Noninvasive Molecular Subtyping of Breast Cancer Using Multimodal Ultrasound Spatiotemporal Transformer

Contents

Supplementary Note 1. Network Architecture 2
Supplementary Note 2. Training Protocol and Implementation Details 6
Supplementary Note 3. Multi-modal US Generation Capabilities 8
Supplementary Note 4. Vision-Language Model Evaluation 11
Supplementary Figure 1. Confusion Matrices and t-SNE Visualizations of MUST-Sub . 13
Supplementary Figure 2. Validation of Spatiotemporal Biomarkers 14
Supplementary Figure 3. Spatial Biomarker Associations with Clinicopathologic Factors 15
Supplementary Figure 4. Temporal Biomarker Associations with Clinicopathologic Factors 16
Supplementary Figure 5. Calibration curves of MUST-Sub across validation cohorts . . . 17
Supplementary Figure 6. Decision Curve Analysis of MUST-Sub 18
Supplementary Figure 7. Threshold-Dependent Performance Trade-offs of MUST-Sub . 19
Supplementary Table 1. Performance Comparison on the Multicenter External Cohort . 20
Supplementary Table 2. Stepwise Ablation Study of MUST-Sub Components 20
Supplementary Table 3. Patient Characteristics across Cohorts 21
Supplementary Table 4. Pairwise DeLong Test across Validation Cohorts. 22

Supplementary Note 1. Network Architecture

Fig. SN1 presents the overview of MUST-Sub comprising four modules: (1) a unified variational autoencoder (VAE) for multimodal compression; (2) a temporal transformer for modeling hemodynamic patterns; (3) a spatial transformer and diffusion multilayer perception (DiffusionMLP) for capturing within-frame morphology; and (4) a classification head for predicting subtypes.

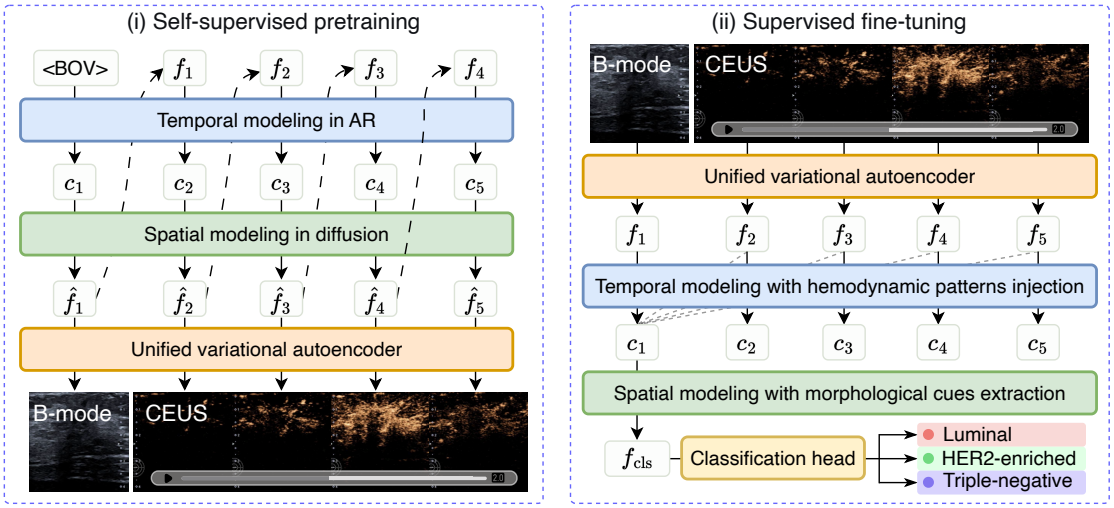


Figure SN1: The overall architecture of MUST-Sub. BOV, begin-of-video; AR, autoregressive; CEUS, contrast-enhanced ultrasound.

