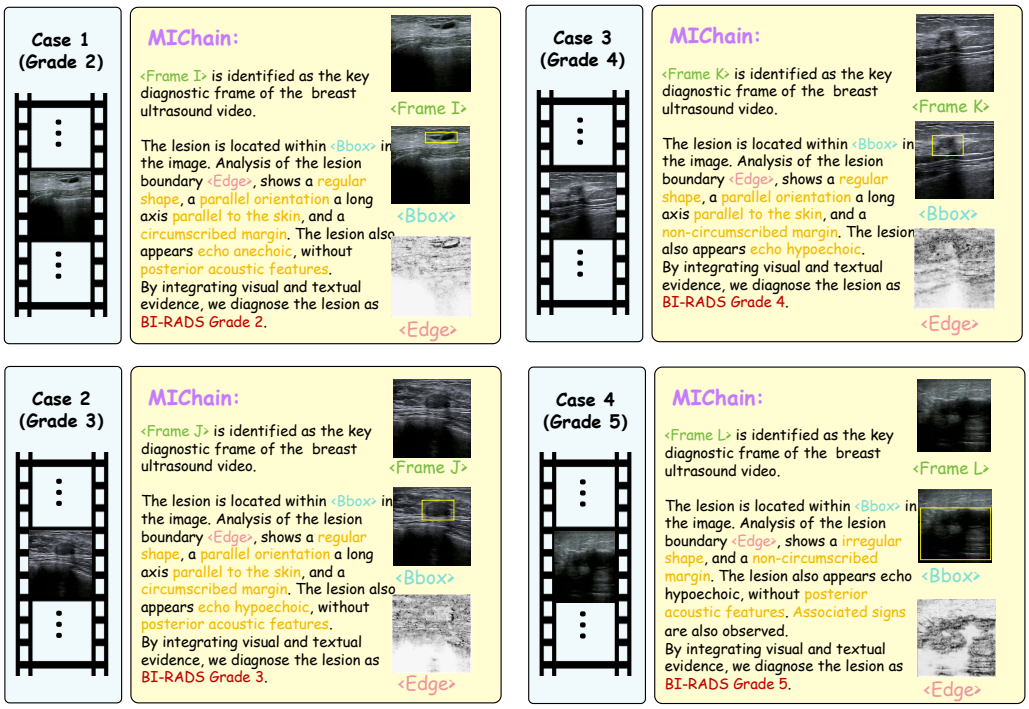


Figure 8: Additional qualitative examples of the multimodal explainable diagnostic chains generated by **MIChain**.

The examples show that **MIChain** provides different levels of diagnostic evidence according to the observed ultrasound features. Cases assigned to lower BI-RADS grades mainly exhibit regular shape, parallel orientation, and circumscribed margins, whereas higher-grade cases contain more suspicious findings, such as non-circumscribed margins, irregular shape, and associated signs. These results allow the final prediction to be traced through the selected frame, lesion localization, boundary evidence, and concept-level textual evidence.

C.2 Hyperparameter Analysis

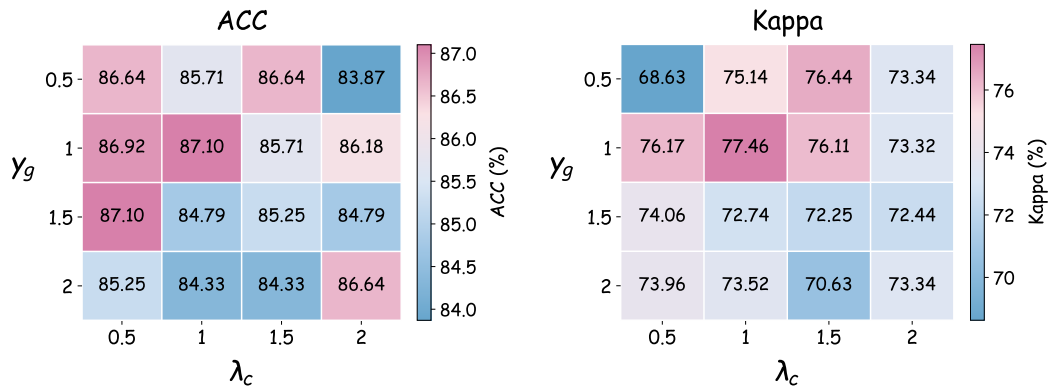
We evaluate four values of γ_g and λ_c from 0.5 to 2.0 with a step of 0.5, and five values of η from 0.1 to 0.5 with a step of 0.1, selecting the final settings based on overall diagnostic performance. As shown in Figs. 9 and 10, $\gamma_g = \lambda_c = 1$ yields 87.10% grading accuracy and the highest Kappa of 77.46%, while $\eta = 0.2$ achieves the best overall performance.

D Construction of the Unpaired Clinical Concept Knowledge Base

D.1 Unpaired Clinical Concept Text Construction

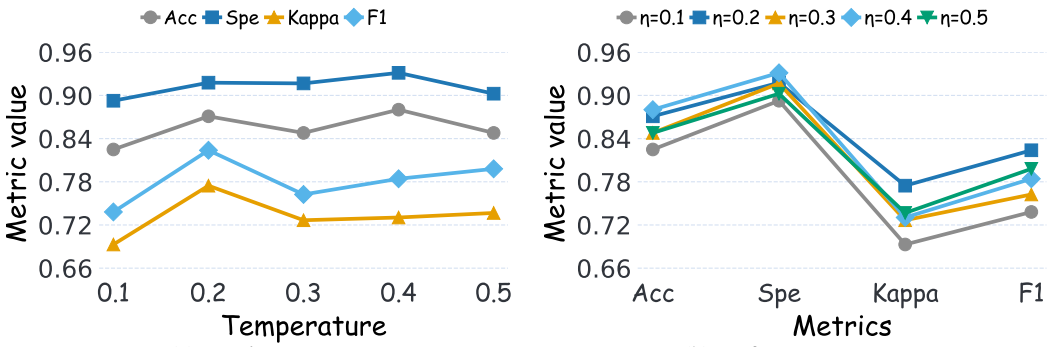
To introduce clinical semantic guidance without relying on paired diagnostic reports, we construct an unpaired clinical concept knowledge base for the 14 ultrasound concepts defined in Appendix A. For each concept, a standardized query is used to request its ultrasound appearance, key characteristics. GPT-5 (Singh et al., 2025) is then used to generate the corresponding concept-level description.

