Medical Image Analysis 112 (2026) 104127

Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

# AsyCMST: Asymmetric cross-modal spatio-temporal learning for multimodal ultrasound nodule recognition☆

Hongcheng Han [a], Zhiqiang Tian [b], Minghao Wang [a], Yutong Zhang [a], Dong Zhang [a], Qinbo Guo [a], Jue Jiang [a], Hui Guo [c], Shaoyi Du [a,*], Juan Wang [a,*]

[a] Department of Ultrasound, the Second Affiliated Hospital of Xi'an Jiaotong University, and State Key Laboratory of Human–Machine Hybrid Augmented Intelligence, and Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University, Xi'an, 710049, Shaanxi, China
[b] School of Software Engineering, Faculty of Electronic and Information Engineering, Xi'an Jiaotong University, Xi'an, 710049, Shaanxi, China
[c] Department of Medical Oncology, and Phase I Clinical Trial Ward, the Second Affiliated Hospital of Xi'an Jiaotong University, and Key Laboratory of Surgical Critical Care and Life Support, Xi'an Jiaotong University, Xi'an, 710049, Shaanxi, China

ARTICLE INFO

*Keywords:*
Multimodal medical imaging
Contrast-enhanced ultrasound
Asymmetric cross-modal attention
Spatio-temporal modeling
Video understanding

ABSTRACT

Multimodal ultrasound combining B-mode ultrasound (BUS) and contrast-enhanced ultrasound (CEUS) has become a powerful tool for diagnosing superficial nodules in the thyroid and breast, leveraging the complementary strengths of BUS spatial structure and CEUS temporal hemodynamics. However, existing fusion methods typically treat both modalities symmetrically or focus solely on modality-specific features, overlooking the inherent asymmetric bidirectional guidance between BUS spatial context and CEUS perfusion dynamics. To address this limitation, we propose AsyCMST, an asymmetric cross-modal spatio-temporal network for multimodal ultrasound nodule diagnosis. First, we design a multi-task learning module to enhance modality-specific representations, where frame self-sorting distills canonical contrast perfusion patterns in CEUS, while nodule segmentation reinforces precise lesion localization in BUS. Second, we propose an asymmetric cross-modal spatio-temporal attention mechanism to enable clinically meaningful directional interaction: BUS spatial cues guide CEUS temporal modeling toward lesion-relevant regions, and CEUS hemodynamic evolution refines ambiguous structural patterns in BUS. This design effectively captures the asymmetric interdependency between structure and function. Experiments on thyroid and breast datasets demonstrate that AsyCMST significantly outperforms state-of-the-art video understanding and multimodal ultrasound fusion methods in accuracy, $F_1$-score, AUC, and cross-dataset generalization. These results validate the effectiveness of knowledge-driven asymmetric fusion and highlight its potential to advance clinical adoption of multimodal ultrasound analysis.

## 1. Introduction

Ultrasound imaging is a highly effective, non-invasive modality for examining superficial organs such as the thyroid (Shen et al., 2025; Kang et al., 2022) and breast (Shen et al., 2021; Yan et al., 2024), owing to its real-time capability and safety. B-mode ultrasound (BUS), the most widely used ultrasound technique, provides clear visualization of static tissue texture, nodule location, morphology, and boundary characteristics (Lin et al., 2024; Zheng et al., 2020). However, BUS is limited in capturing dynamic blood flow information, which is critical for assessing lesion malignancy and often constrains diagnostic performance in complex cases. With the increasing adoption of multimodal ultrasound (Qian et al., 2021), contrast-enhanced ultrasound (CEUS) has gained prominence for its superior ability to depict soft-tissue hemodynamics and perfusion patterns. By integrating BUS and CEUS, multimodal ultrasound enables a more comprehensive evaluation of nodule properties.

As shown in Fig. 1(a), BUS and CEUS are deeply interdependent. BUS provides high-resolution spatial details, such as nodule size, shape,

☆ This work was supported in part by the National Key Research and Development Program of China under Grant No. 2025ZD0217300, Fundamental and Interdisciplinary Disciplines Breakthrough Plan of the Ministry of Education of China under Grant No. JYB2025XDXM504, the National Natural Science Foundation of China under Grant No. 62501462, Guangdong Major Project of Basic and Applied Basic Research under Grant No. 2023B0303000009, the Fundamental Research Funds for the Central Universities under Grant Nos. xtr062025010 and xzy022024010, Xi'an Science and Technology Plan under Grant No. 24ZDCYJSGG0022HZ, and Joint Funds of the Natural Science Foundation of Tianjin under Grant No. 25JCLMJC00280. The program is available at https://github.com/HongchengHan/AsyCMST.

* Corresponding authors.
*E-mail addresses:* dushaoyi@xjtu.edu.cn (S. Du), wangjuan@xjtu.edu.cn (J. Wang).

https://doi.org/10.1016/j.media.2026.104127
Received 17 December 2025; Received in revised form 10 May 2026; Accepted 10 May 2026
Available online 19 May 2026
1361-8415/© 2026 Elsevier B.V. All rights are reserved, including those for text and data mining, AI training, and similar technologies.

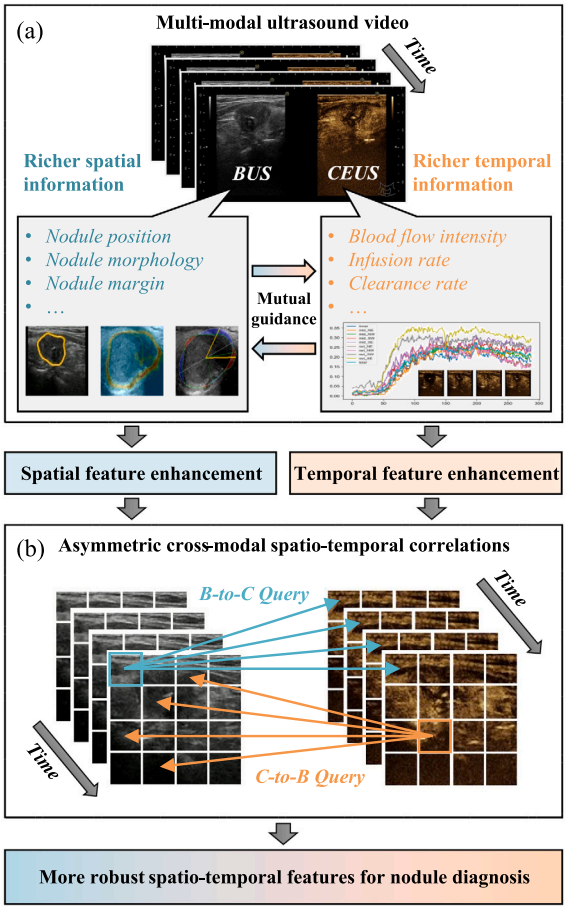


**Fig. 1.** Motivation and concept of the proposed asymmetric cross-modal spatio-temporal correlation mechanism. (a) Spatial and temporal characteristics of BUS and CEUS videos. BUS videos provide richer spatial details, whereas CEUS videos offer more detailed temporal information. (b) Illustration of the proposed cross-modal spatio-temporal correlation mechanism. The B-to-C query associates the feature vector of a B-mode ultrasound (BUS) with feature vectors at corresponding spatial locations in each frame of contrast-enhanced ultrasound (CEUS), using cross-attention to leverage CEUS temporal features for enhanced understanding of BUS structural textures. Conversely, the C-to-B query links a CEUS feature vector with feature vectors across all spatial locations in the corresponding BUS frame, employing cross-attention to utilize BUS structural texture information to guide CEUS temporal feature extraction. This bidirectional association enables complementary modality advantages, yielding robust spatio-temporal features for improved nodule diagnosis.

margins, and echotexture, which are essential for morphological evaluation (Wang et al., 2022). CEUS captures dynamic perfusion patterns, including wash-in/wash-out timing and rim enhancement, critical for assessing vascularity and malignancy (Ruan et al., 2022). Clinically, CEUS interpretation relies on BUS-defined lesion boundaries to localize perfusion analysis, while CEUS clarifies ambiguous BUS features, such as infiltration or cystic components. This bidirectional complementarity enables comprehensive diagnosis but requires advanced cross-modal correlation modeling. Effective fusion needs to align BUS spatial context with CEUS temporal dynamics, allowing structural cues to guide hemodynamic analysis and functional signals to refine morphological understanding. Such synergy motivates the need for robust cross-modal spatio-temporal learning beyond simple feature aggregation.

Recent advances in deep learning have significantly advanced multimodal medical image analysis (Huang et al., 2021; He et al., 2025), particularly in nodule diagnosis (Han et al., 2025; Zhang et al., 2024). Many existing fusion methods employ multi-branch architectures to extract modality-specific features from BUS and CEUS independently, followed by concatenation or summation (Cai et al., 2024; Qu et al., 2025). Such data-driven approaches, however, fail to establish meaningful inter-modal relationships, leading to modality bias, erroneous feature learning, and compromised robustness and generalization. Some methods enhance cross-modal interaction through skip connections or regularization across model stages (Roy et al., 2023; Wu et al., 2025), improving fusion efficiency. Yet, they treat modalities symmetrically and lack task-specific prior guidance, limiting their ability to capture complementary advantages under data scarcity or noise. Although ultrasound-specific methods incorporate domain knowledge, such as key frame selection (Gong et al., 2022), perfusion curves (Chen et al., 2021), or infiltrative region focus (F. Chen et al., 2024), they typically enhance only spatial or temporal features in isolation or establish unidirectional cross-modal links. This results in asymmetric bidirectional dependencies between BUS spatial context and CEUS hemodynamic evolution remaining unmodeled, causing diagnostic failures when one modality is ambiguous.

To address these limitations, we propose AsyCMST, a novel asymmetric cross-modal spatio-temporal learning framework for multimodal ultrasound nodule diagnosis. Inspired by clinical workflows where radiologists jointly interpret BUS spatial structure and CEUS perfusion dynamics, AsyCMST explicitly models their asymmetric yet bidirectional interdependence. BUS provides anatomical context to anchor CEUS temporal analysis, while CEUS resolves structural ambiguities in BUS. Our approach introduces two tightly integrated innovations. First, a multi-task spatio-temporal feature enhancement module simultaneously performs nodule segmentation in BUS to reinforce spatial semantics and frame self-sorting in CEUS to distill perfusion-critical phases, establishing modality-tailored, noise-robust representations. Second, an asymmetric cross-modal spatio-temporal attention mechanism, illustrated in Fig. 1(b), enables directional guidance. BUS spatial features steer CEUS temporal modeling toward lesion-specific hemodynamics, while CEUS dynamics refine subtle or infiltrative patterns in BUS. Unlike symmetric attention, this design respects modality-specific roles, suppresses redundancy, and enhances fusion precision.

By hierarchically aligning structure and function across space and time, AsyCMST achieves superior diagnostic accuracy and clinical interpretability. The main contributions of this work are:

- We propose AsyCMST, an asymmetric cross-modal spatio-temporal network for multimodal ultrasound nodule diagnosis. By designing a novel cross-modal attention mechanism tailored to the characteristics of CEUS and BUS, we enhance inter-modal correlation modeling, improving feature fusion quality and nodule diagnostic accuracy.
- We design a spatio-temporal feature enhancement module based on multi-task learning, utilizing frame self-sorting and nodule segmentation tasks to guide the model in capturing CEUS's hemodynamic patterns and BUS's structural texture information, laying a robust foundation for cross-modal spatio-temporal correlation.
- Experimental results demonstrate that AsyCMST outperforms existing video understanding and multimodal ultrasound analysis methods, providing a reliable and robust solution for nodule recognition through advanced cross-modal spatio-temporal modeling.

## 2. Related work

The task addressed in this work is video-based nodule classification using paired B-mode ultrasound (BUS) and contrast-enhanced ultrasound (CEUS). As such, prior research in multimodal medical data fusion and spatio-temporal modeling for video understanding offers valuable insights. The following subsections briefly review and analyze relevant works from these two perspectives.

### 2.1. Multimodal medical data fusion

Multimodal fusion in medical imaging integrates complementary data sources to improve diagnostic accuracy. Early methods relied on handcrafted features, such as pixel-wise fusion of CT and MRI textures (Zhu et al., 2021) or decision-level ensembles (Du et al., 2016). With deep learning, convolutional neural networks (CNNs) (He et al., 2016) and vision transformers (ViTs) (Dosovitskiy et al., 2021; Liu et al., 2021) enabled automated feature extraction. Cai et al. (2024) proposed a multi-branch CNN for multi-parameter MRI lesion diagnosis, using modality-specific encoders and feature map concatenation. Qin et al. (2019) combined B-mode ultrasound and elastography via early and late fusion networks. Later, Qian et al. (2020) developed a hierarchical model with dual CNNs for B-mode ultrasound and Doppler, followed by MLP-based fusion. While effective in some scenarios, these approaches fail to model inter-modal relationships, leading to modality bias, redundant features, and limited robustness.