Unified variational autoencoder

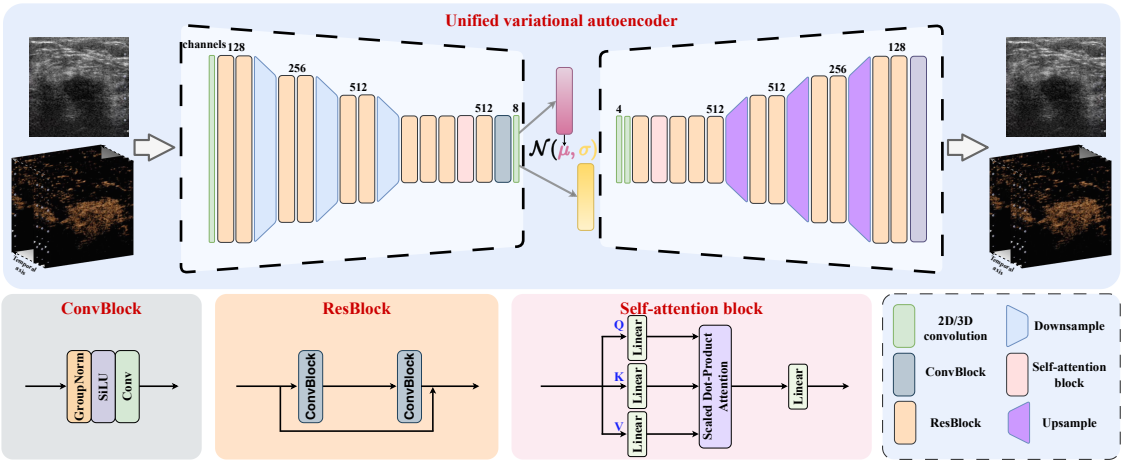


Figure SN2: Architecture of the unified variational autoencoder.

To obtain a shared latent space for both 2D B-mode images and 3D CEUS sequences, we design a unified VAE $\mathcal{V}(\cdot)$ consisting of an encoder $\mathcal{E}_\phi(\cdot)$ and a decoder $\mathcal{D}_\theta(\cdot)$ (Fig. SN2). The encoder maps an input $\boldsymbol{x}$ to a Gaussian distribution parameterized by $(\boldsymbol{\mu}, \boldsymbol{\sigma}) = \mathcal{E}_\phi(\boldsymbol{x})$. A latent representation $\boldsymbol{f}$ is sampled from this distribution via the reparameterization trick:

$$\boldsymbol{f} = \boldsymbol{\mu} + \boldsymbol{\sigma} \odot \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}), \tag{S1}$$

where $\odot$ denotes element-wise multiplication. The decoder reconstructs the corresponding B-mode frame or CEUS sequence from $\boldsymbol{f}$. This unified compression enables subsequent temporal and spatial modeling to be performed in a modality-aligned latent space rather than directly on raw pixels.

The VAE adopts a multi-scale encoder-decoder architecture with four resolution stages (channel dimensions: 128, 256, 512, and 512). Each stage in both the encoder and decoder contains two residual blocks and group normalization (32 groups), with self-attention applied at the bottleneck. For B-mode inputs, the 2D path (Conv2D) maps $3 \times 256 \times 256$ images to $4 \times 32 \times 32$ latent features. For CEUS inputs, the 3D path (Conv3D) compresses $65 \times 3 \times 256 \times 256$ sequences to $17 \times 4 \times 32 \times 32$ latent features, achieving substantial dimensionality reduction while preserving discriminative information.

Temporal transformer

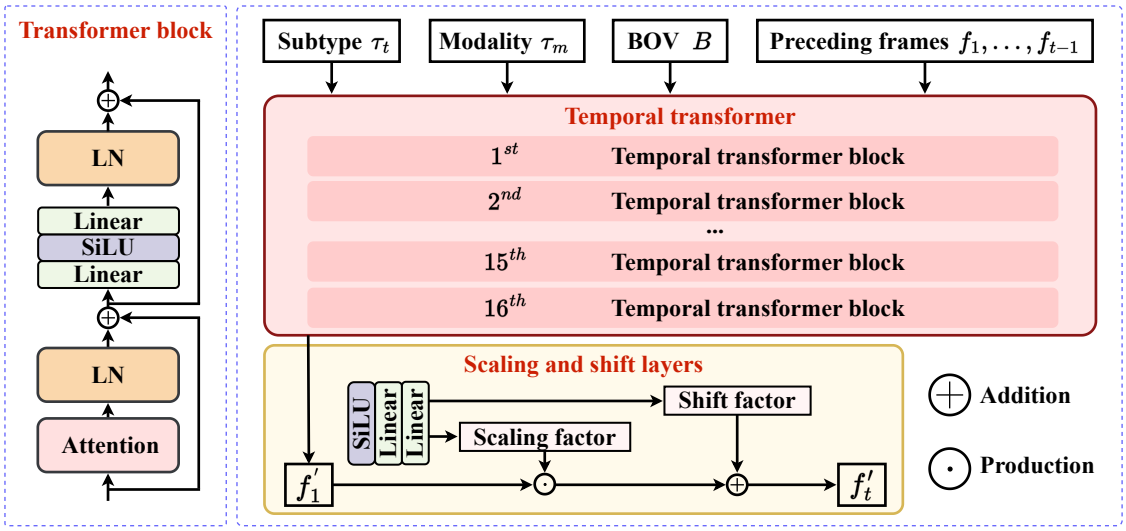


Figure SN3: Architecture of the temporal transformer. LN, layer normalization.