The generated descriptions are further reviewed by radiologists to correct inaccurate or ambiguous terminology and ensure consistency with the clinical definitions of the corresponding concepts. Each entry contains a concept name and a clinical description. Since the construction process does not use any patient-specific video or diagnostic report from BUS-VD, the resulting texts remain independent



(a) Grading Acc (b) Grading Kappa

Figure 9: Results of the parameter analysis on γ_g and λ_c. The figure presents BI-RADS grading Acc and Kappa under different values of γ_g and λ_c.



(a) Metrics across η (b) Performance across η

Figure 10: Results of the parameter analysis on temperature η. The figure presents diagnostic performance under different temperature values from two complementary views.

of individual ultrasound cases and do not require paired video and text. Fig. 11 presents an example of the construction process.

D.2 Annotation of Unpaired Clinical Concept Text

In addition to the clinical description, physicians annotate each concept text with the BI-RADS grades that the described ultrasound feature may support in clinical assessment. Since an individual ultrasound concept may be observed in lesions assigned to more than one BI-RADS grade, the potential grade outcomes indicated by concept text $\mathbf{t}_j$ are represented by a four-dimensional vector:

$$\mathbf{y}_j^{\text{txt}} = \left[y_j^{(2)}, y_j^{(3)}, y_j^{(4)}, y_j^{(5)}\right] \in \{0,1\}^4, \tag{14}$$

where $y_j^{(g)} = 1$ indicates that the ultrasound feature described by $\mathbf{t}_j$ may provide diagnostic support for BI-RADS Grade g.

For example, the label of the Irregular Shape concept text is $[0,0,1,1]$, indicating that an irregular shape may support BI-RADS Grades 4 and 5 in clinical assessment. These labels do not define a deterministic mapping from a single concept to the final grade, since BI-RADS grading is determined by jointly considering multiple ultrasound features. During training, the physician-annotated labels supervise the textual branch to predict the potential BI-RADS grades indicated by each aligned textual concept representation $\mathbf{z}_j^{\text{txt}}$. The complete labels of the 14 concept texts are summarized in Table 7.

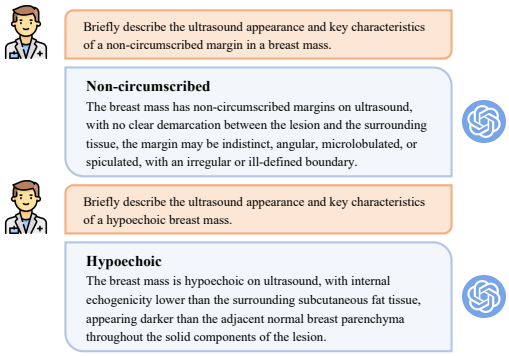


Figure 11: Example of the unpaired clinical concept text construction. A standardized clinical query is used to generate the ultrasound description.

Table 7: BI-RADS grade associations of the clinical concepts used in the unpaired concept knowledge base. A checkmark indicates that the corresponding concept is associated with the given BI-RADS grade in our knowledge-base construction.

Clinical Concept	Grade2	Grade3	Grade4	Grade5
Oval/Round Shape	✓	✓		
Irregular Shape			✓	✓
Parallel Orientation	✓	✓		
Non-parallel Orientation			✓	✓
Circumscribed Margin	✓	✓		
Non-circumscribed Margin			✓	✓
Anechoic	✓	✓		
Hypoechoic		✓	✓	
Other Echo Pattern		✓	✓	
No Posterior Features		✓		
Posterior Enhancement	✓	✓		
Posterior Shadowing			✓	✓
Calcification		✓	✓	✓
Associated Signs			✓	✓

E Proofs of Theoretical Results

This section provides detailed proofs of Theorems 1 and 2. We first establish the required matrix-ordering and Fisher-information preliminaries, and then derive the visual and textual information contributions under the linear Gaussian model. These results are subsequently used to prove the global and directional information reduction induced by unpaired clinical text. Throughout the analysis, we adopt the regularity, identifiability, and conditional-independence assumptions of the main paper, with directions unidentifiable from visual observations alone treated as having unbounded visual-only estimation variance.

E.1 Background and Definitions

In this subsection, we review the mathematical definitions used in our theoretical analysis, including positive semidefinite and positive definite matrices, Loewner ordering, and Fisher information.

Definition 1 (Positive Semidefinite Matrix)**.** *A real symmetric matrix* $\mathbf{M} \in \mathbb{R}^{d\times d}$ *is* positive semidefinite *if* $\mathbf{r}^\top \mathbf{M}\mathbf{r} \geq 0$ *for every* $\mathbf{r} \in \mathbb{R}^d$*. Equivalently, all eigenvalues of* $\mathbf{M}$ *are nonnegative. We denote the set of all* $d \times d$ *symmetric positive semidefinite matrices by* $\mathbb{S}^d_{\succeq 0}$*.*

Definition 2 (Positive Definite Matrix)**.** *A real symmetric matrix* $\mathbf{M} \in \mathbb{R}^{d\times d}$ *is* positive definite *if* $\mathbf{r}^\top \mathbf{M}\mathbf{r} > 0$ *for every nonzero* $\mathbf{r} \in \mathbb{R}^d$*. Equivalently, all eigenvalues of* $\mathbf{M}$ *are strictly positive. We denote the set of all* $d \times d$ *symmetric positive definite matrices by* $\mathbb{S}^d_{\succ 0}$*.*

Definition 3 (Loewner Order)**.** *For two real symmetric matrices* $\mathbf{M}, \mathbf{N} \in \mathbb{R}^{d\times d}$*, we write* $\mathbf{M} \preceq \mathbf{N}$ *if* $\mathbf{N} - \mathbf{M}$ *is positive semidefinite, and* $\mathbf{M} \prec \mathbf{N}$ *if* $\mathbf{N} - \mathbf{M}$ *is positive definite. Equivalently,* $\mathbf{M} \preceq \mathbf{N} \iff \mathbf{N} - \mathbf{M} \succeq \mathbf{0}$ *and* $\mathbf{M} \prec \mathbf{N} \iff \mathbf{N} - \mathbf{M} \succ \mathbf{0}$*. This defines a partial order on the space of real symmetric matrices.*