To enhance interaction, recent works introduce cross-modal connections. Roy et al. (2023) proposed the multimodal fusion Transformer (MFT), using cross-attention between transformer blocks to align corresponding regions across modalities. Black and Souvenir (2024) introduced a hybrid fusion model with mutual distillation, aligning multi-source information via multi-loss supervision. Wu et al. (2025) developed FC-Former, employing fully-connected self-attention to capture multi-scale cross-modal patterns. These methods better exploit modality advantages but lack prior knowledge guidance and struggle with limited data, particularly in fusing BUS spatial structures and CEUS temporal dynamics.

Ultrasound-specific methods incorporate radiologist expertise. Gong et al. (2022) proposed BUS-Net, selecting key frames from BUS via hard example mining, extracting shape features, and fusing with CEUS video encodings. Chen et al. (2021) introduced domain knowledge-powered learning, using physician-labeled time-intensity curves to guide CEUS hemodynamics and comparing BUS-CEUS nodule contours to assess infiltration. F. Chen et al. (2024) further modeled sonographer reasoning with temporal attention on CEUS perfusion, guided by BUS structure to focus on infiltrative regions. While these approaches leverage clinical knowledge, they insufficiently model bidirectional guidance — BUS informing CEUS and vice versa — leaving fusion incomplete when one modality is ambiguous.

Despite these advances, existing approaches predominantly adopt symmetric fusion paradigms that treat BUS and CEUS as equal contributors. This overlooks the directional dependency, BUS provides spatial context essential for interpreting CEUS perfusion patterns, while CEUS offers functional cues to resolve ambiguous BUS textures. Consequently, incomplete bidirectional guidance leads to suboptimal feature alignment, increased noise sensitivity, and reduced robustness in clinically challenging cases.

### 2.2. Spatio-temporal learning for video understanding

Spatio-temporal modeling is foundational to video understanding (Abdar et al., 2024). Early 3D CNNs, such as C3D (Tran et al., 2015), jointly processed spatial and temporal dimensions but incurred high computational overhead. R(2+1)D (Tran et al., 2018) and SlowFast (Feichtenhofer et al., 2019) improved efficiency by decomposing convolutions into separate spatial and temporal pathways, excelling in action recognition. Transformer-based models further advanced the field, ViViT (Arnab et al., 2021) factorizes space–time attention into independent spatial and temporal self-attention modules, while Video Swin Transformer (Liu et al., 2022) employs shifted window attention for hierarchical video representation. These methods effectively capture long-range dependencies in natural videos with structured motion.

However, they assume homogeneous input and symmetric spatio-temporal dynamics, which poorly align with BUS-CEUS video pairs. ViViT, for instance, applies identical spatial attention across frames, failing to distinguish BUS's static structural consistency from CEUS's evolving perfusion patterns. Its temporal attention treats all time steps equally, unable to prioritize contrast wash-in/wash-out phases critical for malignancy assessment. Moreover, ViViT processes each modality independently before late fusion, missing opportunities to use BUS nodule boundaries to guide CEUS perfusion localization.

Multimodal video frameworks like MFT (Roy et al., 2023) and FC-Former (Wu et al., 2025) introduce cross-attention between streams, but their symmetric designs treat BUS and CEUS equivalently, ignoring that BUS provides spatial anchors for CEUS temporal analysis, while CEUS offers functional context to resolve BUS ambiguities, such as isoechoic lesions. In medical video, spatio-temporal models have been applied to echocardiography (Fermann et al., 2024; Wang et al., 2025) and elastography (Kijanka and Urban, 2024), focusing on cardiac motion or tissue deformation. Ultrasound-specific methods, such as BUS-Net (Gong et al., 2022), DKPD (Chen et al., 2021), and DAST (F. Chen et al., 2024), adapt 3D backbones for nodule diagnosis but rely on late fusion, neglecting asymmetric cross-modal guidance, BUS informing CEUS perfusion interpretation and CEUS enhancing BUS boundary delineation in infiltrative cases.

In summary, modality-agnostic designs fail to address ultrasound-specific challenges: high frame redundancy, subtle hemodynamic motion, and the need for directional alignment between BUS spatial anchors and CEUS temporal evolution. This results in inefficient feature utilization and limited diagnostic sensitivity in complex nodule cases.

**In conclusion:** both fields suffer from symmetric fusion paradigms that overlook the asymmetric informational roles of BUS and CEUS. Multimodal medical fusion neglects bidirectional guidance, while video understanding models ignore ultrasound-specific spatio-temporal heterogeneity. Critically, no prior work jointly enables modality-specific enhancement and directional cross-modal interaction across the full hierarchy. To bridge this gap, we propose AsyCMST, which introduces asymmetric cross-modal spatio-temporal attention to enable BUS-guided CEUS temporal modeling and CEUS-enhanced BUS refinement, supported by a multi-task enhancement module. This design achieves robust, interpretable, and clinically meaningful fusion for multimodal ultrasound nodule diagnosis.

## 3. Methodology

### 3.1. Overall framework

The proposed AsyCMST framework is designed to enhance multimodal ultrasound nodule diagnosis by establishing a knowledge-driven, asymmetric cross-modal interaction between B-mode ultrasound (BUS) and contrast-enhanced ultrasound (CEUS). As depicted in Fig. 2, the architecture integrates modality-specific feature enhancement with directional cross-modal fusion, enabling robust alignment of spatial structure and temporal hemodynamics.

Given synchronized video inputs $V_C, V_B \in \mathbb{R}^{T\times H'\times W'\times C}$, where $T$, $H'$, $W'$ and $C$ represent the number of frames, height, width and the number of channels, two ResNet-18 (He et al., 2016) backbones with $D$ output channels are employed as modality-specific encoders. These networks process each frame independently to extract low-level visual features while downsampling spatial resolution, producing feature sequences $F_C, F_B \in \mathbb{R}^{T\times H\times W\times D}$.

To strengthen intra-modal representation learning, a multi-task spatio-temporal feature enhancement module is introduced, as shown in Fig. 2(b). For the CEUS stream, a frame self-sorting (FSS) task supervises the model to recognize canonical contrast perfusion dynamics, such as agent inflow and clearance, thereby attenuating noise from irrelevant fluctuations. This is optimized via loss term $\mathcal{L}_{\text{FSS}}$. In parallel, the BUS stream undergoes nodule segmentation (NS) supervision to precisely delineate lesion boundaries, reducing interference from surrounding anatomical structures including vessels and

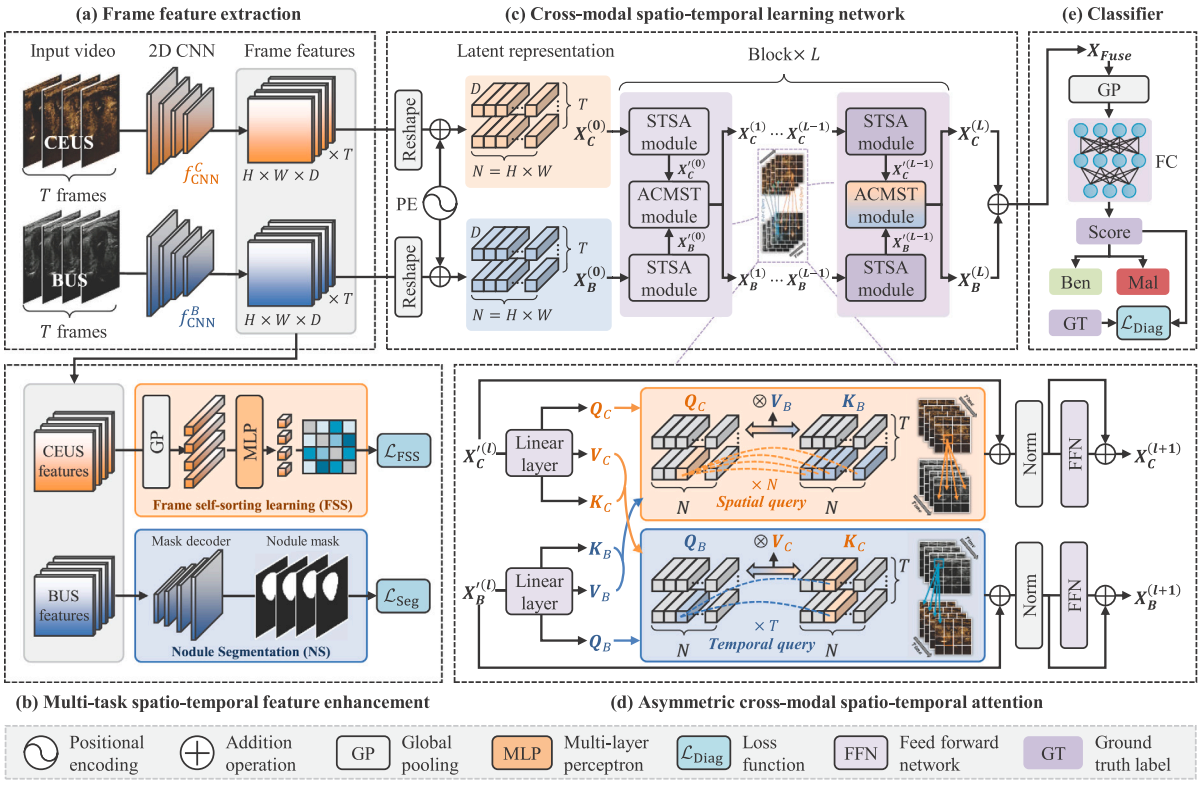


**Fig. 2.** Overall framework of the proposed AsyCMST. (a) Frame feature extraction module. CEUS and BUS video frames ($T$ frames each) are processed by two 2D ResNet18-based encoders to extract per-frame features. (b) Spatio-temporal feature enhancement module. CEUS features are enhanced via frame self-sorting to capture hemodynamic temporal patterns, BUS features are supervised by nodule segmentation to learn structural and texture details. (c) Cross-modal spatio-temporal correlation network. Flattened 2D CNN features ($T \times N \times D$) with positional encoding are processed through blocks of spatio-temporal self-attention and cross-modal spatio-temporal attention to produce fused CEUS and BUS features. (d) Asymmetric cross-modal spatio-temporal attention module. BUS tokens query all temporal frames at the same spatial location in CEUS, enabling perfusion dynamics to guide structural representation. Conversely, CEUS tokens query all spatial tokens in the corresponding BUS frame, allowing anatomical context to steer temporal perfusion modeling. This design establishes clinically aligned cross-modal associations while reducing redundant computation. (e) Classifier. A fully connected layer performs benign-malignant nodule classification.

parenchyma, guided by $\mathcal{L}_{\text{Seg}}$. These auxiliary tasks collectively refine modality-specific feature quality prior to cross-modal integration.

Following enhancement, frame-level features are spatially flattened and augmented with learnable absolute positional encodings, yielding latent sequences $X_C^{(0)}, X_B^{(0)} \in \mathbb{R}^{T\times N\times D}$, where $N = H \times W$. These are processed by a dual-branch transformer backbone comprising $L$ stages. Within each stage, intra-modal spatio-temporal correlations are modeled using spatio-temporal self-attention (STSA) modules adapted from ViViT (Arnab et al., 2021).

Cross-modal fusion is achieved through an asymmetric cross-modal spatio-temporal attention (ACMST) module, illustrated in Fig. 2(d). The ACMST employs asymmetric query-key interactions to enforce clinically informed guidance: BUS-derived spatial context directs CEUS temporal modeling toward lesion-relevant perfusion, while CEUS hemodynamic evolution refines ambiguous or infiltrative patterns in BUS. This directional mechanism contrasts with symmetric attention, better capturing modality-specific roles and mitigating information redundancy.

At the final stage, features $X_C^{(L)}$ and $X_B^{(L)}$ are aggregated via element-wise summation and performed global average pooling. Then, it is fed into a fully connected classifier, as shown in Fig. 2(e). Classification is supervised using cross-entropy loss $\mathcal{L}_{\text{Diag}}$.

By hierarchically integrating enhanced modality-specific representations with asymmetric cross-modal spatio-temporal attention, AsyCMST establishes a robust information fusion pathway that aligns anatomical structure with functional perfusion, significantly improving diagnostic accuracy and clinical interpretability in multimodal ultrasound analysis.