We employ a temporal transformer ($\mathcal{T}_T$; Fig. SN3) to model perfusion dynamics in CEUS sequences. $\mathcal{T}_T$ consists of 16 transformer blocks with an embedding dimension of 1024. Temporal and spatial locations are encoded using 1D and 2D sine-cosine positional embeddings [1]. To improve robustness and temporal continuity, we adopt learnable scaling and shift layers [2] to implement relative, rather than absolute, temporal embeddings. Each transformer block follows a pre-norm design and is implemented sequentially with LayerNorm [3], AdaLN [4], a linear transformation, SiLU activation [5], and a final linear projection.

$\mathcal{T}_T$ operates under two attention regimes. During self-supervised pretraining, block-wise causal masking is applied to preserve temporal causality for autoregressive sequence modeling. Under this causal formulation, tokens at time t can attend to: (1) the subtype condition token τ_t (including a dedicated token for unlabeled data), (2) the modality condition token τ_m, (3) a learnable begin-of-video (BOV) embedding B, and (4) tokens from preceding frames. The resulting autoregressive factorization is

$$p(\tau_t, \tau_m, B, f_1, \ldots, f_{T+1}) = \prod_{t=1}^{T+1} p\left(f_t \mid \tau_t, \tau_m, B, f_1, \ldots, f_{t-1}\right), \tag{S2}$$

where f_t denotes the tokenized representation of the t-th frame. During supervised fine-tuning, the causal mask is removed and $\mathcal{T}_T$ uses bidirectional attention to support multimodal feature fusion.

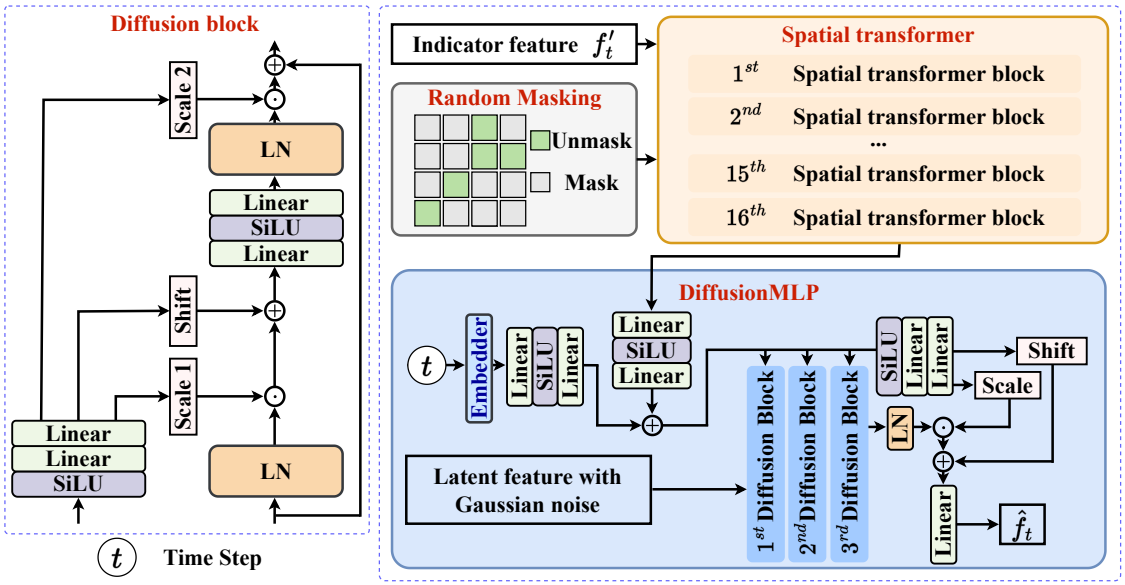


Figure SN4: Architecture of the spatial transformer and DiffusionMLP.

Spatial transformer and DiffusionMLP

We employ a spatial transformer ($\mathcal{T}_S$) and DiffusionMLP (ϵ_{diff}; Fig. SN4) to capture within-frame morphological patterns in US images. $\mathcal{T}_S$ shares the same backbone configuration as $\mathcal{T}_T$, and spatial locations are encoded using 2D sine-cosine positional embeddings. For each frame, spatial tokens are grouped into K token sets $(f_{(t,1)}, \ldots, f_{(t,K)})$ [6].