Definition 4 (Fisher Information Matrix)**.** *Given a parametric family of densities* $p(\mathbf{x};\boldsymbol{\theta})$*, the Fisher information matrix at parameter* $\boldsymbol{\theta}$ *is defined as*

$$\mathcal{I}(\boldsymbol{\theta}) = \mathbb{E}_{\mathbf{x}\sim p(\cdot;\boldsymbol{\theta})}\left[\nabla_{\boldsymbol{\theta}} \log p(\mathbf{x};\boldsymbol{\theta}) \nabla_{\boldsymbol{\theta}} \log p(\mathbf{x};\boldsymbol{\theta})^\top\right], \tag{C1-1}$$

Equivalently, under the standard regularity conditions,

$$\mathcal{I}(\boldsymbol{\theta}) = -\mathbb{E}_{\mathbf{x}\sim p(\cdot;\boldsymbol{\theta})}\left[\nabla^2_{\boldsymbol{\theta}} \log p(\mathbf{x};\boldsymbol{\theta})\right]. \tag{C1-2}$$

E.2 Least-Squares Estimators and Fisher Contributions

In this subsection, we revisit the linear-Gaussian model introduced in the main paper, formulate the visual-only, text-only, and joint estimators, and derive their Fisher information contributions to the shared clinical concept factors.

Linear-Gaussian Shared Factor model. Recall that the visual and textual embeddings are modeled as different observations of a shared clinical concept space. Let $\boldsymbol{\theta} \in \mathbb{R}^d$ denote the latent parameter vector, decomposed as $\boldsymbol{\theta} = [\boldsymbol{\theta}_c, \boldsymbol{\theta}_v, \boldsymbol{\theta}_t]^\top$, where $\boldsymbol{\theta}_c \in \mathbb{R}^{d_c}$ represents the clinical concept factors shared by the visual and textual modalities, $\boldsymbol{\theta}_v \in \mathbb{R}^{d_v}$ represents visual-specific factors, and $\boldsymbol{\theta}_t \in \mathbb{R}^{d_t}$ represents text-specific factors, with $d_c + d_v + d_t = d$.

We consider N_v visual observations $\{\mathbf{h}_i^{\text{vis}}\}_{i=1}^{N_v}$ and N_t textual observations $\{\mathbf{h}_j^{\text{txt}}\}_{j=1}^{N_t}$. Following the model in the main paper, each observation is expressed as a linear combination of the shared and modality-specific factors:

$$\begin{aligned} \mathbf{h}_i^{\text{vis}} &= \mathbf{A}_{c,i}\boldsymbol{\theta}_c + \mathbf{A}_{v,i}\boldsymbol{\theta}_v + \boldsymbol{\epsilon}_{v,i}, \quad \boldsymbol{\epsilon}_{v,i} \sim \mathcal{N}\left(\mathbf{0}, \sigma_v^2\mathbf{I}\right), \\ \mathbf{h}_j^{\text{txt}} &= \mathbf{B}_{c,j}\boldsymbol{\theta}_c + \mathbf{B}_{t,j}\boldsymbol{\theta}_t + \boldsymbol{\epsilon}_{t,j}, \quad \boldsymbol{\epsilon}_{t,j} \sim \mathcal{N}\left(\mathbf{0}, \sigma_t^2\mathbf{I}\right). \end{aligned} \tag{C2-1}$$

Here, $\mathbf{A}_{c,i}$ and $\mathbf{B}_{c,j}$ characterize how the visual and textual observations probe the shared concept factors, whereas $\mathbf{A}_{v,i}$ and $\mathbf{B}_{t,j}$ capture modality-specific variation. The noise variables are assumed to be independent across observations and modalities.

In this linear setting, the estimation precision of the shared concept factors is characterized by the Fisher information $\mathcal{I}(\boldsymbol{\theta})$. Intuitively, larger curvature of the log-likelihood along a direction in $\boldsymbol{\theta}_c$ indicates that the observations more strongly constrain the corresponding concept factor, thereby reducing its estimation uncertainty. Our analysis therefore compares the information contributed to $\boldsymbol{\theta}_c$ by visual observations alone and by visual observations together with unpaired clinical text.

Unimodal Estimators. We first consider estimation from the visual observations alone. Stacking $\{\mathbf{h}_i^{\text{vis}}\}_{i=1}^{N_v}$ gives a design matrix whose block rows are $[\mathbf{A}_{c,i}, \mathbf{A}_{v,i}, \mathbf{0}]$. The corresponding least-squares estimator is

$$\widehat{\boldsymbol{\theta}}^{\,v} = \arg\min_{\boldsymbol{\theta}} \sum_{i=1}^{N_v} \left\| \mathbf{h}_i^{\text{vis}} - \mathbf{A}_{c,i}\boldsymbol{\theta}_c - \mathbf{A}_{v,i}\boldsymbol{\theta}_v \right\|_2^2. \tag{C2-2}$$

The objective is independent of $\boldsymbol{\theta}_t$; hence, visual observations provide no information about the text-specific factors.

Analogously, stacking $\{\mathbf{h}_j^{\text{txt}}\}_{j=1}^{N_t}$ gives block rows $[\mathbf{B}_{c,j}, \mathbf{0}, \mathbf{B}_{t,j}]$, and the text-only estimator is

$$\widehat{\boldsymbol{\theta}}^{\,t} = \arg\min_{\boldsymbol{\theta}} \sum_{j=1}^{N_t} \left\| \mathbf{h}_j^{\text{txt}} - \mathbf{B}_{c,j}\boldsymbol{\theta}_c - \mathbf{B}_{t,j}\boldsymbol{\theta}_t \right\|_2^2. \tag{C2-3}$$

This objective is independent of $\boldsymbol{\theta}_v$ and thus provides no information about the visual-specific factors. Importantly, although each modality alone is blind to the factors specific to the other modality, both estimators depend on the same shared factors $\boldsymbol{\theta}_c$.