### 3.2. Spatio-temporal feature enhancement based on multi-task learning

To establish robust and clinically meaningful modality-specific representations prior to cross-modal fusion, we introduce a spatio-temporal feature enhancement module grounded in multi-task learning. This module leverages radiological prior knowledge to suppress noise and strengthen discriminative feature learning in both B-mode ultrasound (BUS) and contrast-enhanced ultrasound (CEUS) streams. By aligning model behavior with established diagnostic principles, it provides a solid foundation for subsequent asymmetric cross-modal interaction.

In CEUS, contrast agent dynamics follow a canonical temporal pattern: rapid enhancement during arterial inflow, peak intensity, and gradual washout. Capturing this progression is essential for distinguishing malignant from benign perfusion behaviors. In BUS, accurate nodule localization is critical for isolating lesion-specific structural features from surrounding anatomy, including vessels, parenchyma, and artifacts. To embed these priors, we design two complementary auxiliary tasks: frame self-sorting (FSS) for CEUS and nodule segmentation (NS) for BUS.

For the FSS in the CEUS stream, as Fig. 3 shows, given CEUS feature maps $F_C \in \mathbb{R}^{T\times H\times W\times D}$, we extract per-frame representations via a hybrid global pooling operation. Each frame $F_C^t \in \mathbb{R}^{H\times W\times D}$ is processed as:

$$v_t = \frac{1}{2} f_{\text{GAP}}(F_C^t) + \frac{1}{2} f_{\text{GMP}}(F_C^t), \tag{1}$$

where $f_{\text{GAP}}$ and $f_{\text{GMP}}$ denote global average and max pooling, respectively. The resulting sequence $\{v_t\}_{t=1}^{T} \in \mathbb{R}^D$ preserves both dominant

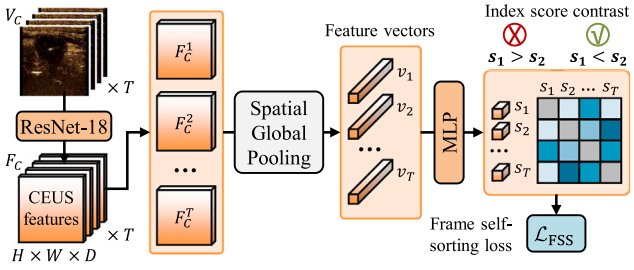


**Fig. 3.** Illustration of the frame self-sorting learning in the multi-task spatio-temporal feature enhancement module.

**Algorithm 1** Frame self-sorting for CEUS temporal feature enhancement

**Require:** CEUS feature maps $F_C \in \mathbb{R}^{T\times H\times W\times D}$, MLP $f_{\text{MLP}}$, soft clamping threshold $\varphi$
**Ensure:** Temporal ranking loss $\mathcal{L}_{\text{FSS}}$
Per-frame features: $F_C^t \leftarrow X_C[t,:,:,:]$ for $t = 1$ to $T$
Hybrid pooling: $v_t \leftarrow \frac{1}{2} f_{\text{GAP}}(F_C^t) + \frac{1}{2} f_{\text{GMP}}(F_C^t)$
Predict temporal index: $s_t \leftarrow f_{\text{MLP}}(v_t)$
Initialize loss: $\mathcal{L}_{\text{FSS}} \leftarrow 0$
**for** $t = 1$ to $T-1$ **do**
  **for** $i = t+1$ to $T$ **do**
    $f_{\text{Sort}}(s_t, s_i) \leftarrow -\log(\min(\sigma(s_i - s_t), \varphi))$
    $\mathcal{L}_{\text{FSS}} \leftarrow \mathcal{L}_{\text{FSS}} + f_{\text{Sort}}(s_t, s_i)$
  **end for**
**end for**
Normalize: $\mathcal{L}_{\text{FSS}} \leftarrow \mathcal{L}_{\text{FSS}} / \frac{1}{2}(T \cdot (T-1))$
**return** $\mathcal{L}_{\text{FSS}}$

and salient activation patterns across frames. A two-layer multi-layer perception $f_{\text{MLP}}$ then predicts a scalar temporal index score:

$$s_t = f_{\text{MLP}}(v_t) \in \mathbb{R}. \tag{2}$$

Instead of regressing absolute frame indices, we adopt a pairwise ranking objective to learn relative temporal ordering. This approach is robust to sequence-level shifts such as delayed contrast arrival and enforces global consistency across multiple frame comparisons, ensuring the model captures the canonical perfusion progression despite local noise. The frame self-sorting loss is defined as:

$$\mathcal{L}_{\text{FSS}} = \frac{2}{T\cdot(T-1)} \sum_{t=1}^{T-1} \left( \sum_{i=t+1}^{T} f_{\text{Sort}}(s_t, s_i) \right), \tag{3}$$

$$f_{\text{Sort}}(s_t, s_i) = -\log\left( \min\left( \frac{1}{1+\exp(s_t - s_i)}, \varphi \right) \right), \tag{4}$$

where $f_{\text{Sort}}(s_t, s_i)$ applies a sigmoid with soft clamping thresholds $\varphi = 0.9$. This prevents outlier frames, caused by motion, gain changes, or artifacts, from inducing conflicting ranking signals and unstable gradients. By limiting overconfidence in ambiguous pairs, it stabilizes training while preserving the dominant physiological perfusion trend such as wash-in/wash-out, ensuring robust and clinically coherent temporal learning. The full FSS procedure is outlined in Algorithm 1, where $\sigma$ represents the sigmoid function.

For the BUS stream, frames with annotated nodule masks are processed by a lightweight mask decoder (J. Chen et al., 2024). The decoder upsamples spatial features using transposed convolutions. Since the goal of segmentation is to guide the model in learning spatial information within BUS, skip connections from the encoder to the decoder are omitted. Segmentation is supervised using the Dice loss:

$$\mathcal{L}_{\text{Seg}} = 1 - \frac{2\sum p_i g_i + \varepsilon}{\sum p_i + \sum g_i + \varepsilon}, \tag{5}$$

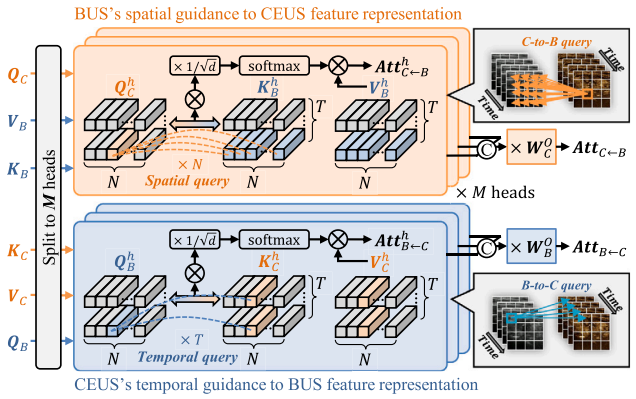


**Fig. 4.** Illustration of the asymmetric cross-modal spatio-temporal attention between CEUS and BUS feature maps.

where $p_i$ and $g_i$ denote the predicted probabilities and ground-truth labels, respectively, and $\varepsilon = 10^{-6}$ is a normalization coefficient introduced to prevent division by zero. This task enforces precise lesion localization, filtering out confounding background structures.

Through joint optimization of $\mathcal{L}_{\text{FSS}}$ and $\mathcal{L}_{\text{Seg}}$, the module significantly enhances temporal coherence in CEUS and spatial specificity in BUS. These refined representations serve as high-quality inputs to the subsequent asymmetric cross-modal attention mechanism, enabling more effective and clinically aligned information fusion across modalities.

### 3.3. Asymmetric cross-modal spatio-temporal attention

The fusion of CEUS and BUS in nodule diagnosis relies on their complementary information: CEUS captures rich temporal hemodynamic patterns through contrast agent dynamics, while BUS provides clear spatial structural context. These modalities are inherently interdependent. Spatial variations in CEUS perfusion intensity can guide the localization of subtle nodules in BUS that are isoechoic with surrounding tissue. Conversely, well-defined anatomical boundaries in BUS help focus CEUS analysis on lesion-relevant regions, filtering out irrelevant vascular signals. This bidirectional guidance is asymmetric in nature: spatial context from BUS anchors temporal interpretation in CEUS, while functional perfusion from CEUS refines structural ambiguity in BUS. Symmetric fusion mechanisms fail to capture this directional dependency, limiting diagnostic precision.

To address this, we propose the asymmetric cross-modal spatio-temporal attention (ACMST) module, which establishes clinically meaningful, directional interactions between modalities. As Fig. 4 shows, given input features from layer $l$, $X_C^{(l)}, X_B^{(l)} \in \mathbb{R}^{T\times N\times D}$, each stream first passes through its respective spatio-temporal self-attention module to produce refined intra-modal representations $X_C^{\prime(l)}, X_B^{\prime(l)}$. Linear projections then generate query, key, and value tensors:

$$[Q_C, K_C, V_C] = f_{\text{Linear}}^{C}(X_C^{\prime(l)}), \tag{6}$$

$$[Q_B, K_B, V_B] = f_{\text{Linear}}^{B}(X_B^{\prime(l)}), \tag{7}$$

where $f_{\text{Linear}}^{C}$ and $f_{\text{Linear}}^{B}$ are linear layers, all with dimension $D$. Then, they are split into $M$ tensors along the depth dimension for multi-head attention to extract the multiple correlation relationships between BUS and CEUS, where $M$ is the number of heads, and each tensor naturally has a depth dimension of $d = D/M$. The core innovation lies in the asymmetric attention computation. To enable CEUS temporal perfusion to guide BUS structural representation, which is realized through BUS-to-CEUS query, in the $h$-th attention head, the token at time $t$ and

**Algorithm 2** Asymmetric Cross-Modal Spatio-Temporal Attention (ACMST)

**Require:** $X_C^{'(l)}, X_B^{'(l)} \in \mathbb{R}^{T\times N\times D}$, number of attention heads $M$, linear projections $f^C_{\text{Linear}}, f^B_{\text{Linear}}$, output projections $W_C^O$ and $W_B^O$
**Ensure:** $X_C^{(l+1)}, X_B^{(l+1)}$

Compute Q, K and V: $[Q_C, K_C, V_C] \leftarrow f^C_{\text{Linear}}(X_C^{'(l)})$, $[Q_B, K_B, V_B] \leftarrow f^B_{\text{Linear}}(X_B^{'(l)})$
Split $Q_C, K_C, V_C, Q_B, K_B, V_B$ to $M$ heads along depth dimension, gets $Q_C^h, K_C^h, V_C^h, Q_B^h, K_B^h, V_B^h$, $h \in [1, M]$, $d = D/M$
# $\text{Att}_{B\leftarrow C}$: Temporal perfusion guides spatial structure
**for** each $h \in [1, M], t \in [1, T], n \in [1, N]$ **do**
    $\text{score} \leftarrow Q_B^h[t, n] \cdot (K_C^h[:, n])^\top / \sqrt{d}$
    $\text{Att}^h_{B\leftarrow C}(t, n) \leftarrow \text{softmax}(\text{score}) \cdot V_C^h[t, n]$
**end for**
$\text{Att}_{B\leftarrow C} \leftarrow \text{Concat}(\{\text{Att}^h_{B\leftarrow C} | h \in [1, M]\}) \cdot W_B^O$
# $\text{Att}_{C\leftarrow B}$: Spatial context anchors temporal dynamics
**for** each $h \in [1, M], t \in [1, T], n \in [1, N]$ **do**
    $\text{score} \leftarrow Q_C^h[t, n] \cdot (K_B^h[t, :])^\top / \sqrt{d}$
    $\text{Att}^h_{C\leftarrow B}(t, n) \leftarrow \text{softmax}(\text{score}) \cdot V_B^h[t, n]$
**end for**
$\text{Att}_{C\leftarrow B} \leftarrow \text{Concat}(\{\text{Att}^h_{C\leftarrow B} | h \in [1, M]\}) \cdot W_C^O$
# Update outputs:
$X_C^{(l+1)} \leftarrow \text{FFN}(\text{LN}(X_C^{'(l)} + \text{Att}_{C\leftarrow B}))$
$X_B^{(l+1)} \leftarrow \text{FFN}(\text{LN}(X_B^{'(l)} + \text{Att}_{B\leftarrow C}))$
**return** $X_C^{(l+1)}, X_B^{(l+1)}$

spatial position $n$ in BUS uses its query $Q_B^h[t, n]$ to attend over all temporal frames at the same spatial location in CEUS:

$$\text{Att}^h_{B\leftarrow C}(t, n) = \text{softmax}\left(\frac{Q_B^h[t, n] \cdot K_C^h[:, n]^\top}{\sqrt{d}}\right) \cdot V_C^h[:, n], \tag{8}$$

$$\text{Att}_{B\leftarrow C} = \text{Concat}(\{\text{Att}^h_{B\leftarrow C} | h \in [1, M]\}) \cdot W_B^O, \tag{9}$$

where $K_C^h[:, n] \in \mathbb{R}^{T\times d}$ denotes the key vectors of all $T$ frames at spatial location $n$ in CEUS, Concat means concatenation along depth dimension and $W_B^O$ is the output projection matrix to align the feature spaces of multiple heads. This allows the full temporal perfusion profile at position $n$ in CEUS to inform the structural representation in BUS, enhancing detection of infiltrative or functionally active lesions.