Similar to $\mathcal{T}_T$, $\mathcal{T}_S$ operates under two attention regimes. During self-supervised pretraining, we adopt a progressive masked-token decoding strategy conditioned on temporal context [6]. Specifically, the output of $\mathcal{T}_T$ for frame f_t serves as an indicator feature f'_t that conditions the sequential decoding of token sets:

$$p\left(f'_t, f_{(t,1)}, \ldots, f_{(t,K)}\right) = \prod_{k=1}^{K} p\left(f_{(t,k)} \mid f'_t, f_{(t,1)}, \ldots, f_{(t,k-1)}\right). \tag{S3}$$

During supervised fine-tuning, the causal mask is removed and $\mathcal{T}_S$ uses standard bidirectional attention to support within-frame feature interaction for downstream classification.

We further introduce a DiffusionMLP that iteratively denoises an initial Gaussian sample into the corresponding clean latent features, conditioned on the spatially decoded token sets. The DiffusionMLP is implemented as a multilayer perceptron comprising three blocks with 1280-dimensional hidden features.

Table SN1: Parameter analysis of MUST-Sub.

Component	Parameters (Million)
Variational autoencoder	239.19
Temporal transformer	201.61
Spatial transformer	234.45
Classification head	0.66
Total	675.91

Classification head

A multilayer perceptron classifier is used as the classification head to map fused spatiotemporal features to subtype probabilities. It consists of two hidden layers with dimensions 512 and 256, followed by an output layer producing probabilities over the three molecular subtypes.

Trainable parameters

The total number of trainable parameters for each component of MUST-Sub is summarized in Table SN1.

Supplementary Note 2. Training Protocol and Implementation Details

VAE objectives

The unified VAE is optimized using a composite objective:

$$\mathcal{L}_{\text{VAE}} = \mathcal{L}_{\text{ELBO}} + \lambda_1 \mathcal{L}_{\text{perceptual}} + \lambda_2 \mathcal{L}_{\text{GAN}}, \tag{S4}$$

where λ_1 and λ_2 weight the contributions of the auxiliary losses. $\mathcal{L}_{\text{ELBO}}$ denotes the negative evidence lower bound, consisting of a reconstruction term and a Kullback-Leibler divergence regularizer. $\mathcal{L}_{\text{perceptual}}$ is a perceptual loss based on LPIPS metric [7] to preserve structural similarity. $\mathcal{L}_{\text{GAN}}$ is an adversarial loss in the least-squares form, implemented with a patch discriminator D [8]:

$$\mathcal{L}_{\text{GAN}} = \frac{1}{2}\mathbb{E}_{\boldsymbol{x}}[(D(\boldsymbol{x}) - 1)^2] + \frac{1}{2}\mathbb{E}_{\boldsymbol{x}}[D(\mathcal{V}(\boldsymbol{x}))^2]. \tag{S5}$$

Two-stage learning

MUST-Sub is trained in two stages: (1) self-supervised pretraining on paired B-mode US and CEUS to learn transferable multimodal representations; and (2) supervised fine-tuning on labeled data for subtype classification.

Pretraining objective. The temporal transformer, spatial transformer, and DiffusionMLP are jointly optimized with a diffusion-style denoising loss [6]. Given the ground-truth latent representation of the t-th frame f_t and its masked prediction $\hat{f}_t$, we sample Gaussian noise $\epsilon \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$ and a diffusion step l, and construct a noisy latent

$$f_t^{\,l} = \sqrt{\bar{\alpha}_l}\, f_t + \sqrt{1 - \bar{\alpha}_l}\, \epsilon, \tag{S6}$$

where $\bar{\alpha}_l$ follows a predefined noise schedule [9, 10, 11]. The denoising loss is then defined as

$$\mathcal{L}\left(f_t \mid \hat{f}_t\right) = \mathbb{E}_{\epsilon,t}\left[\left\|\epsilon - \epsilon_{diff}\left(f_t^l \mid t, \hat{f}_t\right)\right\|^2\right]. \tag{S7}$$

Fine-tuning objective. During supervised fine-tuning, the temporal transformer, spatial transformer, and classification head are optimized end-to-end using focal loss [12] with $\gamma = 2.0$:

$$\mathcal{L}_{\text{cls}} = -\sum_{c=1}^{C} y_c\,(1 - p_c)^{\gamma} \log(p_c), \tag{S8}$$

where y_c and p_c denote the ground-truth label and predicted probability for class c, respectively.

Training configuration

The framework was implemented in PyTorch [13] and trained on NVIDIA V100 GPUs (40 GB) with Adam [14] optimizer. Early stopping [15] was applied based on internal validation performance. The VAE used a cosine decay schedule with warm-up, whereas the transformer pretraining and supervised fine-tuning stages used constant learning rates.