Joint Estimator from Unpaired Observations. The clinical texts used by **MIChain** are not diagnostic reports paired with individual ultrasound videos. Thus, no correspondence between $\mathbf{h}_i^{\text{vis}}$ and $\mathbf{h}_j^{\text{txt}}$ is assumed. Nevertheless, both sets of observations depend on the common shared factors $\boldsymbol{\theta}_c$. Since the visual and textual observations are conditionally independent given the latent factors, their joint likelihood factorizes as

$$\begin{aligned} &p\left(\{\mathbf{h}_i^{\text{vis}}\}_{i=1}^{N_v}, \{\mathbf{h}_j^{\text{txt}}\}_{j=1}^{N_t} \mid \boldsymbol{\theta}\right) \\ &\quad = \prod_{i=1}^{N_v} p\left(\mathbf{h}_i^{\text{vis}} \mid \boldsymbol{\theta}_c, \boldsymbol{\theta}_v\right) \prod_{j=1}^{N_t} p\left(\mathbf{h}_j^{\text{txt}} \mid \boldsymbol{\theta}_c, \boldsymbol{\theta}_t\right). \end{aligned} \tag{C2-4}$$

Maximizing this likelihood is equivalent, under the Gaussian model, to the joint least-squares problem

$$\begin{aligned} \widehat{\boldsymbol{\theta}}^{\,v,t} = \arg\min_{\boldsymbol{\theta}} \Bigg\{ &\sum_{i=1}^{N_v} \left\| \mathbf{h}_i^{\text{vis}} - \mathbf{A}_{c,i}\boldsymbol{\theta}_c - \mathbf{A}_{v,i}\boldsymbol{\theta}_v \right\|_2^2 \\ &+ \sum_{j=1}^{N_t} \left\| \mathbf{h}_j^{\text{txt}} - \mathbf{B}_{c,j}\boldsymbol{\theta}_c - \mathbf{B}_{t,j}\boldsymbol{\theta}_t \right\|_2^2 \Bigg\}. \end{aligned} \tag{C2-5}$$

Therefore, sample-wise video–text pairing is not required for joint estimation. The two modalities are coupled through the common parameters $\boldsymbol{\theta}_c$: visual observations and clinical texts independently constrain the same shared concept factors while remaining unpaired at the sample level.

Fisher Information. Under the linear-Gaussian model, each modality contributes a block-structured Fisher information matrix. For the visual observations, we have

$$\mathcal{I}_v(\boldsymbol{\theta}) = \sum_{i=1}^{N_v} \begin{bmatrix} \mathbf{A}_{c,i}^\top \mathbf{A}_{c,i} & \mathbf{A}_{c,i}^\top \mathbf{A}_{v,i} & \mathbf{0} \\ \mathbf{A}_{v,i}^\top \mathbf{A}_{c,i} & \mathbf{A}_{v,i}^\top \mathbf{A}_{v,i} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} \end{bmatrix}, \tag{C2-6}$$

whereas the textual observations contribute

$$\mathcal{I}_t(\boldsymbol{\theta}) = \sum_{j=1}^{N_t} \begin{bmatrix} \mathbf{B}_{c,j}^\top \mathbf{B}_{c,j} & \mathbf{0} & \mathbf{B}_{c,j}^\top \mathbf{B}_{t,j} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{B}_{t,j}^\top \mathbf{B}_{c,j} & \mathbf{0} & \mathbf{B}_{t,j}^\top \mathbf{B}_{t,j} \end{bmatrix}. \tag{C2-7}$$

Because the visual and textual observations are conditionally independent, their log-likelihood contributions add, and hence their Fisher information matrices are additive:

$$\mathcal{I}_{v,t}(\boldsymbol{\theta}) = \mathcal{I}_v(\boldsymbol{\theta}) + \mathcal{I}_t(\boldsymbol{\theta}). \tag{C2-8}$$

Since our analysis focuses on the shared clinical concept factors $\boldsymbol{\theta}_c$, we consider the corresponding shared-factor blocks of the Fisher information. Consistent with the notation in the main paper, we denote

$$\begin{aligned} \mathcal{J}_v &:= (\mathcal{I}_v(\boldsymbol{\theta}))_{\theta_c,\theta_c} = \sum_{i=1}^{N_v} \mathbf{A}_{c,i}^\top \mathbf{A}_{c,i}, \\ \mathcal{J}_t &:= (\mathcal{I}_t(\boldsymbol{\theta}))_{\theta_c,\theta_c} = \sum_{j=1}^{N_t} \mathbf{B}_{c,j}^\top \mathbf{B}_{c,j}. \end{aligned} \tag{C2-9}$$

Therefore, the shared-factor block of the joint Fisher information is

$$\begin{aligned} \mathcal{J}_{v,t} &:= (\mathcal{I}_{v,t}(\boldsymbol{\theta}))_{\theta_c,\theta_c} \\ &= (\mathcal{I}_v(\boldsymbol{\theta}))_{\theta_c,\theta_c} + (\mathcal{I}_t(\boldsymbol{\theta}))_{\theta_c,\theta_c} \\ &= \mathcal{J}_v + \mathcal{J}_t \\ &= \sum_{i=1}^{N_v} \mathbf{A}_{c,i}^\top \mathbf{A}_{c,i} + \sum_{j=1}^{N_t} \mathbf{B}_{c,j}^\top \mathbf{B}_{c,j}. \end{aligned} \tag{C2-10}$$

This decomposition shows that the unpaired textual observations contribute an additional Fisher information term $\mathcal{J}_t$ to the shared clinical concept factors. It forms the basis of the global and directional information gains established in Theorems 1 and 2.

E.3 Supporting Lemmas

We now prove Theorems 1 and 2 using the shared-factor Fisher information decomposition derived above.

Lemma 1 (Loewner Order Reversal for Inverses)**.** *Let* $\mathbf{M}, \mathbf{N} \in \mathbb{R}^{d\times d}$ *be positive-definite matrices. If* $\mathbf{M} \prec \mathbf{N}$*, then* $\mathbf{N}^{-1} \prec \mathbf{M}^{-1}$.

Proof. Since $\mathbf{N} \succ \mathbf{0}$, $\mathbf{N}^{-1/2}$ exists and is nonsingular. Define

$$\mathbf{C} := \mathbf{N}^{-1/2}\mathbf{M}\mathbf{N}^{-1/2}. \tag{C3-1}$$

Because congruence by an invertible matrix preserves positive definiteness, $\mathbf{M} \prec \mathbf{N}$ implies

$$\mathbf{0} \prec \mathbf{C} \prec \mathbf{I}. \tag{C3-2}$$

Hence, every eigenvalue of $\mathbf{C}$ lies in $(0, 1)$, which implies

$$\mathbf{C}^{-1} \succ \mathbf{I}. \tag{C3-3}$$

Using

$$\mathbf{M} = \mathbf{N}^{1/2}\mathbf{C}\mathbf{N}^{1/2}, \tag{C3-4}$$

we obtain

$$\begin{aligned} \mathbf{M}^{-1} &= \mathbf{N}^{-1/2}\mathbf{C}^{-1}\mathbf{N}^{-1/2} \\ &\succ \mathbf{N}^{-1/2}\mathbf{I}\mathbf{N}^{-1/2} = \mathbf{N}^{-1}. \end{aligned} \tag{C3-5}$$

Therefore, $\mathbf{N}^{-1} \prec \mathbf{M}^{-1}$.