Conversely, to enable BUS spatial context to guide CEUS temporal modeling, which is realized by CEUS-to-BUS query, a token in CEUS at time $t$ and position $n$ uses $Q_C^h[t, n]$ to attend over all spatial tokens within the same frame in BUS:

$$\text{Att}^h_{C\leftarrow B}(t, n) = \text{softmax}\left(\frac{Q_C^h[t, n] \cdot K_B^h[t, :]^\top}{\sqrt{d}}\right) \cdot V_B^h[t, :], \tag{10}$$

$$\text{Att}_{C\leftarrow B} = \text{Concat}(\{\text{Att}^h_{C\leftarrow B} | h \in [1, M]\}) \cdot W_C^O, \tag{11}$$

where $K_B^h[t, :] \in \mathbb{R}^{N\times d}$ gathers key vectors of all $N$ spatial positions in frame $t$ of BUS, $W_C^O$ is the output projection matrix. This enables BUS spatial context to steer CEUS temporal modeling toward anatomically relevant regions, suppressing noise from extraneous vasculature.

The attended features are combined with residuals, normalized (LN), and passed through feed-forward networks (FFN):

$$X_C^{(l+1)} = \text{FFN}(\text{LN}(X_C^{'(l)} + \text{Att}_{C\leftarrow B})), \tag{12}$$

$$X_B^{(l+1)} = \text{FFN}(\text{LN}(X_B^{'(l)} + \text{Att}_{B\leftarrow C})). \tag{13}$$

The full ACMST procedure is detailed in Algorithm 2.

By explicitly modeling directional, knowledge-guided cross-modal interactions, the ACMST module significantly enhances multimodal semantic alignment, reduces irrelevant signal interference, and improves diagnostic accuracy in complex nodule cases.

### 3.4. Optimization objective

The AsyCMST framework is trained under a composite loss function that jointly optimizes nodule classification, temporal modeling, and spatial localization, ensuring robust multimodal feature learning. The overall objective is defined as:

$$\mathcal{L}_{\text{Overall}} = \mathcal{L}_{\text{Diag}} + w_1\mathcal{L}_{\text{FSS}} + w_2\mathcal{L}_{\text{Seg}}, \tag{14}$$

where $\mathcal{L}_{\text{Diag}}$ supervises the primary diagnostic task, $\mathcal{L}_{\text{FSS}}$ enforces temporal coherence in CEUS via frame self-sorting, and $\mathcal{L}_{\text{Seg}}$ promotes precise nodule delineation in BUS. The weights $w_1 = 0.4$ and $w_2 = 0.4$ are determined through ablation studies to balance auxiliary task contributions.

The classification head operates on the fused representation $X_{\text{fused}} \in \mathbb{R}^{T\times N\times D}$, obtained by summing the final CEUS and BUS feature maps $X_C^{(L)}$ and $X_B^{(L)}$ followed by global average pooling (GAP) across spatio-temporal dimensions and fully connected network classifier (FC). The predicted probability is:

$$\hat{y} = \text{softmax}\left(f_{\text{FC}}(f_{\text{GAP}}(X_C^{(L)} + X_B^{(L)}))\right), \tag{15}$$

and the diagnosis loss is the binary cross-entropy:

$$\mathcal{L}_{\text{Diag}} = -y \log \hat{y} - (1 - y) \log(1 - \hat{y}), \tag{16}$$

with $y \in \{0, 1\}$ as the ground-truth.

The auxiliary losses $\mathcal{L}_{\text{FSS}}$ and $\mathcal{L}_{\text{Seg}}$, detailed in the spatio-temporal feature enhancement module, guide modality-specific representation learning. $\mathcal{L}_{\text{FSS}}$ encourages the model to capture canonical contrast perfusion dynamics, while $\mathcal{L}_{\text{Seg}}$ enforces accurate lesion boundary localization. Although the ACMST module introduces directional cross-modal dependencies, no explicit inter-modal loss is required; its learning emerges implicitly through gradient flow from $\mathcal{L}_{\text{Diag}}$, reinforced by the enhanced inputs from the multi-task spatio-temporal feature enhancement module.

This unified optimization strategy ensures that spatio-temporal priors are effectively embedded into modality-specific features, which are then dynamically aligned via asymmetric attention. The balanced composite loss enables AsyCMST to achieve clinically interpretable and diagnostically accurate multimodal fusion.

## 4. Results

### 4.1. Experimental setup

#### 4.1.1. Data preparation

To evaluate the performance of nodule diagnosis using BUS and CEUS, a thyroid dataset and a breast dataset are collected from the Second Affiliated Hospital of Xi'an Jiaotong University. The thyroid dataset comprises 2910 samples collected from 2910 unique patients between 2014 and 2021, and the breast dataset includes 1246 samples collected from 1246 unique patients between 2022 and 2024. Specifically, each patient contributed exactly one sample to the dataset. Each sample consists of paired CEUS and BUS videos of a nodule, with benign and malignant labels confirmed by pathological examination and nodule boundaries manually delineated by four experienced radiologists (all have over five years of experience, and two of them have over ten years). Specifically, the two radiologists with over five years of experience have each diagnosed more than 20,000 patients with thyroid and breast diseases, while the two with over ten years of experience have each diagnosed more than 30,000 patients. The attributes of the patients and the corresponding nodules is presented in Table 1. The age distribution of the patients is illustrated in Fig. 5. Benign nodules encompass representative conditions such as fibroadenoma, fibrocystic disease, and chronic granulomatous inflammation. Malignant nodules include representative cancers such as carcinoma in situ, malignant phyllodes tumor, and invasive carcinoma.

**Table 1**
The number of samples in the collected dataset. F and M stand for female and male, respectively. Ben. indicates benign sample and Mal. means malignant sample.

| Dataset | Sex | | Age | Number of samples | | |
|---|---|---|---|---|---|---|
| | F | M | range | Ben. | Mal. | Overall |
| Thyroid | 2304 | 606 | 10–83 | 1606 | 1304 | 2910 |
| Breast | 1233 | 13 | 15–88 | 540 | 706 | 1246 |
| Total | 2537 | 619 | 10–88 | 2146 | 2010 | 4156 |

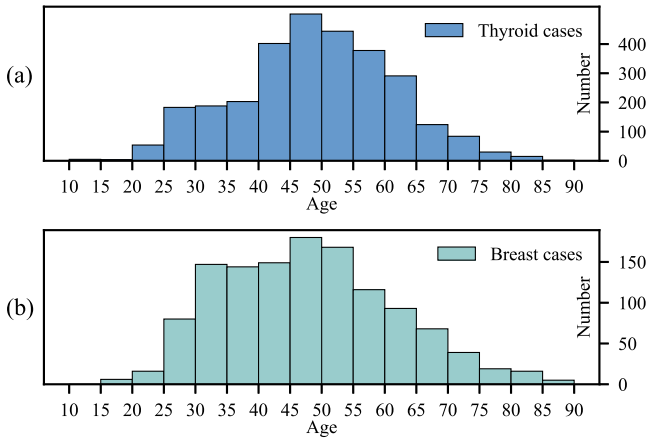


**Fig. 5.** Age distribution of the patients in the collected datasets. (a) Age distribution of the thyroid patients. (b) Age distribution of the breast patients. The age distributions align well with the epidemiological patterns of thyroid and breast diseases, supporting the rationality and representativeness of the dataset construction (Li et al., 2024; Priyadarshini and Panda, 2024).

The study received ethical approval from the medical ethics committee of the Second Affiliated Hospital of Xi'an Jiaotong University on Dec. 8th 2022 (Approval Number: 2022259). Written informed consent was obtained from all participants before the study. Given that CEUS examinations are still being promoted in clinical practice, CEUS samples remain relatively scarce. To the best of our knowledge, no large-scale public dataset of paired BUS and CEUS videos for nodule diagnosis is currently available. Therefore, the datasets employed in this study represent a large collection of CEUS and BUS videos accessible for nodule diagnosis and surpass the scale of the dataset used by Chen et al. (2021), which contains 221 breast lesions.

#### 4.1.2. Implementation details

The proposed method is implemented using PyTorch 2.4.0 on a system running Ubuntu 22.04, equipped with four RTX 3090 GPUs (each with 24 GB memory). Following preprocessing, the input BUS and CEUS videos are resized to 16 frames × 224 × 224. The backbone and the ACMST modules in the cross-modal spatio-temporal correlation network utilizes the architecture with an embedding dimension of $D = 384$, a depth of $L = 6$ and a number of heads $M = 6$. The batch size is set to 32. The RMSProp optimizer is employed for training, with a weight decay of 0.0001 and momentum of 0.9. The initial learning rate is set to 0.001 and is gradually decreased as the number of steps progresses. Data augmentation techniques are applied to the input BUS and CEUS videos during training, including random rotation, random cropping, vertical and horizontal flipping, brightness adjustment, and contrast transformation.

The dataset was randomly partitioned into training, validation, and test sets with a ratio of 5:2:3. To prevent any information leakage, the model was trained exclusively on the training set, hyperparameters were tuned according to performance on the validation set, and final evaluation was performed on the held-out test set. Each model was trained for approximately 80–100 epochs using early stopping based on validation loss convergence. To reduce the impact of a particular data split on performance assessment and to obtain a more robust estimate of model generalization, we repeated the entire experimental process using three independent random splits. Importantly, the three test sets were designed to be mutually disjoint, ensuring that every sample appears in exactly one test set across the three folds. This setup provides a reliable evaluation while controlling for partition-induced variance. All reported performance metrics are therefore presented as mean and standard deviation (mean ± std) computed over the three independent test sets.

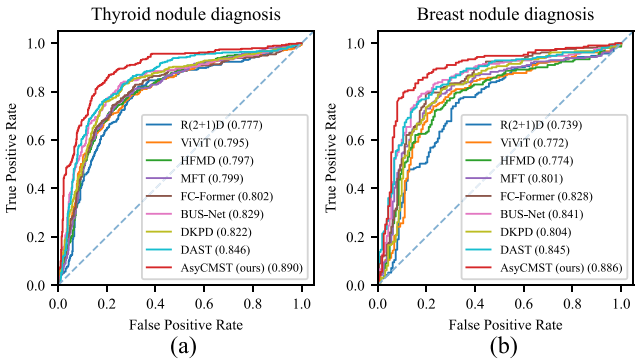


**Fig. 6.** ROC curves of nodule diagnosis of different methods. (a) Results on the thyroid dataset. (b) Results on the breast dataset. The value in parentheses in the legend represents the area under the curve (AUC), higher is better. AsyCMST achieves higher AUC values on both the thyroid dataset and the breast dataset compared to the comparison methods.

#### 4.1.3. Evaluation metrics

Nodule diagnosis performance is comprehensively evaluated using accuracy (*Acc*), recall (*Rec*), precision (*Pre*), $F_1$-score, and area under the receiver operating characteristic curve (AUC).

Accuracy reflects overall correctness, recall measures sensitivity to malignant lesions, precision indicates reliability of positive predictions, and $F_1$-score harmonically balances precision and recall. AUC quantifies discriminative power across all decision thresholds and is threshold-independent. ROC curves are plotted to visualize trade-offs between sensitivity and specificity. All metrics and curves are computed using the scikit-learn Python package.

#### 4.1.4. Competing methods

The task addressed in this study involves multimodal video fusion classification for nodule diagnosis. For comparison with the proposed method, several representative methods in video understanding and multi-source data fusion are selected and categorized into distinct groups. First, prominent video understanding methods, namely R(2+1)D (Tran et al., 2018) and ViViT (Arnab et al., 2021), are adapted into dual-branch structures for nodule diagnosis. Late fusion of BUS and CEUS is achieved through feature concatenation in these models. Second, hybrid fusion methods, including HFMD (Black and Souvenir, 2024), MFT (Roy et al., 2023), and FC-Former (Wu et al., 2025), are chosen for evaluation. Fusion interactions are performed at various model stages in these methods to enhance inter-modal correlations. Additionally, ultrasound-tailored methods, such as BUS-Net (Gong et al., 2022), DKPD (Chen et al., 2021), and DAST (F. Chen et al., 2024), are included as key competing methods, as they are specifically designed for combined CEUS and BUS nodule diagnosis, offering high task specificity. For methods originally developed for image fusion, adaptation to the video-based task is accomplished by extending the input dimensions of their image encoders to accommodate video inputs.