The VAE was trained with $\lambda_1 = \lambda_2 = 0.1$, a learning rate of 2×10^{-5}, and a batch size of 2. The temporal transformer, spatial transformer, and DiffusionMLP were pretrained with a learning rate of 1×10^{-4} and a batch size of 8, whereas supervised fine-tuning of the transformer and classification head used a learning rate of 1×10^{-5} and a batch size of 8. During self-supervised pretraining, we used a dynamic masking ratio ranging from 0.7 to 1.0 and a 1000-step diffusion noise schedule [10]. Comprehensive hyperparameter settings for all three training stages are provided in Table SN2.

Table SN2: Hyperparameters of the proposed framework.

Parameter	Value	Parameter	Value
Variational Autoencoder (Pretraining)			
Architecture		*Optimization*	
Input channels	3	Random seed	42
Latent channels	4	Optimizer	Adam
Block out channels	(128, 256, 512, 512)	Base learning rate	2×10^{-5}
Layers per block	2	Learning rate schedule	Cosine decay
Norm groups	32	Weight decay	0.01
Activation function	SiLU	Batch size	2
Latent shape (B-mode)	$4 \times 32 \times 32$	Max epochs	200
Latent shape (CEUS)	$17 \times 4 \times 32 \times 32$	Perceptual loss weight	0.1
Scaling factor	0.18215	Discriminator weight	0.1
Shift factor	None	KL loss weight	1.0×10^{-6}
		Discriminator start iter	50,001
Spatiotemporal Transformer (Pretraining)			
Architecture		*Optimization*	
Transformer depth	16	Random seed	42
Transformer embed dim	1024	Optimizer	Adam
Transformer num heads	16	Base learning rate	1×10^{-4}
Patch size	2/4	Learning rate schedule	ConstantLR
DiffusionMLP depth	3	Weight decay	0.01
DiffusionMLP embed dim	1280	Batch size	8
Beta schedule	squaredcos_cap_v2	Max epochs	1000
Prediction type	epsilon	Diffusion timesteps	1000
Spatiotemporal Transformer (Supervised Fine-Tuning)			
Architecture		*Optimization*	
Classifier hidden dims	[512, 256]	Random seed	42
Classifier dropout	0.1	Optimizer	Adam
Classifier input dim	1024	Base learning rate	1×10^{-5}
Num classes	3	Learning rate schedule	ConstantLR
Transformer dropout	0.1	Weight decay	0.01
		Batch size	8
		Max epochs	100 with early stop

Data preprocessing and cohort summary

To capture the clinically relevant early enhancement phase, CEUS cine loops were clipped to the 12–20 s post-injection window, corresponding to the early-to-peak arterial phase where wash-in dynamics, time-to-peak and enhancement rate, are most rapid and discriminative; later venous and washout phases were excluded as they exhibit slower signal changes that would introduce redundancy for subtype classification. The labeled development cohort comprised 624 patients (Luminal: 446; HER2-enriched: 92; triple-negative: 86), of whom 547 were used for training and 77 were held out as an internal validation set for early stopping and model selection; the latter were not used for performance reporting. The internal test cohort included 159 patients (Luminal: 113; HER2-enriched: 24; triple-negative: 22). The prospective cohort included 149 patients (Luminal: 116; HER2-enriched: 15; triple-negative: 18). The multicenter external cohort included 122 patients (Luminal: 96; HER2-enriched: 13; triple-negative: 13).

Training data were augmented with spatial perturbations (random horizontal and vertical flips, random rotation ±90°, scaling 0.8–1.2×, translation ±10%, elastic deformation), temporal perturbations (random shift of the perfusion window by ±5 seconds), and intensity perturbations (brightness/contrast adjustment ±10%).

Supplementary Note 3. Multi-modal US Generation Capabilities

The performance of MUST-Sub depends on its ability to learn meaningful spatiotemporal representations from B-mode morphology and CEUS perfusion dynamics. To verify that the proposed spatiotemporal modules capture these modality-specific patterns, we evaluated their generative capability through B-mode image synthesis and CEUS sequence generation. We emphasize that these generation experiments serve solely as technical probes of the learned spatiotemporal representations, and are not intended to produce clinically deployable synthetic data; any such application would require dedicated prospective validation beyond the scope of this work.

B-mode US synthesis

MUST-Sub generates spatially coherent B-mode images while preserving subtype-associated morphological cues (Fig. SN5a). With classifier-free guidance at a scale of 1.1 [16], the synthesized B-mode images reflect clinically relevant patterns used for subtype differentiation, including spiculated margins in Luminal cases, indistinct borders in HER2-enriched lesions, and relatively circumscribed margins in triple-negative tumors.