Lemma 2 (Directional Inverse Monotonicity)**.** *Let $\mathbf{M}, \mathbf{N} \in \mathbb{R}^{d\times d}$ be positive-definite matrices satisfying $\mathbf{M} \preceq \mathbf{N}$. Let*

$$\mathbf{\Delta} := \mathbf{N} - \mathbf{M} \succeq \mathbf{0}. \tag{C3-6}$$

Then, for any nonzero $\mathbf{r} \in \mathbb{R}^d$,

$$\mathbf{r}^\top \mathbf{N}^{-1}\mathbf{r} \leq \mathbf{r}^\top \mathbf{M}^{-1}\mathbf{r}. \tag{C3-7}$$

Moreover, the inequality is strict if

$$\mathbf{\Delta}\mathbf{M}^{-1}\mathbf{r} \neq \mathbf{0}. \tag{C3-8}$$

If $\mathbf{r}^\top \mathbf{\Delta}\mathbf{r} > 0$, then there also exists at least one nonzero direction $\mathbf{u}$ such that

$$\mathbf{u}^\top \mathbf{N}^{-1}\mathbf{u} < \mathbf{u}^\top \mathbf{M}^{-1}\mathbf{u}. \tag{C3-9}$$

Proof. Let

$$\mathbf{C} := \mathbf{M}^{-1/2}\mathbf{\Delta}\mathbf{M}^{-1/2} \succeq \mathbf{0}. \tag{C3-10}$$

Then

$$\mathbf{N} = \mathbf{M}^{1/2}(\mathbf{I} + \mathbf{C})\mathbf{M}^{1/2}, \tag{C3-11}$$

and therefore

$$\mathbf{N}^{-1} = \mathbf{M}^{-1/2}(\mathbf{I} + \mathbf{C})^{-1}\mathbf{M}^{-1/2}. \tag{C3-12}$$

Hence,

$$\begin{aligned} \mathbf{M}^{-1} - \mathbf{N}^{-1} &= \mathbf{M}^{-1/2}\left[\mathbf{I} - (\mathbf{I} + \mathbf{C})^{-1}\right]\mathbf{M}^{-1/2} \\ &= \mathbf{M}^{-1/2}\mathbf{C}(\mathbf{I} + \mathbf{C})^{-1}\mathbf{M}^{-1/2}. \end{aligned} \tag{C3-13}$$

Since $\mathbf{C} \succeq \mathbf{0}$, the matrix $\mathbf{C}(\mathbf{I} + \mathbf{C})^{-1}$ is positive semidefinite. Therefore,

$$\mathbf{r}^\top \left(\mathbf{M}^{-1} - \mathbf{N}^{-1}\right)\mathbf{r} \geq 0, \tag{C3-14}$$

which proves the non-increasing relation.

The inequality is strict exactly when

$$\mathbf{C}\mathbf{M}^{-1/2}\mathbf{r} \neq \mathbf{0}. \tag{C3-15}$$

Using the definition of $\mathbf{C}$ and the invertibility of $\mathbf{M}^{-1/2}$, this condition is equivalent to

$$\mathbf{\Delta}\mathbf{M}^{-1}\mathbf{r} \neq \mathbf{0}. \tag{C3-16}$$

Finally, if $\mathbf{r}^\top \mathbf{\Delta}\mathbf{r} > 0$, then $\mathbf{\Delta} \neq \mathbf{0}$ and hence $\mathbf{C} \neq \mathbf{0}$. Thus, $\mathbf{C}$ has at least one strictly positive eigenvalue. Choosing a corresponding eigenvector and mapping it back through $\mathbf{M}^{1/2}$ yields a nonzero direction $\mathbf{u}$ for which the above inequality is strict.

E.4 Proof of Theorem 1

Restatement of Theorem 1 (Variance Reduction from Unpaired Clinical Text). Let $\widehat{\boldsymbol{\theta}}_c^{\,v}$ and $\widehat{\boldsymbol{\theta}}_c^{\,v,t}$ denote the least-squares estimators of the shared concept factors obtained from visual observations alone and jointly from visual observations and unpaired clinical text, respectively. Under the standard regularity and identifiability conditions, if at least one textual design matrix $\mathbf{B}_{c,j}$ has full column rank, then $\mathcal{J}_t \succ \mathbf{0}$, and

$$\mathcal{J}_{v,t} = \left(\mathcal{I}_v(\boldsymbol{\theta}) + \mathcal{I}_t(\boldsymbol{\theta})\right)_{\theta_c,\theta_c} \succ \left(\mathcal{I}_v(\boldsymbol{\theta})\right)_{\theta_c,\theta_c} = \mathcal{J}_v. \tag{C4-1}$$

Consequently, the covariance of the shared concept factor estimator satisfies the strict Loewner ordering

$$\mathrm{Var}\left(\widehat{\boldsymbol{\theta}}_c^{\,v,t}\right) \prec \mathrm{Var}\left(\widehat{\boldsymbol{\theta}}_c^{\,v}\right). \tag{C4-2}$$

Proof. We first establish the positive-semidefinite property of Fisher information. Let

$$\mathbf{s}(\boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}} \log p(\mathbf{x}; \boldsymbol{\theta}) \tag{C4-3}$$

denote the score function. For any $\mathbf{r} \in \mathbb{R}^d$,

$$\begin{aligned} \mathbf{r}^\top \mathcal{I}(\boldsymbol{\theta})\mathbf{r} &= \mathbf{r}^\top \mathbb{E}\left[\mathbf{s}(\boldsymbol{\theta})\mathbf{s}(\boldsymbol{\theta})^\top\right]\mathbf{r} \\ &= \mathbb{E}\left[\left(\mathbf{r}^\top \mathbf{s}(\boldsymbol{\theta})\right)^2\right] \geq 0. \end{aligned} \tag{C4-4}$$

Hence, every Fisher information matrix is positive semidefinite.