**Table 2**
Performance of various methods for thyroid and breast nodule diagnosis. The best result in each column is shown in bold, and the second-best results are underlined.

| Method | Nodule diagnosis performance (%) | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | Thyroid dataset | | | | Breast dataset | | | |
| | *Acc* ↑ | *Rec* ↑ | *Pre* ↑ | $F_1$ ↑ | *Acc* ↑ | *Rec* ↑ | *Pre* ↑ | $F_1$ ↑ |
| R(2+1)D (Tran et al., 2018) | 72.0 ± 1.3 | 71.6 ± 1.8 | 67.9 ± 1.6 | 69.7 ± 1.3 | 70.8 ± 2.2 | 70.4 ± 4.2 | 76.2 ± 1.0 | 73.1 ± 2.7 |
| ViViT (Arnab et al., 2021) | 75.2 ± 1.1 | 74.6 ± 1.3 | 71.4 ± 1.3 | 73.0 ± 1.2 | 75.4 ± 3.4 | 76.9 ± 4.1 | 79.0 ± 2.7 | 77.9 ± 3.3 |
| HFMD (Black and Souvenir, 2024) | 72.3 ± 1.7 | 72.5 ± 1.8 | 68.0 ± 2.1 | 70.1 ± 1.7 | 72.4 ± 2.0 | 72.9 ± 2.0 | 77.1 ± 2.0 | 75.0 ± 1.8 |
| MFT (Roy et al., 2023) | 76.3 ± 1.3 | 76.0 ± 1.4 | 72.5 ± 1.5 | 74.2 ± 1.4 | 76.1 ± 1.4 | 75.2 ± 1.5 | 81.3 ± 1.5 | 78.1 ± 1.3 |
| FC-Former (Wu et al., 2025) | 74.1 ± 0.9 | 74.5 ± 1.6 | 69.8 ± 1.6 | 72.1 ± 0.8 | 72.8 ± 3.4 | 71.8 ± 3.6 | 78.2 ± 3.1 | 74.9 ± 3.2 |
| BUS-Net (Gong et al., 2022) | 79.1 ± 1.0 | 78.7 ± 0.9 | 75.7 ± 1.6 | 77.2 ± 0.9 | 78.7 ± 2.0 | 79.4 ± 2.1 | 82.4 ± 2.2 | 80.8 ± 1.8 |
| DKPD (Chen et al., 2021) | 78.5 ± 0.7 | 77.7 ± 1.2 | 75.1 ± 0.8 | 76.4 ± 0.8 | 78.2 ± 2.3 | 78.8 ± 2.6 | 82.0 ± 2.5 | 80.3 ± 2.1 |
| DAST (F. Chen et al., 2024) | 77.8 ± 1.2 | 78.2 ± 1.4 | 73.9 ± 1.4 | 75.9 ± 1.3 | 77.0 ± 1.5 | 76.7 ± 2.1 | 81.6 ± 2.4 | 79.0 ± 1.2 |
| AsyCMST (ours) | **82.0 ± 1.4** | **83.1 ± 1.9** | **78.2 ± 1.3** | **80.6 ± 1.6** | **81.8 ± 1.7** | **80.9 ± 2.8** | **86.1 ± 1.6** | **83.4 ± 1.7** |

**Table 3**
Cross-dataset performance comparison of various methods for thyroid and breast nodule diagnosis. The best result in each column is shown in bold, and the second-best results are underlined.

| Method | Nodule diagnosis performance (%) | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | Breast→Thyroid[a] | | | | Thyroid→Breast[a] | | | |
| | *Acc* ↑ | *Rec* ↑ | *Pre* ↑ | $F_1$ ↑ | *Acc* ↑ | *Rec* ↑ | *Pre* ↑ | $F_1$ ↑ |
| R(2+1)D (Tran et al., 2018) | 69.0 ± 1.4 | 68.9 ± 1.1 | 64.5 ± 1.6 | 66.6 ± 1.3 | 70.9 ± 3.6 | 71.7 ± 5.5 | 75.4 ± 2.2 | 73.5 ± 3.9 |
| ViViT (Arnab et al., 2021) | 70.2 ± 0.1 | 68.9 ± 1.6 | 66.1 ± 0.6 | 67.4 ± 0.4 | 70.7 ± 2.4 | 71.9 ± 3.7 | 75.2 ± 2.3 | 73.5 ± 2.5 |
| HFMD (Black and Souvenir, 2024) | 70.9 ± 1.8 | 71.2 ± 1.2 | 66.4 ± 2.2 | 68.7 ± 1.7 | 70.3 ± 0.7 | 71.6 ± 1.0 | 74.9 ± 1.2 | 73.2 ± 0.5 |
| MFT (Roy et al., 2023) | 72.7 ± 2.5 | 74.3 ± 2.6 | 67.9 ± 2.7 | 71.0 ± 2.6 | 70.9 ± 1.2 | 71.2 ± 1.8 | 75.8 ± 0.9 | 73.4 ± 1.3 |
| FC-Former (Wu et al., 2025) | 71.0 ± 1.3 | 70.7 ± 2.9 | 66.7 ± 1.6 | 68.6 ± 1.7 | 71.2 ± 0.9 | 71.1 ± 1.4 | 76.4 ± 1.1 | 73.6 ± 0.9 |
| BUS-Net (Gong et al., 2022) | 75.2 ± 1.5 | 75.1 ± 0.7 | 71.2 ± 2.0 | 73.1 ± 1.4 | 73.5 ± 0.6 | 72.7 ± 2.0 | 78.8 ± 0.7 | 75.6 ± 0.9 |
| DKPD (Chen et al., 2021) | 74.8 ± 0.8 | 75.1 ± 2.6 | 70.6 ± 0.2 | 72.8 ± 1.3 | 73.0 ± 1.1 | 73.0 ± 3.1 | 77.9 ± 0.9 | 75.3 ± 1.5 |
| DAST (F. Chen et al., 2024) | 72.9 ± 1.4 | 72.0 ± 3.2 | 69.0 ± 1.1 | 70.4 ± 1.9 | 72.9 ± 1.4 | 73.8 ± 2.9 | 77.3 ± 0.6 | 75.5 ± 1.6 |
| AsyCMST (ours) | **77.7 ± 1.3** | **77.8 ± 1.5** | **73.8 ± 1.5** | **75.8 ± 1.4** | **79.9 ± 0.2** | **79.0 ± 1.2** | **84.5 ± 0.9** | **81.6 ± 0.3** |

[a] A → B denotes training on the training set of dataset A and evaluating on the testing set of dataset B.

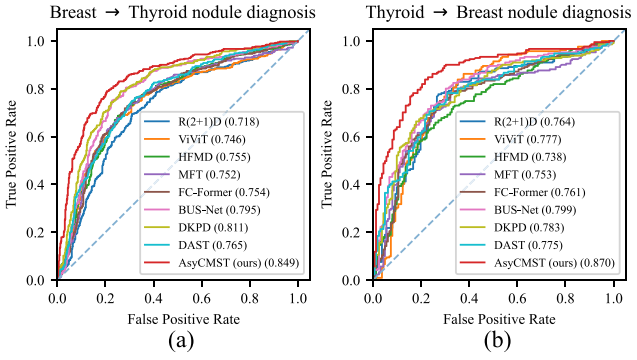


**Fig. 7.** ROC curves for nodule diagnosis in cross-dataset evaluations. (a) Performance on the Breast-to-Thyroid transfer task. (b) Performance on the Thyroid-to-Breast transfer task. The value in parentheses in the legend represents the area under the curve (AUC), higher is better. AsyCMST attains higher AUC values in both cross-dataset diagnosis tasks when compared to the comparison methods.

### 4.2. Comparison results on nodule diagnosis

Table 2 presents the nodule classification performance of AsyCMST and state-of-the-art methods on thyroid and breast ultrasound datasets. AsyCMST achieves the highest accuracy of 82.0% on the thyroid dataset and 81.8% on the breast dataset, outperforming the strongest baseline BUS-Net by 2.9% and 3.1%, respectively. Similar gains are observed in $F_1$-score which shows 80.6% vs. 77.2% on thyroid and 83.4% vs. 80.8% on breast. These improvements are statistically significant (paired $t$-test, $p < 0.01$). The superiority is further confirmed by ROC curves in Fig. 6, where AsyCMST yields the highest AUC on both datasets.

Video models employing late fusion, such as R(2+1)D and ViViT, show limited performance due to their inability to model inter-modal dependencies. Multimodal architectures with cross-attention, including MFT and FC-Former, provide modest improvements but remain constrained by symmetric fusion designs. Methods incorporating ultrasound-specific priors, such as BUS-Net, DKPD, and DAST, perform significantly better by exploiting domain knowledge. Nevertheless, their reliance on heuristic or unidirectional guidance limits full utilization of the inherent asymmetry between BUS spatial structure and CEUS temporal perfusion. By combining multi-task modality-specific enhancement with asymmetric cross-modal spatio-temporal attention, AsyCMST effectively captures directional, clinically meaningful interactions between anatomical context and functional dynamics. This integrated design substantially outperforms all competitors, establishing new state-of-the-art results and validating the importance of knowledge-driven, asymmetric fusion in multimodal ultrasound nodule diagnosis.

### 4.3. Cross-dataset transfer evaluation

To evaluate domain generalization, we perform cross-dataset transfer experiments by training on one dataset and testing on the other. Table 3 summarizes accuracy, recall, precision, and $F_1$-score under this setting. AsyCMST consistently outperforms all competing methods. When trained on the breast dataset and evaluated on thyroid data, it achieves 77.7% accuracy and 75.8% $F_1$-score, exceeding the strongest baseline BUS-Net by 2.5% and 2.7%, respectively. In the opposite direction, thyroid to breast, AsyCMST attains 79.9% accuracy and 81.6% $F_1$-score, improving upon BUS-Net by 6.4% in accuracy and 6.0% in $F_1$-score. The ROC curves for cross-dataset evaluation, shown in Fig. 7, further confirm the superior discriminative power of AsyCMST in both transfer scenarios.

General video models and symmetric multimodal frameworks (R(2+1)D, ViViT, MFT, FC-Former) exhibit relatively poor transferability, with accuracy generally below 73%. Ultrasound-specific approaches incorporating fixed priors (BUS-Net, DKPD, DAST) improve

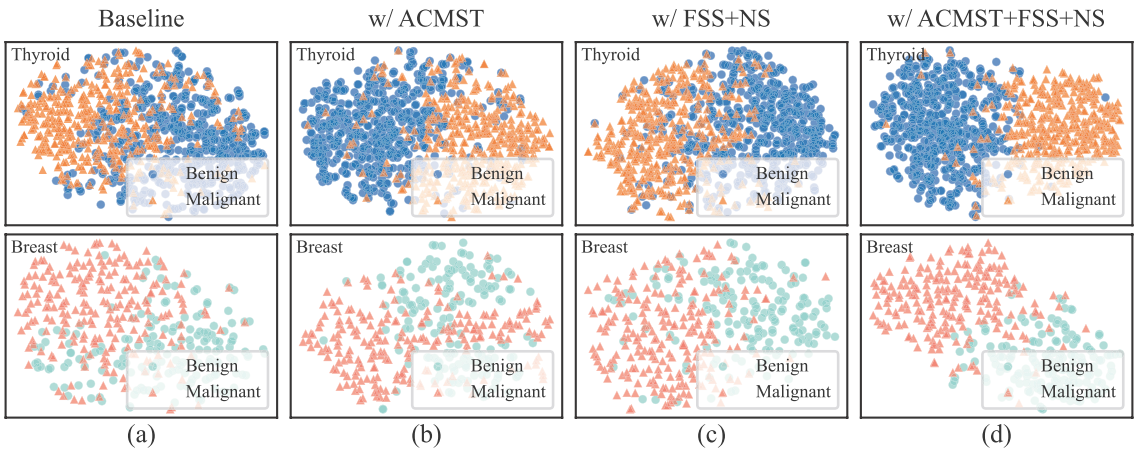


**Fig. 8.** Visualization of t-SNE dimensionality reduction results for the final layer features from different models in the ablation study. (a) Baseline model without the proposed modules. (b) Model incorporating the proposed ACMST module. (c) Model with the FSS and NS modules. (d) Full model integrating all proposed modules. The ACMST module and the FSS+NS modules both enhance feature separability relative to the baseline, while the full model further improves separability significantly, confirming the complementary benefits of integrating all proposed modules.