Clinical validation was performed using a visual Turing test (Fig. SN5b). A board-certified radiologist with over 10 years of experience reviewed real and synthetic images under blinded and randomized settings and was asked to distinguish synthetic from real images. The identification accuracies (real versus synthetic) were: Unlabeled (100% versus 90%), Luminal (100% versus 80%), HER2-enriched (100% versus 70%), and triple-negative (100% versus 100%). The same expert further rated images on a 1-5 scale for overall image quality, lesion conspicuity, and subtype consistency. The mean scores (real versus synthetic) were: Unlabeled (5.0 versus 3.3), Luminal (4.9 versus 3.2), HER2-enriched (4.8 versus 2.8), and triple-negative (4.9 versus 3.5), suggesting that the synthesized images preserve clinically relevant morphology.

We additionally quantified perceptual realism using Kernel Inception Distance (KID) [17] (Fig. SN5c), where lower values indicate greater similarity to real images. For unlabeled images, KID was 37.3 for synthetic-versus-real, compared with 0.2 for the real-versus-real baseline. Subtype-specific KID results (synthetic-versus-real versus real-versus-real baseline) were: Luminal (53.0 versus 54.5), HER2-enriched (47.2 versus 55.5), and triple-negative (60.2 versus 63.5).

Finally, augmenting training with the synthesized B-mode images improved downstream classification performance (Fig. SN5d). The macro-F1 score increased from 42.3 to 45.8, with larger gains for HER2-enriched versus others (23.2 to 29.5) and triple-negative versus others (22.2 to 26.0), indicating that the generated spatial concepts provide useful morphological signals for the downstream model.

CEUS sequence generation

The temporal module of MUST-Sub depends on accurate modeling of perfusion dynamics in CEUS sequences. Fig. SN6a shows representative synthesized CEUS sequences that reproduce subtype-associated enhancement patterns: Luminal cases tend to present gradual and sustained enhancement, HER2-enriched cases show a moderate wash-in followed by a relatively faster wash-out, and triple-negative cases exhibit rapid wash-in and wash-out dynamics.

Quantitative results are summarized in Fig. SN6b. When real CEUS was replaced with synthesized CEUS as the temporal input, MUST-Sub retained most of its performance. On the external cohort, the macro-average F1 reached 0.65 compared with 0.81 using real CEUS, and substantially exceeded the CEUS-free baseline (macro-average F1: 0.41). Similar trends were observed across the internal and prospective cohorts, indicating that the synthesized perfusion dynamics provide effective temporal information for subtype classification. This ability is clinically relevant in settings where contrast administration is unavailable or contraindicated, allowing the model to leverage temporally plausible CEUS patterns without requiring an actual contrast study.