Using the shared-factor Fisher information derived above,

$$\mathcal{J}_t = \sum_{j=1}^{N_t} \mathbf{B}_{c,j}^\top \mathbf{B}_{c,j}. \tag{C4-5}$$

For every j and every $\mathbf{r} \in \mathbb{R}^{d_c}$,

$$\mathbf{r}^\top \mathbf{B}_{c,j}^\top \mathbf{B}_{c,j} \mathbf{r} = \|\mathbf{B}_{c,j}\mathbf{r}\|_2^2 \geq 0, \tag{C4-6}$$

so each $\mathbf{B}_{c,j}^\top \mathbf{B}_{c,j}$ is positive semidefinite.

By assumption, there exists at least one index j^* such that $\mathbf{B}_{c,j^*}$ has full column rank. Therefore, for every nonzero $\mathbf{r} \in \mathbb{R}^{d_c}$,

$$\mathbf{r}^\top \mathbf{B}_{c,j^*}^\top \mathbf{B}_{c,j^*} \mathbf{r} = \|\mathbf{B}_{c,j^*}\mathbf{r}\|_2^2 > 0. \tag{C4-7}$$

Thus, $\mathbf{B}_{c,j^*}^\top \mathbf{B}_{c,j^*}$ is positive definite. Since every summand in $\mathcal{J}_t$ is positive semidefinite and at least one summand is positive definite,

$$\mathcal{J}_t \succ \mathbf{0}. \tag{C4-8}$$

From the shared-factor Fisher-information decomposition established in the previous subsection,

$$\mathcal{J}_{v,t} = \mathcal{J}_v + \mathcal{J}_t. \tag{C4-9}$$

Since $\mathcal{J}_t \succ \mathbf{0}$,

$$\mathcal{J}_{v,t} - \mathcal{J}_v = \mathcal{J}_t \succ \mathbf{0}, \tag{C4-10}$$

and therefore

$$\mathcal{J}_v \prec \mathcal{J}_{v,t}. \tag{C4-11}$$

Equivalently,

$$\left(\mathcal{I}_v(\boldsymbol{\theta}) + \mathcal{I}_t(\boldsymbol{\theta})\right)_{\theta_c,\theta_c} \succ \left(\mathcal{I}_v(\boldsymbol{\theta})\right)_{\theta_c,\theta_c}. \tag{C4-12}$$

We next relate this strict information gain to the covariance of the shared concept factor estimator. Under the linear-Gaussian model, after whitening with respect to the Gaussian noise covariance, maximizing the likelihood is equivalent to minimizing the least-squares objective. Thus, the least-squares estimator coincides with the maximum-likelihood estimator.

Under the standard regularity conditions, the maximum-likelihood estimator around the true parameter $\boldsymbol{\theta}_0$ is asymptotically normal:

$$\sqrt{n}\left(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0\right) \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \mathcal{I}(\boldsymbol{\theta}_0)^{-1}\right), \tag{C4-13}$$

where n denotes the relevant sample size. Therefore, on identifiable parameter directions, the estimator covariance is characterized by the inverse Fisher information. Applying this relation to the shared concept factors, their covariance is determined by the inverse of the corresponding shared-factor Fisher information matrix.

Under the visual identifiability condition, $\mathcal{J}_v \succ \mathbf{0}$. Since $\mathcal{J}_t \succ \mathbf{0}$, we also have

$$\mathcal{J}_{v,t} = \mathcal{J}_v + \mathcal{J}_t \succ \mathbf{0}. \tag{C4-14}$$

Therefore, the corresponding covariance matrices satisfy

$$\mathrm{Var}\left(\widehat{\boldsymbol{\theta}}_c^{\,v}\right) = \mathcal{J}_v^{-1}, \qquad \mathrm{Var}\left(\widehat{\boldsymbol{\theta}}_c^{\,v,t}\right) = \mathcal{J}_{v,t}^{-1}. \tag{C4-15}$$

Because $\mathcal{J}_v \prec \mathcal{J}_{v,t}$, Lemma 1 gives

$$\mathcal{J}_{v,t}^{-1} \prec \mathcal{J}_v^{-1}. \tag{C4-16}$$

Hence,

$$\mathrm{Var}\left(\widehat{\boldsymbol{\theta}}_c^{\,v,t}\right) \prec \mathrm{Var}\left(\widehat{\boldsymbol{\theta}}_c^{\,v}\right). \tag{C4-17}$$

E.5 Proof of Theorem 2

Restatement of Theorem 2 (Directional Information Gain from Unpaired Clinical Text). Let $\mathbf{r} \in \mathbb{R}^{d_c}$ be a nonzero direction in the shared concept space. If there exists at least one textual observation $j \in \{1, \ldots, N_t\}$ such that $\mathbf{B}_{c,j}\mathbf{r} \neq \mathbf{0}$, then

$$\mathbf{r}^\top \mathcal{J}_{v,t}\mathbf{r} = \mathbf{r}^\top \left[\mathcal{I}_v(\boldsymbol{\theta}) + \mathcal{I}_t(\boldsymbol{\theta})\right]_{\theta_c,\theta_c} \mathbf{r} > \mathbf{r}^\top \left[\mathcal{I}_v(\boldsymbol{\theta})\right]_{\theta_c,\theta_c} \mathbf{r} = \mathbf{r}^\top \mathcal{J}_v \mathbf{r}. \tag{C4-18}$$

Furthermore, if $\mathbf{r} \notin \mathrm{range}(\mathcal{J}_v)$ and the corresponding direction becomes identifiable under the joint observations, then the estimation uncertainty along $\mathbf{r}$ is reduced from unbounded under visual-only observations to finite under the joint observations.

Proof. For any nonzero $\mathbf{r} \in \mathbb{R}^{d_c}$, the textual Fisher information along $\mathbf{r}$ is

$$\mathbf{r}^\top \mathcal{J}_t \mathbf{r} = \sum_{j'=1}^{N_t} \|\mathbf{B}_{c,j'}\mathbf{r}\|_2^2. \tag{C4-19}$$

By assumption, there exists at least one textual observation j such that $\mathbf{B}_{c,j}\mathbf{r} \neq \mathbf{0}$. Therefore,

$$\mathbf{r}^\top \mathcal{J}_t \mathbf{r} \geq \|\mathbf{B}_{c,j}\mathbf{r}\|_2^2 > 0. \tag{C4-20}$$

Since

$$\mathcal{J}_{v,t} = \mathcal{J}_v + \mathcal{J}_t, \tag{C4-21}$$

we obtain

$$\begin{aligned} \mathbf{r}^\top \mathcal{J}_{v,t}\mathbf{r} &= \mathbf{r}^\top \mathcal{J}_v \mathbf{r} + \mathbf{r}^\top \mathcal{J}_t \mathbf{r} \\ &> \mathbf{r}^\top \mathcal{J}_v \mathbf{r}. \end{aligned} \tag{C4-22}$$

This proves the directional Fisher-information gain.