**Table 4**
Ablation analysis results of the ACMST module, FSS module and NS module in the proposed method. The best result in each column is shown in bold.

| Module | | | *Acc* (%) ↑ | | *p* value |
|---|---|---|---|---|---|
| ACMST | FSS | NS | Thyroid | Breast | |
| | | | 76.4 ± 1.1 | 75.9 ± 1.3 | – |
| ✓ | | | 79.9 ± 1.0 | 79.3 ± 1.8 | < 0.01 |
| | ✓ | | 77.8 ± 1.3 | 77.1 ± 2.0 | < 0.01 |
| | | ✓ | 78.0 ± 0.8 | 77.7 ± 1.6 | < 0.01 |
| | ✓ | ✓ | 78.6 ± 1.5 | 78.2 ± 1.8 | < 0.01 |
| ✓ | ✓ | ✓ | **82.0 ± 1.4** | **81.8 ± 1.7** | < 0.01 |

robustness but remain limited by dataset-specific heuristics. In contrast, AsyCMST learns clinically coherent, asymmetric cross-modal interactions and modality-specific representations that generalize effectively across anatomical domains. By aligning universal diagnostic principles of structure and perfusion through multi-task enhancement and directional attention, AsyCMST achieves markedly stronger cross-dataset performance, demonstrating excellent generalization capability for real-world multimodal ultrasound deployment.

### 4.4. Ablation analysis

An ablation study is conducted to evaluate the contributions of the proposed components: ACMST (asymmetric cross-modal spatio-temporal attention), FSS (frame self-sorting), and NS (nodule segmentation). Results on thyroid and breast datasets are shown in Table 4, with $p$-values computed via paired $t$-tests against the baseline (no modules).

The baseline achieves 76.4 ± 1.1% and 75.9 ± 1.3% accuracy on thyroid and breast datasets, respectively. Adding ACMST alone improves performance to 79.9 ± 1.0% and 79.3 ± 1.8% ($p < 0.01$), confirming its central role in modeling bidirectional spatio-temporal correlations between BUS and CEUS, thus enhancing cross-modal information fusion.

FSS alone yields 77.8 ± 1.3% and 77.1 ± 2.0%, while NS alone reaches 78.0 ± 0.8% and 77.7 ± 1.6% ($p < 0.01$ for both), validating their effectiveness in strengthening temporal coherence (FSS) and spatial localization (NS). Combining FSS and NS without ACMST results in 78.6 ± 1.5% and 78.2 ± 1.8%, showing additive intra-modal benefits.

The full model with all three modules achieves the highest accuracies: 82.0 ± 1.4% (thyroid) and 81.8 ± 1.7% (breast), significantly outperforming all partial configurations ($p < 0.01$). This demonstrates that while FSS and NS enhance modality-specific representations, ACMST is essential for effective cross-modal integration. The results affirm ACMST as the primary innovation, with FSS and NS providing critical support, collectively enabling robust and accurate multimodal ultrasound nodule diagnosis.

Furthermore, Fig. 8 visualizes the feature vectors fed into the fully connected classifier on the thyroid dataset and the breast dataset, projected into 2D using t-SNE (Maaten and Hinton, 2008) under different configurations. Compared to the baseline, the inclusion of ACMST, FSS, and NS markedly improves the separability between benign and malignant samples. The full model exhibits the clearest cluster separation, further validating the reliability and effectiveness of the proposed components in enhancing diagnostic discrimination.

### 4.5. Discussion on the fusion of BUS and CEUS

To verify the contribution of multimodal fusion, we evaluate four input configurations on the thyroid dataset and the breast dataset using dual-branch R(2+1)D, ViViT, and the proposed AsyCMST. The configurations are: both branches receive CEUS only, both branches receive BUS only, CEUS in the upper branch and BUS in the lower branch, and BUS in the upper branch and CEUS in the lower branch. Results are shown in Fig. 9.

For all three models, configurations that combine CEUS and BUS consistently outperform single-modality settings, confirming that integrating the spatial structural clarity of BUS with the temporal hemodynamic richness of CEUS substantially improves diagnostic accuracy and highlighting the critical value of multimodal information fusion.

R(2+1)D and ViViT employ symmetric branch interactions, yielding nearly identical performance when the modalities are swapped between branches, with accuracy differences falling within statistical variation. This symmetry prevents effective exploitation of modality-specific strengths. In contrast, AsyCMST exhibits a pronounced performance gap: as Fig. 9(a) shows, on the thyroid dataset, placing CEUS in the upper branch and BUS in the lower branch achieves 82.0% accuracy, markedly surpassing the reverse configuration at 79.5%; as Fig. 9(b) shows, the corresponding accuracies on the breast dataset are 81.8% and 79.1%, further demonstrating the effectiveness of modality-specific asymmetric attention. This substantial difference demonstrates that the proposed asymmetric spatio-temporal attention successfully leverages BUS spatial context to guide CEUS temporal modeling while allowing CEUS perfusion dynamics to refine ambiguous BUS structural features.

To further interpret the fusion mechanism, Grad-CAM++ (Chattopadhay et al., 2018) visualizations of class activation maps are presented in Fig. 10. In samples where the nodule boundary is indistinct

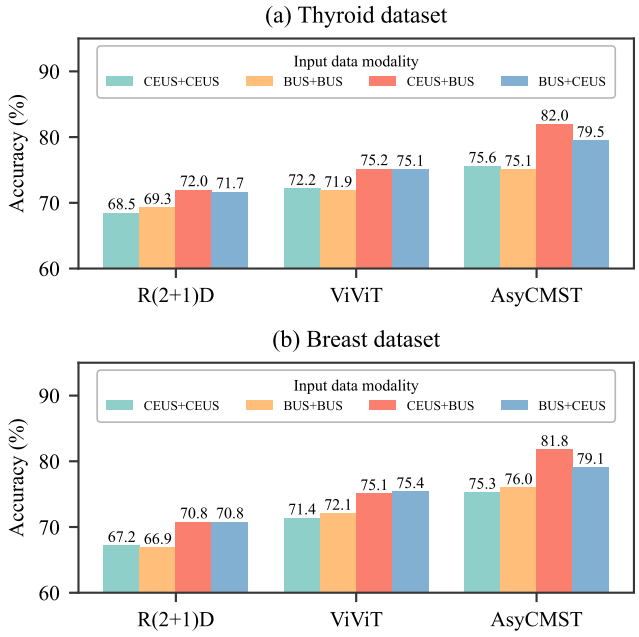


**Fig. 9.** Performance of various methods under different input modality configurations. (a) Results on the thyroid dataset. (b) Results on the breast dataset. CEUS+CEUS and BUS+BUS indicates that both branches receive CEUS or BUS video input, respectively, CEUS+BUS indicates that the upper branch receives CEUS and the lower branch receives BUS, BUS+CEUS indicates that the upper branch receives BUS and the lower branch receives CEUS. Multimodal fusion (CEUS+BUS) consistently outperforms single-modality inputs. Unlike symmetric models, AsyCMST achieves higher accuracy when CEUS is in the upper branch and BUS in the lower branch, demonstrating the benefit of asymmetric attention.

**Table 5**
Comparison of the accuracy and the number of interactions between query and key vectors in cross-modal attention under different strategies. The best result in each column is shown in bold.

| Strategy | *Acc* (%) ↑ | | Interaction number ↓ |
|---|---|---|---|
| | Thyroid | Breast | |
| None | 78.6 ± 1.5 | 78.2 ± 1.8 | – |
| Global | 80.4 ± 0.9 | 79.8 ± 1.1 | $(N\times T)(N\times T\times 2)$ |
| Symmetric | 80.9 ± 0.7 | 81.0 ± 0.6 | $(N\times T)(2N+2T-2)$ |
| **Asymmetric** | **82.0 ± 1.4** | **81.8 ± 1.7** | $(N\times T)(N+T)$ |
| Spatial | 79.6 ± 1.3 | 79.4 ± 1.5 | $(N\times T)\times N\times 2$ |
| Temporal | 79.3 ± 1.4 | 78.9 ± 1.2 | $(N\times T)\times T\times 2$ |

in BUS yet necrosis is evident in CEUS (second column), the model correctly focuses on the necrotic region in CEUS and transfers this attention to the corresponding BUS area, increasing the likelihood of a benign diagnosis. Conversely, when CEUS perfusion is ambiguous while BUS clearly delineates the lesion (eighth column), BUS structural cues effectively steer CEUS attention toward the true lesion region, enabling accurate assessment of perfusion homogeneity. These visualizations confirm that AsyCMST dynamically exploits the complementary strengths of both modalities, achieving clinically coherent and mutually guided focus that significantly enhances diagnostic reliability and performance in multimodal ultrasound nodule assessment.

### 4.6. Discussion of the cross-attention strategy

To further investigate the effectiveness of asymmetric cross-modal spatio-temporal attention in multimodal ultrasound video understanding, we compared five different cross-modal attention strategies.

The first strategy, *Global* spatio-temporal attention, allows every query vector in one modality to interact with all feature vectors in the other modality, expressed as $Q_B[t,n] \leftrightarrow K_C[:,:]$ and $Q_C[t,n] \leftrightarrow K_B[:,:]$. This bidirectional design produces $(N\times T)\times(N\times T)\times 2$ interactions, providing the most comprehensive information exchange but at a high computational cost.

The second strategy, *Symmetric* spatio-temporal attention, restricts interactions to spatially and temporally corresponding positions, formulated as $Q_B[t,n] \leftrightarrow \text{concat}(K_C[:,n], K_C[t,:n-1], K_C[t,n+1:])$ and vice versa. This results in $(N\times T)\times(N+T-1)\times 2$ interactions, significantly reducing computational complexity compared to the global approach.

The proposed *Asymmetric* spatio-temporal attention further simplifies the interaction by leveraging the complementary nature of the two modalities, formulated as $Q_B[t,n] \leftrightarrow K_C[:,n]$ and $Q_C[t,n] \leftrightarrow K_B[t,:]$. This design requires only $(N\times T)\times(N+T)$ interactions.

For comparison, we also implemented *Spatial-only* and *Temporal-only* cross-attention strategies, which perform $(N\times T)\times N\times 2$ and $(N\times T)\times T\times 2$ interactions, respectively. A baseline without any cross-modal attention (*None*) was included as well.

As shown in Table 5, the Symmetric strategy already achieves higher accuracy than the Global approach with substantially lower computational cost, suggesting that removing redundant interactions helps improve model robustness. The proposed Asymmetric attention further reduces the number of interactions while yielding the best diagnostic performance on both thyroid and breast datasets. In contrast, Spatial-only and Temporal-only strategies provide only marginal improvements over the no-cross-attention baseline, indicating that both spatial and temporal interactions contribute meaningfully to performance.

These results demonstrate that designing modality-specific asymmetric cross-modal interactions, rather than exhaustive symmetric or global attention, enables better exploitation of complementary information between BUS and CEUS. Such a design not only improves computational efficiency but also enhances diagnostic accuracy in multimodal ultrasound video analysis.

### 4.7. Discussion on the impact of label noise

In this study, the ground-truth benign/malignant labels were established according to the clinical gold standard through fine-needle aspiration biopsy or postoperative histopathological examination. For nodule contours, even when radiologists provide relatively reliable annotations, considering the limited data sources in the current work, a certain degree of label noise is inevitable when scaling to larger and more diverse multi-center cohorts. Therefore, it is necessary to analyze the impact of label noise on the performance of the proposed model.

To simulate label errors, as Fig. 11(a) shows, we randomly inverted the benign/malignant labels of varying proportions of samples in the training set and retrained the model. The corresponding diagnosis performance on the test set is shown in Fig. 12(a) and (b). As the proportion of flipped labels increases, AUC declines substantially. These results indicate that excessive incorrect labels can severely disrupt the model's ability to learn discriminative representations and significantly degrade diagnostic performance.