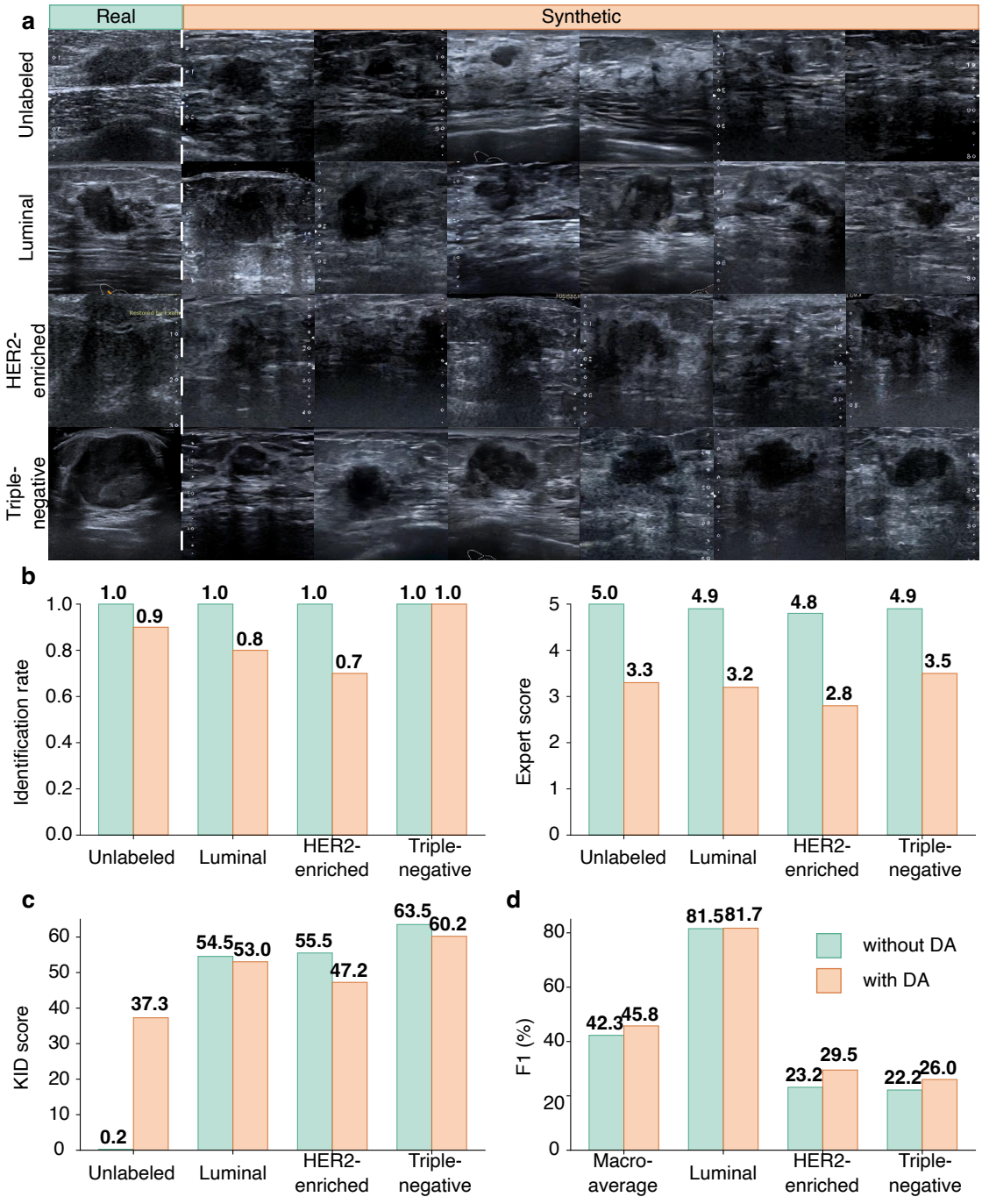


Figure SN5: Evaluation of MUST-Sub's B-mode images generation. **(a)** Representative subtype-conditional B-mode images synthesized by MUST-Sub with CFG scale 1.1, shown alongside real clinical images. **(b)** Visual Turing test: clinician accuracy in distinguishing real from generated images (left) and expert ratings (scale 1-5) of image quality, lesion conspicuity, and subtype consistency (right). **(c)** Quantitative perceptual similarity measured by KID between the internal test cohort (n=159) and MUST-Sub-generated samples (n=256). **(d)** Effect of augmenting training data with synthesized B-mode images on molecular subtyping performance. CFG, classifier-free guidance; KID, kernel inception distance; DA, data augmentation.