Case 1: $\mathbf{r} \notin \mathrm{range}(\mathcal{J}_v)$. Let

$$\mathcal{S}_v := \mathrm{range}(\mathcal{J}_v). \tag{C4-23}$$

Since $\mathcal{J}_v$ is symmetric,

$$\mathcal{S}_v^\perp = \ker(\mathcal{J}_v). \tag{C4-24}$$

Hence,

$$\mathbf{r} = \mathbf{r}_{\mathcal{S}} + \mathbf{r}_0, \tag{C4-25}$$

where

$$\mathbf{r}_{\mathcal{S}} \in \mathcal{S}_v, \qquad \mathbf{r}_0 \in \ker(\mathcal{J}_v), \qquad \mathbf{r}_0 \neq \mathbf{0}. \tag{C4-26}$$

Since

$$\mathcal{J}_v \mathbf{r}_0 = \mathbf{0}, \tag{C4-27}$$

the visual observations provide no information for constraining this component. Moreover,

$$\mathbf{r}^\top \mathbf{r}_0 = \|\mathbf{r}_0\|_2^2 > 0, \tag{C4-28}$$

so perturbations along $\mathbf{r}_0$ change $\mathbf{r}^\top \boldsymbol{\theta}_c$ while remaining unidentifiable from the visual observations. Thus,

$$\mathbf{r}^\top \operatorname{Var}\left(\widehat{\boldsymbol{\theta}}_c^{\,v}\right) \mathbf{r} = \infty. \tag{C4-29}$$

Under the joint-identifiability condition adopted in the main paper, the corresponding direction is identifiable from the combined visual and textual observations, and therefore

$$\mathbf{r}^\top \operatorname{Var}\left(\widehat{\boldsymbol{\theta}}_c^{\,v,t}\right) \mathbf{r} < \infty. \tag{C4-30}$$

Hence,

$$\mathbf{r}^\top \operatorname{Var}\left(\widehat{\boldsymbol{\theta}}_c^{\,v,t}\right) \mathbf{r} < \mathbf{r}^\top \operatorname{Var}\left(\widehat{\boldsymbol{\theta}}_c^{\,v}\right) \mathbf{r}. \tag{C4-31}$$

This establishes the variance-reduction statement of Theorem 2.

Case 2: $\mathbf{r} \in \operatorname{range}(\mathcal{J}_v)$. For completeness, we further characterize directions that are already identifiable from the visual observations.

Let

$$\mathcal{S}_v := \operatorname{range}(\mathcal{J}_v), \tag{C4-32}$$

and let $\mathbf{P}_v$ denote the orthogonal projector onto $\mathcal{S}_v$. We additionally assume that $\mathcal{S}_v$ is invariant under the joint Fisher information $\mathcal{J}_{v,t}$. Since $\mathcal{J}_v$ already leaves $\mathcal{S}_v$ invariant, this condition is equivalently written as

$$\mathbf{P}_v \mathcal{J}_t \left(\mathbf{I} - \mathbf{P}_v\right) = \mathbf{0}. \tag{C4-33}$$

Because $\mathcal{J}_t$ is symmetric, this also implies

$$\left(\mathbf{I} - \mathbf{P}_v\right) \mathcal{J}_t \mathbf{P}_v = \mathbf{0}. \tag{C4-34}$$

Thus, the visually identifiable subspace and its orthogonal complement are not coupled by the textual Fisher information.

Restricted to $\mathcal{S}_v$, define

$$\widetilde{\mathcal{J}}_v := \mathbf{P}_v \mathcal{J}_v \mathbf{P}_v|_{\mathcal{S}_v}, \tag{C4-35}$$

and

$$\widetilde{\mathcal{J}}_{v,t} := \mathbf{P}_v \mathcal{J}_{v,t} \mathbf{P}_v|_{\mathcal{S}_v}. \tag{C4-36}$$

Since $\mathcal{J}_v$ is positive semidefinite and $\mathcal{S}_v = \operatorname{range}(\mathcal{J}_v)$, the restriction $\widetilde{\mathcal{J}}_v$ is positive definite on $\mathcal{S}_v$.

Let

$$\boldsymbol{\Delta}_v := \mathbf{P}_v \mathcal{J}_t \mathbf{P}_v|_{\mathcal{S}_v} \succeq \mathbf{0}. \tag{C4-37}$$

Then

$$\widetilde{\mathcal{J}}_{v,t} = \widetilde{\mathcal{J}}_v + \boldsymbol{\Delta}_v \succeq \widetilde{\mathcal{J}}_v \succ \mathbf{0}. \tag{C4-38}$$

Under the invariance condition above, the joint Fisher information is block diagonal with respect to $\mathcal{S}_v \oplus \mathcal{S}_v^\perp$. Therefore, for directions in $\mathcal{S}_v$, the corresponding estimator covariance is characterized by the inverse of the restricted Fisher information. Applying Lemma 2 gives

$$\mathbf{r}^\top \widetilde{\mathcal{J}}_{v,t}^{-1} \mathbf{r} \leq \mathbf{r}^\top \widetilde{\mathcal{J}}_v^{-1} \mathbf{r}. \tag{C4-39}$$

Moreover, the inequality is strict whenever

$$\boldsymbol{\Delta}_v \widetilde{\mathcal{J}}_v^{-1} \mathbf{r} \neq \mathbf{0}. \tag{C4-40}$$

Equivalently,

$$\mathbf{r}^\top \mathrm{Var}\left(\widehat{\boldsymbol{\theta}}_c^{v,t}\right)\mathbf{r} < \mathbf{r}^\top \mathrm{Var}\left(\widehat{\boldsymbol{\theta}}_c^{v}\right)\mathbf{r} \tag{C4-41}$$

whenever the textual Fisher information acts nontrivially on the inverse-information-weighted direction $\widetilde{\mathcal{J}}_v^{-1}\mathbf{r}$.