Furthermore, nodule contour annotation inherently involves subjectivity. Differences in radiologists' perception and annotation habits may cause control-point offsets or variations in overall contour size, especially in regions with blurry boundaries. To investigate the influence of such variability, as Fig. 11(b) shows, we applied three perturbation strategies to the contour labels across the entire dataset: (1) randomly shifting each control point inward or outward by 0%–15%; (2) uniformly shrinking all contours by 15%; and (3) uniformly expanding all contours by 15%. The model performance before and after these perturbations is compared in Fig. 12(c) and (d). The results show that none of the three contour perturbation strategies led to significant degradation in nodule classification performance, with only a slight

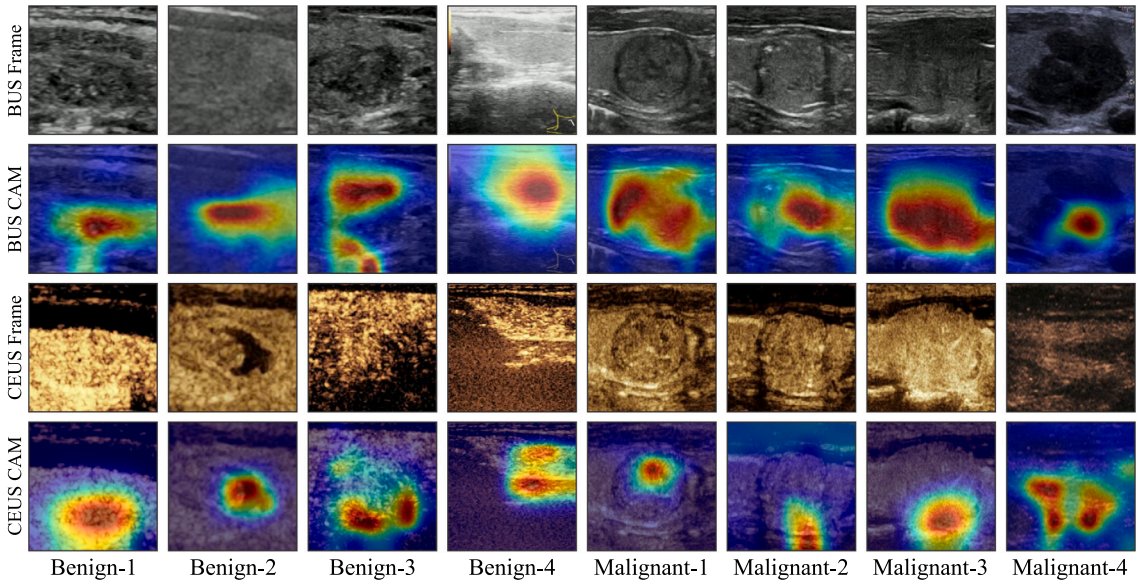


**Fig. 10.** Visualized CAM for a subset of samples from the thyroid dataset. Each column represents one sample, the first to fourth columns correspond to benign nodules, while the fifth to eighth columns represent malignant nodules. For each sample, the first and third rows display the key frames from the BUS and CEUS videos, respectively, while the second and fourth rows present the corresponding CAMs. AsyCMST dynamically leverages complementary modality strengths: when BUS is ambiguous, CEUS cues guide attention; when CEUS is unclear, BUS structure refines focus. This mutually guided attention enables clinically coherent interpretation and enhances diagnostic reliability.

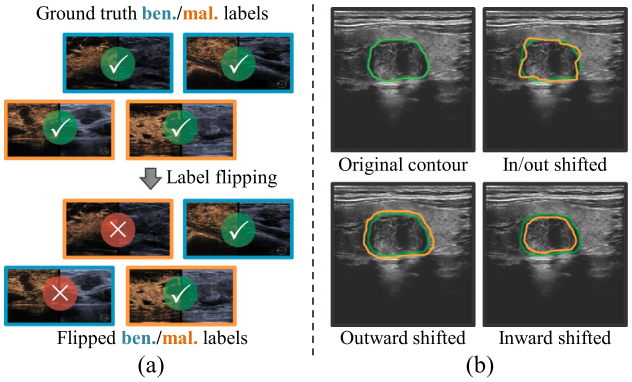


**Fig. 11.** Illustration of label noise strategies. (a) Flipping of the benign-malignant labels. A part benign samples are mislabeled as malignant, and conversely, a part of malignant samples are mislabeled as benign. (b) Shift of the nodule contour annotations. The green contours represent the ground truth contours, and the orange ones are the contours shifted by different strategies.

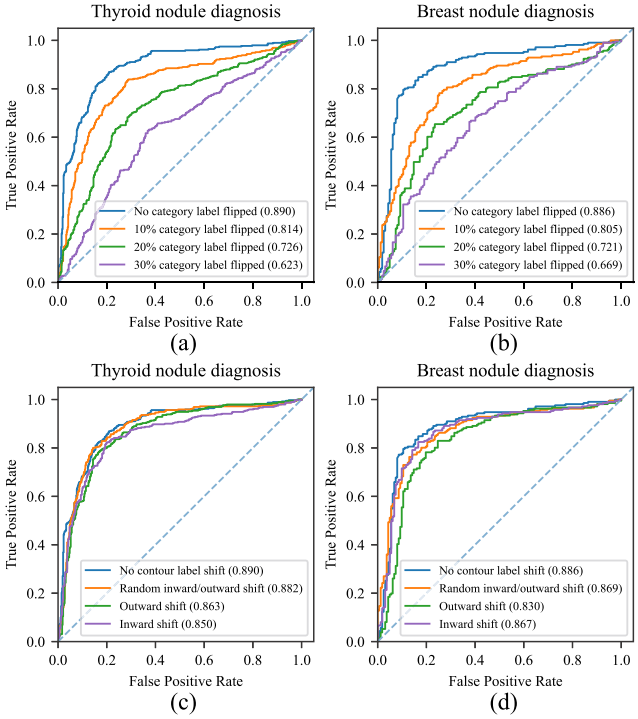


**Fig. 12.** ROC curves of nodule diagnosis under different label noise strategies. (a) Results of category label flipping on the thyroid dataset. (b) Results of category label flipping on the breast dataset. (c) Results of nodule contour label shift on the thyroid dataset. (d) Results of nodule contour label shift on the breast dataset. The value in parentheses in the legend represents the area under the curve (AUC), higher is better.

decrease in AUC observed. This suggests that, as auxiliary supervision, small-range contour deviations have limited impact on the primary diagnostic task.

In future work, we will continue to strictly adhere to pathology-confirmed gold-standard malignancy labels and employ multi-rater consensus protocols to further improve the quality and consistency of contour annotations, thereby enhancing the reliability of model training and evaluation.

### *4.8. Discussion on the layer configuration of the ACMST modules*

To determine the optimal placement of the proposed ACMST modules within the backbone network, an ablation study is conducted by progressively inserting the modules from the shallowest to the deepest layers (layers 1 through 6). The nodule diagnosis performance on both the thyroid and breast datasets is reported in Fig. 13. As shown, model performance consistently improves with the addition of ACMST modules at deeper layers. On the thyroid dataset, accuracy increases steadily from the baseline (no ACMST) to the full configuration with modules in all six layers. A similar upward trend is observed on the breast dataset, with incremental gains at each stage. This progressive

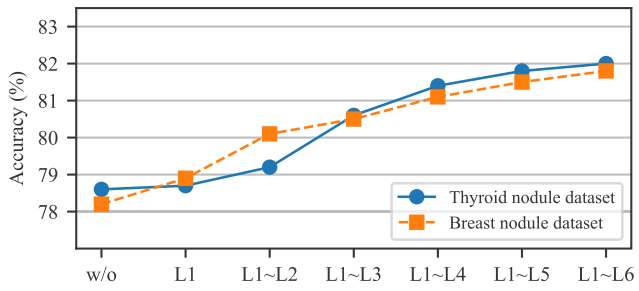


**Fig. 13.** Performance of models incorporating the ACMST module under various layer configurations. The ACMST modules are progressively integrated into the backbone network from shallow to deep layers. L1–Ln indicates the placement of ACMST modules from the first to the $n$th layer.

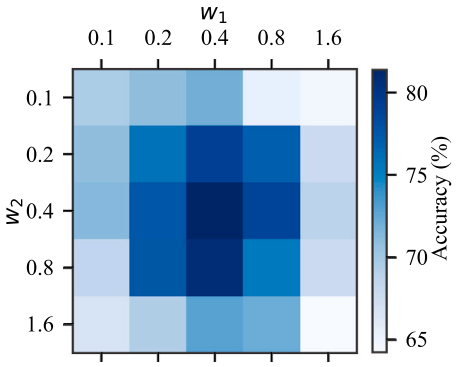


**Fig. 14.** Accuracy performance of the proposed model with various loss coefficients in thyroid nodule diagnosis based on BUS and CEUS. The horizontal and vertical coordinates are the values set for $w_1$ and $w_2$ in $\mathcal{L}_{\text{Overall}}$. The darker colors indicating higher accuracy.

enhancement indicates that cross-modal spatio-temporal interactions benefit from being applied across multiple abstraction levels—from low-level texture and motion cues in early layers to high-level semantic and hemodynamic patterns in later layers. The above results indicate that introducing cross-modal interaction at both the shallow, low-level texture-shape temporal features and the high-level semantic features of BUS and CEUS is beneficial for improving nodule diagnostic performance. Consequently, the final AsyCMST model incorporates ACMST modules in every layer of the backbone network. This design ensures comprehensive inter-modal alignment throughout the feature hierarchy, maximizing the fusion of BUS spatial structures and CEUS temporal dynamics. The results validate that pervasive asymmetric cross-modal attention is essential for achieving optimal diagnostic performance in multimodal ultrasound analysis.

### 4.9. Analysis of loss bias coefficients

In the overall loss function $\mathcal{L}_{\text{Overall}}$ of the proposed model, hyperparameters $w_1$ and $w_2$ control the optimization bias toward the frame self-sorting (FSS) and nodule segmentation (NS) tasks within the multi-task spatio-temporal feature enhancement module. Proper settings must balance $\mathcal{L}_{\text{Diag}}$, $\mathcal{L}_{\text{FSS}}$, and $\mathcal{L}_{\text{Seg}}$ to enable effective feature learning for nodule diagnosis.

To investigate the impact of these coefficients, $w_1$ and $w_2$ are varied logarithmically over $\{0.1, 0.2, 0.4, 0.8, 1.6\}$, yielding 25 combinations. Diagnostic performance on the thyroid dataset is shown in Fig. 14. Accuracy increases initially with $w_1$ and $w_2$, then declines, forming a convex trend. Excessively small weights suppress the auxiliary tasks, limiting spatio-temporal enhancement. Overly large weights over-emphasize FSS or segmentation, diverting optimization from the primary classification objective and degrading performance.

These results indicate that, for nodule diagnosis, the weight setting should prioritize the classification loss. The spatio-temporal feature enhancement module can improve the extraction of temporal and spatial features conducive to diagnosis when assigned an appropriate weight. However, an excessively large weight may cause the model to deviate from the diagnostic accuracy objective, increase the risk of overfitting, and potentially degrade diagnostic performance.

The optimal configuration, $w_1 = 0.4$ and $w_2 = 0.4$, achieves peak accuracy, ensuring balanced contributions from all loss components. This setting enables robust temporal modeling via FSS and precise spatial localization via segmentation, while preserving focus on diagnostic classification.

### 4.10. Analysis of computational efficiency

Table 6 compares diagnostic accuracy with inference time per sample ($t_{inf}$) across all evaluated methods on thyroid and breast ultrasound datasets. Inference time is measured on RTX 3090 GPUs with identical input resolution and frame number. AsyCMST achieves the highest accuracy of 82.0% on thyroid and 81.8% on breast data while maintaining competitive computational efficiency at 33.1 ms per sample. Compared with symmetric multimodal frameworks such as MFT (39.5 ms) and FC-Former (35.5 ms), AsyCMST reduces inference latency by 16% and 7%, respectively, despite delivering substantially higher diagnostic performance. Ultrasound-specific baselines BUS-Net and DKPD exhibit lower latency (30.2 ms and 28.6 ms) yet sacrifice 2.9%–3.1% accuracy relative to AsyCMST.

**Table 6**
Comparison of accuracy and inference efficiency in terms of time per sample across different methods. The best result in each column is shown in bold, and the second-best results are underlined.

| Method | $Acc$ (%) ↑ | | $t_{inf}$ (ms) ↓ |
|---|---|---|---|
| | Thyroid | Breast | |
| R(2+1)D (Tran et al., 2018) | 72.0 ± 1.3 | 70.8 ± 2.2 | **25.7** |
| ViViT (Arnab et al., 2021) | 75.2 ± 1.1 | 75.4 ± 3.4 | 28.8 |
| HFMD (Black and Souvenir, 2024) | 72.3 ± 1.7 | 72.4 ± 2.0 | <u>27.6</u> |
| MFT (Roy et al., 2023) | 76.3 ± 1.3 | 76.1 ± 1.4 | 39.5 |
| FC-Former (Wu et al., 2025) | 74.1 ± 0.9 | 72.8 ± 3.4 | 35.5 |
| BUS-Net (Gong et al., 2022) | <u>79.1 ± 1.0</u> | <u>78.7 ± 2.0</u> | 30.2 |
| DKPD (Chen et al., 2021) | 78.5 ± 0.7 | 78.2 ± 2.3 | 28.6 |
| DAST (F. Chen et al., 2024) | 77.8 ± 1.2 | 77.0 ± 1.5 | 38.2 |
| AsyCMST(ours) | **82.0 ± 1.4** | **81.8 ± 1.7** | 33.1 |

The favorable efficiency of AsyCMST stems from its asymmetric cross-modal spatio-temporal attention mechanism. By selectively establishing directional attention pathways guided by clinical priors — BUS spatial context informing CEUS temporal modeling and vice versa — the module avoids exhaustive pairwise interactions required in symmetric designs. This targeted fusion strategy eliminates redundant cross-modal computations while preserving critical diagnostic signals, yielding a more compact yet expressive representation. Consequently, AsyCMST strikes an effective balance between state-of-the-art accuracy and real-time inference capability, making it well-suited for clinical deployment where both diagnostic reliability and throughput are essential.

## 5. Conclusion

This paper presents AsyCMST, an asymmetric cross-modal spatio-temporal learning framework for multimodal ultrasound nodule diagnosis in thyroid and breast lesions. Clinically, B-mode ultrasound (BUS) is the primary screening tool but often lacks specificity for indeterminate nodules, leading to diagnostic uncertainty and unnecessary biopsies. Contrast-enhanced ultrasound (CEUS) complements BUS by

providing critical microvascular information that improves malignancy differentiation.

AsyCMST addresses the fusion challenge through the ACMST module with adaptive asymmetric inter-modal attention and a multi-task spatio-temporal enhancement module that incorporates frame self-sorting and auxiliary nodule segmentation. Experiments on thyroid and breast datasets show that AsyCMST outperforms state-of-the-art video models, hybrid fusion strategies, and ultrasound-specific methods in accuracy, AUC, and cross-dataset generalization. Ablation studies confirm the importance of asymmetric attention and auxiliary tasks.

The proposed method offers potential to reduce inter-observer variability, support reliable risk stratification, and aid biopsy and treatment decisions. This work advances interpretable multimodal fusion in ultrasound, contributing to more objective AI-assisted nodule evaluation and better patient outcomes.

Our future work will focus on multicenter prospective validation, integration of additional modalities, more precise lesion risk assessment, and improved clinical explainability.

## CRediT authorship contribution statement

**Hongcheng Han:** Writing – original draft, Software, Methodology, Conceptualization. **Zhiqiang Tian:** Writing – review & editing, Conceptualization. **Minghao Wang:** Visualization, Software. **Yutong Zhang:** Data curation. **Dong Zhang:** Writing – review & editing. **Qinbo Guo:** Software. **Jue Jiang:** Project administration. **Hui Guo:** Data curation. **Shaoyi Du:** Writing – review & editing, Project administration, Funding acquisition. **Juan Wang:** Project administration.

## Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Acknowledgments

We would like to express our gratitude to the numerous non-author radiologists from multiple centers for their efforts in data collection and labeling. We also appreciate the valuable discussions with non-author clinical doctors, which have significantly contributed to our research.

## References

Abdar, M., Kollati, M., Kuraparthi, S., Pourpanah, F., McDuff, D., Ghavamzadeh, M., Yan, S., Mohamed, A., Khosravi, A., Cambria, E., et al., 2024. A review of deep learning for video captioning. IEEE Trans. Pattern Anal. Mach. Intell..

Arnab, A., Dehghani, M., Heigold, G., Sun, C., Lučić, M., Schmid, C., 2021. ViViT: A video vision transformer. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. ICCV, pp. 6816–6826.

Black, S., Souvenir, R., 2024. Multi-view classification using hybrid fusion and mutual distillation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. WACV, pp. 270–280.

Cai, J.C., Nakai, H., Kuanar, S., Froemming, A.T., Bolan, C.W., Kawashima, A., Takahashi, H., Mynderse, L.A., Dora, C.D., Humphreys, M.R., et al., 2024. Fully automated deep learning model to detect clinically significant prostate cancer at MRI. Radiology 312 (2), e232635.

Chattopadhay, A., Sarkar, A., Howlader, P., Balasubramanian, V.N., 2018. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. WACV, pp. 839–847.

Chen, F., Han, H., Wan, P., Chen, L., Kong, W., Liao, H., Wen, B., Liu, C., Zhang, D., 2024. Do as sonographers think: Contrast-enhanced ultrasound for thyroid nodules diagnosis via microvascular infiltrative awareness. IEEE Trans. Med. Imaging 43 (11), 3881–3894.

Chen, J., Mei, J., Li, X., Lu, Y., Yu, Q., Wei, Q., Luo, X., Xie, Y., Adeli, E., Wang, Y., et al., 2024. TransUNet: Rethinking the U-Net architecture design for medical image segmentation through the lens of transformers. Med. Image Anal. 97, 103280.

Chen, C., Wang, Y., Niu, J., Liu, X., Li, Q., Gong, X., 2021. Domain knowledge powered deep learning for breast cancer diagnosis based on contrast-enhanced ultrasound videos. IEEE Trans. Med. Imaging 40 (9), 2439–2451.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al., 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations. ICLR, pp. 1–12.

Du, J., Li, W., Lu, K., Xiao, B., 2016. An overview of multi-modal medical image fusion. Neurocomputing 215, 3–20.

Feichtenhofer, C., Fan, H., Malik, J., He, K., 2019. Slowfast networks for video recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. ICCV, pp. 6202–6211.

Fermann, B.S., Nyberg, J., Remme, E.W., Grue, J.F., Grue, H., Haaland, R., Lovstakken, L., Dalen, H., Grenne, B., Aase, S.A., et al., 2024. Cardiac valve event timing in echocardiography using deep learning and triplane recordings. IEEE J. Biomed. Health Inform. 28 (5), 2759–2768.

Gong, X., Zhao, X., Fan, L., Li, T., Guo, Y., Luo, J., 2022. BUS-Net: A bimodal ultrasound network for breast cancer diagnosis. Int. J. Mach. Learn. Cybern. 13 (11), 3311–3328.

Han, H., Tian, Z., Guo, Q., Jiang, J., Du, S., Wang, J., 2025. HSC-T: B-ultrasound-to-elastography translation via hierarchical structural consistency learning for thyroid cancer diagnosis. IEEE J. Biomed. Health Inform. 29 (2), 799–806.

He, D., Li, W., Wang, G., Huang, Y., Liu, S., 2025. MMIF-INet: Multimodal medical image fusion by invertible network. Inf. Fusion 114, 102666.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. CVPR, pp. 770–778.

Huang, R., Lin, Z., Dou, H., Wang, J., Miao, J., Zhou, G., Jia, X., Xu, W., Mei, Z., Dong, Y., et al., 2021. AW3M: An auto-weighting and recovery framework for breast cancer diagnosis using multi-modal ultrasound. Med. Image Anal. 72, 102137.

Kang, Q., Lao, Q., Li, Y., Jiang, Z., Qiu, Y., Zhang, S., Li, K., 2022. Thyroid nodule segmentation and classification in ultrasound images through intra- and inter-task consistent learning. Med. Image Anal. 79, 102443.

Kijanka, P., Urban, M.W., 2024. Ultrasound shear elastography with expanded bandwidth (USEWEB): A novel method for 2D shear phase velocity imaging of soft tissues. IEEE Trans. Med. Imaging 43 (5), 1910–1922.

Li, M., Dal Maso, L., Pizzato, M., Vaccarella, S., 2024. Evolving epidemiological patterns of thyroid cancer and estimates of overdiagnosis in 2013–17 in 63 countries worldwide: A population-based study. Lancet Diabetes Endocrinol. 12 (11), 824–836.

Lin, M., Zhang, Z., Gao, X., Bian, Y., Wu, R.S., Park, G., Lou, Z., Zhang, Z., Xu, X., Chen, X., et al., 2024. A fully integrated wearable ultrasound system to monitor deep tissues in moving subjects. Nature Biotechnol. 42 (3), 448–457.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Guo, B., 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. ICCV, pp. 10012–10022.

Liu, Z., Ning, J., Cao, Y., Wei, Y., Zhang, Z., Lin, S., Hu, H., 2022. Video swin transformer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. CVPR, pp. 3202–3211.

Maaten, L.V.D., Hinton, G., 2008. Visualizing data using T-SNE. J. Mach. Learn. Res. 9 (Nov), 2579–2605.

Priyadarshini, S., Panda, K.B., 2024. Trends in age-specific incidence, mortality, and DALYs of female breast cancer from 1990 to 2021. Aging Med. 7 (6), 770–780.

Qian, X., Pei, J., Zheng, H., Xie, X., Yan, L., Zhang, H., Han, C., Gao, X., Zhang, H., Zheng, W., et al., 2021. Prospective assessment of breast cancer risk from multimodal multiview ultrasound images via clinically applicable deep learning. Nat. Biomed. Eng. 5 (6), 522–532.

Qian, X., Zhang, B., Liu, S., Wang, Y., Chen, X., Liu, J., Yang, Y., Chen, X., Wei, Y., Xiao, Q., et al., 2020. A combined ultrasonic B-mode and color Doppler system for the classification of breast masses using neural network. Eur. Radiol. 30 (5), 3023–3033.

Qin, P., Wu, K., Hu, Y., Zeng, J., Chai, X., 2019. Diagnosis of benign and malignant thyroid nodules using combined conventional ultrasound and ultrasound elasticity imaging. IEEE J. Biomed. Health Inform. 24 (4), 1028–1036.

Qu, J., Huang, D., Shi, Y., Liu, J., Tang, W., 2025. Entropy-aware dynamic path selection network for multi-modality medical image fusion. Inf. Fusion 123, 103312.

Roy, S.K., Deria, A., Hong, D., Rasti, B., Plaza, A., Chanussot, J., 2023. Multimodal fusion transformer for remote sensing image classification. IEEE Trans. Geosci. Remote Sens. 61, 1–20.

Ruan, J., Xu, X., Cai, Y., Zeng, H., Luo, M., Zhang, W., Liu, R., Lin, P., Xu, Y., Ye, Q., et al., 2022. A practical CEUS thyroid reporting system for thyroid nodules. Radiology 305 (1), 149–159.

Shen, Y., Shamout, F.E., Oliver, J.R., Witowski, J., Kannan, K., Park, J., Wu, N., Huddleston, C., Wolfson, S., Millet, A., et al., 2021. Artificial intelligence system reduces false-positive findings in the interpretation of breast ultrasound exams. Nat. Commun. 12 (1), 5645.

Shen, P., Yang, Z., Sun, J., Wang, Y., Qiu, C., Wang, Y., Ren, Y., Liu, S., Cai, W., Lu, H., et al., 2025. Explainable multimodal deep learning for predicting thyroid cancer lateral lymph node metastasis using ultrasound imaging. Nat. Commun. 16 (1), 7052.

Tran, D., Bourdev, L., Fergus, R., Torresani, L., Paluri, M., 2015. Learning spatiotemporal features with 3D convolutional networks. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. ICCV, pp. 4489–4497.

Tran, D., Wang, H., Torresani, L., Ray, J., LeCun, Y., Paluri, M., 2018. A closer look at spatiotemporal convolutions for action recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. CVPR, pp. 6450–6459.

Wang, M., Du, S., Wang, J., Han, H., Huo, H., Zhang, D., Yu, S., Jiang, J., 2025. A segment anything model for transesophageal echocardiography based on bidirectional spatiotemporal context fusion. Inf. Fusion 127, 103771.

Wang, J., Jiang, J., Zhang, D., Zhang, Y.-Z., Guo, L., Jiang, Y., Du, S., Zhou, Q., 2022. An integrated AI model to improve diagnostic accuracy of ultrasound and output known risk features in suspicious thyroid nodules. Eur. Radiol. 1–10.

Wu, X., Cao, Z.-H., Huang, T.-Z., Deng, L.-J., Chanussot, J., Vivone, G., 2025. Fully-connected transformer for multi-source image fusion. IEEE Trans. Pattern Anal. Mach. Intell. 47 (3), 2071–2088.

Yan, P., Gong, W., Li, M., Zhang, J., Li, X., Jiang, Y., Luo, H., Zhou, H., 2024. TDF-Net: Trusted dynamic feature fusion network for breast cancer diagnosis using incomplete multimodal ultrasound. Inf. Fusion 112, 102592.

Zhang, M., Zhang, Y., Liu, S., Han, Y., Cao, H., Qiao, B., 2024. Dual-attention transformer-based hybrid network for multi-modal medical image segmentation. Sci. Rep. 14 (1), 25704.

Zheng, X., Yao, Z., Huang, Y., Yu, Y., Wang, Y., Liu, Y., Mao, R., Li, F., Xiao, Y., Wang, Y., et al., 2020. Deep learning radiomics can predict axillary lymph node status in early-stage breast cancer. Nat. Commun. 11 (1), 1236.

Zhu, R., Li, X., Zhang, X., Wang, J., 2021. HID: the hybrid image decomposition model for MRI and CT fusion. IEEE J. Biomed. Health Inform. 26 (2), 727–739.