Furthermore, because $\mathbf{r} \in \mathcal{S}_v$, we have $\mathbf{P}_v\mathbf{r} = \mathbf{r}$, and therefore

$$\begin{aligned}\mathbf{r}^\top \boldsymbol{\Delta}_v \mathbf{r} &= \mathbf{r}^\top \mathbf{P}_v \mathcal{J}_t \mathbf{P}_v \mathbf{r} \\ &= \mathbf{r}^\top \mathcal{J}_t \mathbf{r} > 0.\end{aligned} \tag{C4-42}$$

Hence, $\boldsymbol{\Delta}_v \neq \mathbf{0}$. By Lemma 2, there exists at least one nonzero direction $\mathbf{u} \in \mathcal{S}_v$ such that

$$\mathbf{u}^\top \widetilde{\mathcal{J}}_{v,t}^{-1}\mathbf{u} < \mathbf{u}^\top \widetilde{\mathcal{J}}_v^{-1}\mathbf{u}. \tag{C4-43}$$

Thus, within the visually identifiable subspace, incorporating unpaired clinical text cannot increase estimation uncertainty and strictly reduces it whenever the textual information contributes to the corresponding inverse-information-weighted direction. In addition, nonzero textual information on this subspace guarantees strict variance reduction along at least one visually identifiable direction.

F Visual Evidence Generation

To further provide traceable visual evidence, **MIChain** uses the concept representations to generate lesion localization and boundary cues. Specifically, the concept representations are projected into the feature spaces required by the localization and boundary branches, and are then integrated with GroundingDINO (Liu et al., 2024) and PiDiNet (Su et al., 2021), respectively. The following sections provide the implementation details of these two branches.

F.1 Projection to Visual Expert Spaces

For frame t, the visual concept representation is given by $\mathbf{z}_t^{vis} = \{\mathbf{C}_t^i\}_{i=1}^n$, where $\mathbf{C}_t^i$ denotes the embedding of the i-th clinical concept. Since the localization and boundary branches operate in different feature spaces, we introduce two task-specific projection modules:

$$\mathbf{Q}_t^d = \mathcal{P}_d\left(\mathbf{z}_t^{vis}\right), \qquad \mathbf{Q}_t^e = \mathcal{P}_e\left(\mathbf{z}_t^{vis}\right), \tag{C4-44}$$

where $\mathcal{P}_d$ and $\mathcal{P}_e$ denote the projection modules for lesion localization and boundary generation, respectively. Both modules first map the concept embeddings into the feature dimension required by the corresponding visual branch and then aggregate the projected concept features through attention.

F.2 Concept-Guided Lesion Localization

To locate the lesion from each ultrasound frame, we use the projected concept representations $\mathbf{Q}_t^d$ as decoder-side prompts for GroundingDINO (Liu et al., 2024). The original text prompt is retained and encoded into the textual memory $\mathbf{M}^{txt}$. The concept prompts are then appended to the textual memory as:

$$\widetilde{\mathbf{M}}_t^{txt} = \left[\mathbf{M}^{txt}; \mathbf{Q}_t^d\right]. \tag{C4-45}$$

Meanwhile, the ultrasound frame $\mathbf{x}_t$ is encoded by the GroundingDINO visual backbone to obtain the corresponding image features $\mathbf{M}_t^{img}$. The decoder then performs lesion localization using the object queries, image features, and concept-augmented textual memory:

$$\mathbf{H}_t^d = \mathcal{D}_d\left(\mathbf{Q}^{obj}, \mathbf{M}_t^{img}, \widetilde{\mathbf{M}}_t^{txt}\right), \tag{C4-46}$$

where $\mathcal{D}_d$ denotes the GroundingDINO decoder and $\mathbf{Q}^{obj}$ denotes the object queries. The resulting decoder features are further mapped to lesion confidence scores and bounding boxes:

$$\left(\widehat{\mathbf{s}}_t, \widehat{\mathbf{b}}_t\right) = \mathcal{H}_d\left(\mathbf{H}_t^d\right). \tag{C4-47}$$

Here, the concept representations are used as additional decoder-side prompts rather than replacing the original textual memory or object queries. In this way, lesion localization is guided by the same clinical concept representations used for diagnostic reasoning.

Localization Supervision. The localization branch is supervised using the expert-verified lesion bounding boxes described in Appendix A. The detection loss combines focal classification loss, bounding-box regression loss, and generalized IoU loss:

$$\mathcal{L}_{loc} = \lambda_{cls}\mathcal{L}_{focal} + \lambda_{box}\mathcal{L}_1 + \lambda_{giou}\mathcal{L}_{GIoU}. \tag{C4-48}$$

F.3 Concept-Guided Lesion Boundary Generation

To further provide lesion boundary information, we use PiDiNet (Su et al., 2021) to extract edge features from each ultrasound frame. The boundary projection module $\mathcal{P}_e$ transforms the frame-level concept representations into a set of boundary tokens:

$$\mathbf{Q}_t^e = \left\{\mathbf{q}_{t,k}^e\right\}_{k=1}^{K}, \tag{C4-49}$$

where K denotes the number of boundary tokens.

Given the ultrasound frame $\mathbf{x}_t$, PiDiNet extracts the intermediate edge representation:

$$\mathbf{F}_t^e = \mathcal{E}_e\left(\mathbf{x}_t\right) \in \mathbb{R}^{d_e \times H_e \times W_e}, \tag{C4-50}$$

where $\mathcal{E}_e$ denotes the PiDiNet edge encoder. Each boundary token is then used to obtain a spatial edge response from $\mathbf{F}_t^e$:

$$\mathbf{R}_{t,k}(h,w) = \left\langle \mathbf{q}_{t,k}^e, \mathbf{F}_t^e(:,h,w) \right\rangle. \tag{C4-51}$$

The responses of all boundary tokens are aggregated to generate the predicted boundary map:

$$\widehat{\mathbf{E}}_t = \sigma\left(\frac{1}{K}\sum_{k=1}^{K}\mathbf{R}_{t,k}\right), \tag{C4-52}$$

where $\sigma(\cdot)$ denotes the sigmoid function. In this way, the clinical concept representations are associated with the spatial edge features to generate the lesion boundary evidence.

Boundary Supervision. The boundary ground truth is generated from the expert-annotated lesion mask and the corresponding ultrasound frame. Specifically, the lesion mask is first used to extract the lesion contour and determine the local region around the lesion. Local edge responses are then extracted from the ultrasound image within this region and combined with the mask-derived contour to obtain the final boundary target.