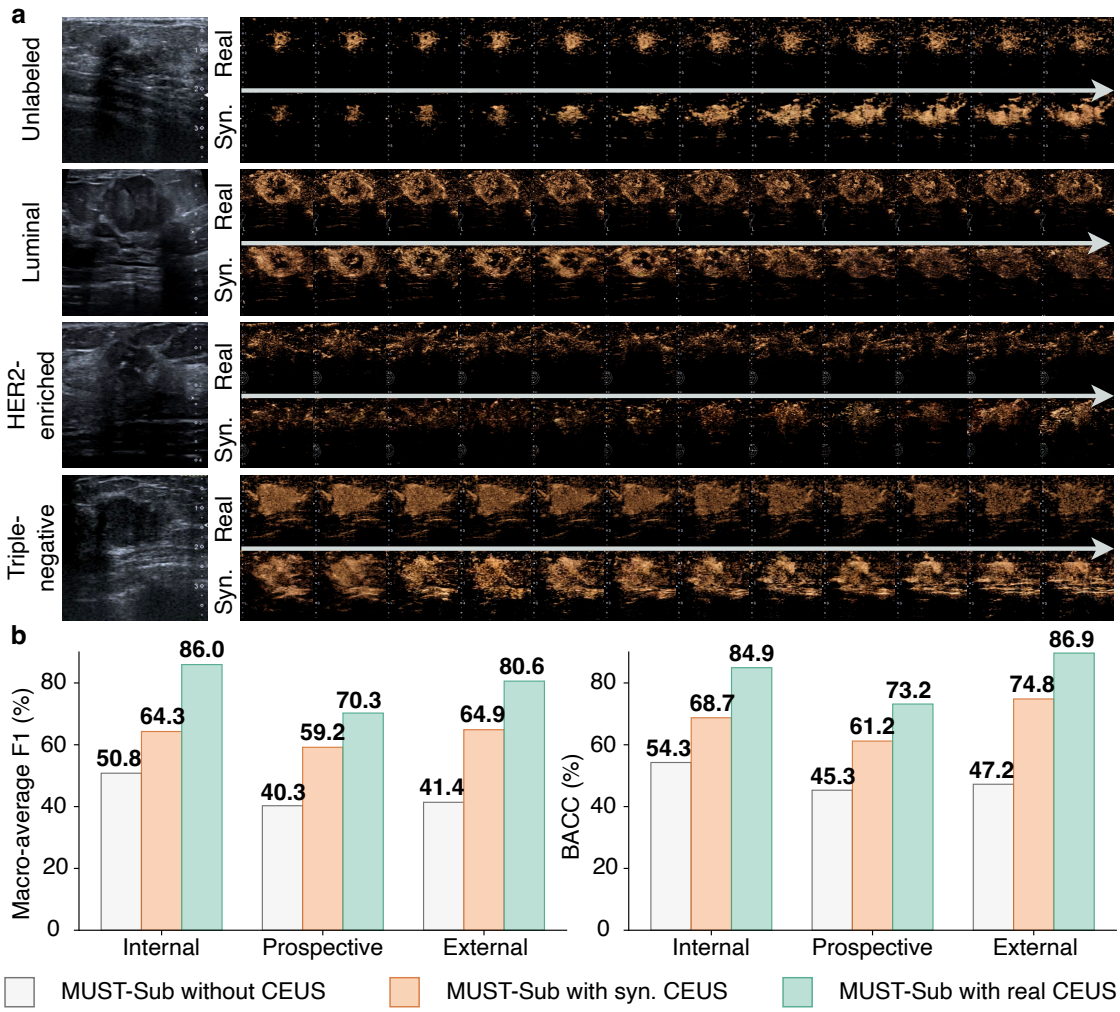


Figure SN6: Evaluation of MUST-Sub's CEUS sequence generation. **(a)** Representative synthesized CEUS sequences illustrating subtype-associated enhancement dynamics. **(b)** Quantitative comparison of model performance when using synthesized versus real CEUS sequences across validation cohorts. BACC, balanced accuracy.

Supplementary Note 4. Vision-Language Model Evaluation

We evaluated the capability of large vision-language models (VLMs) for breast cancer molecular subtyping. The evaluated models include the open-source Qwen3-VL-8B [18], the closed-source Claude 4.5, and the medical-domain model Hulu-Med-4B [19].

Prompt 1: B-mode ultrasound only.

> You are a board-certified breast radiologist with extensive experience in ultrasound interpretation and breast cancer imaging. Your task is to classify the breast cancer molecular subtype using ONLY the provided B-mode ultrasound image.
>
> **Target classes (choose exactly ONE):** (1) Luminal, (2) HER2-enriched, (3) triple-negative.
>
> **Instructions:**
>
> - Base your decision strictly on imaging morphology visible on this B-mode frame (e.g., lesion shape, margins, orientation, echo pattern, posterior acoustic features, surrounding tissue reaction, heterogeneity).
> - Do NOT use patient age, clinical history, pathology, Doppler/CEUS/MRI, or any metadata unless explicitly shown in the image.
> - You may think step-by-step internally, but do not reveal detailed chain-of-thought. Instead, provide a brief evidence summary.
>
> **Output format (must follow exactly):**
> Final subtype: `<Luminal | HER2-enriched | Triple-negative>`
> Evidence (1-3 bullet points): (1) ... (2) ... (3) ...
>
> **Fallback rule:** If the image quality is insufficient or the lesion is not clearly visible, still choose the most likely subtype among the three and note the uncertainty in the evidence bullets.

Prompt 2: B-mode ultrasound + CEUS sequence.

> You are a board-certified breast radiologist with extensive experience in ultrasound interpretation and breast cancer imaging. Your task is to classify the breast cancer molecular subtype using ONLY: (1) the provided B-mode ultrasound image, and (2) the provided CEUS sequence (interpreted as a time series of contrast enhancement).
>
> **Target classes (choose exactly ONE):** (1) Luminal, (2) HER2-enriched, (3) triple-negative.
>
> **Instructions:**
>
> - Base your decision strictly on imaging findings visible in the B-mode frame and the enhancement kinetics in the CEUS sequence.
> - From B-mode, consider morphology (shape, margins, orientation, echogenicity/heterogeneity, posterior acoustic features, surrounding tissue reaction).
> - From CEUS, assess kinetic patterns (wash-in speed, peak intensity, homogeneity, rim enhancement if present, wash-out speed, persistence), and integrate them with B-mode findings.
> - Do NOT use patient age, clinical history, pathology, Doppler/MRI, or any metadata unless explicitly shown in the provided inputs.
> - You may think step-by-step internally, but do not reveal detailed chain-of-thought. Provide only a concise evidence summary.
>
> **Output format (must follow exactly):**
> Final subtype: `<Luminal | HER2-enriched | Triple-negative>`
> Evidence (1-3 bullet points): (1) B-mode: ... (2) CEUS: ... (3) Integrated impression: ...
>
> **Fallback rule:** If either input is low quality or not clearly interpretable, still choose the most likely subtype among the three and explicitly state the limitation.

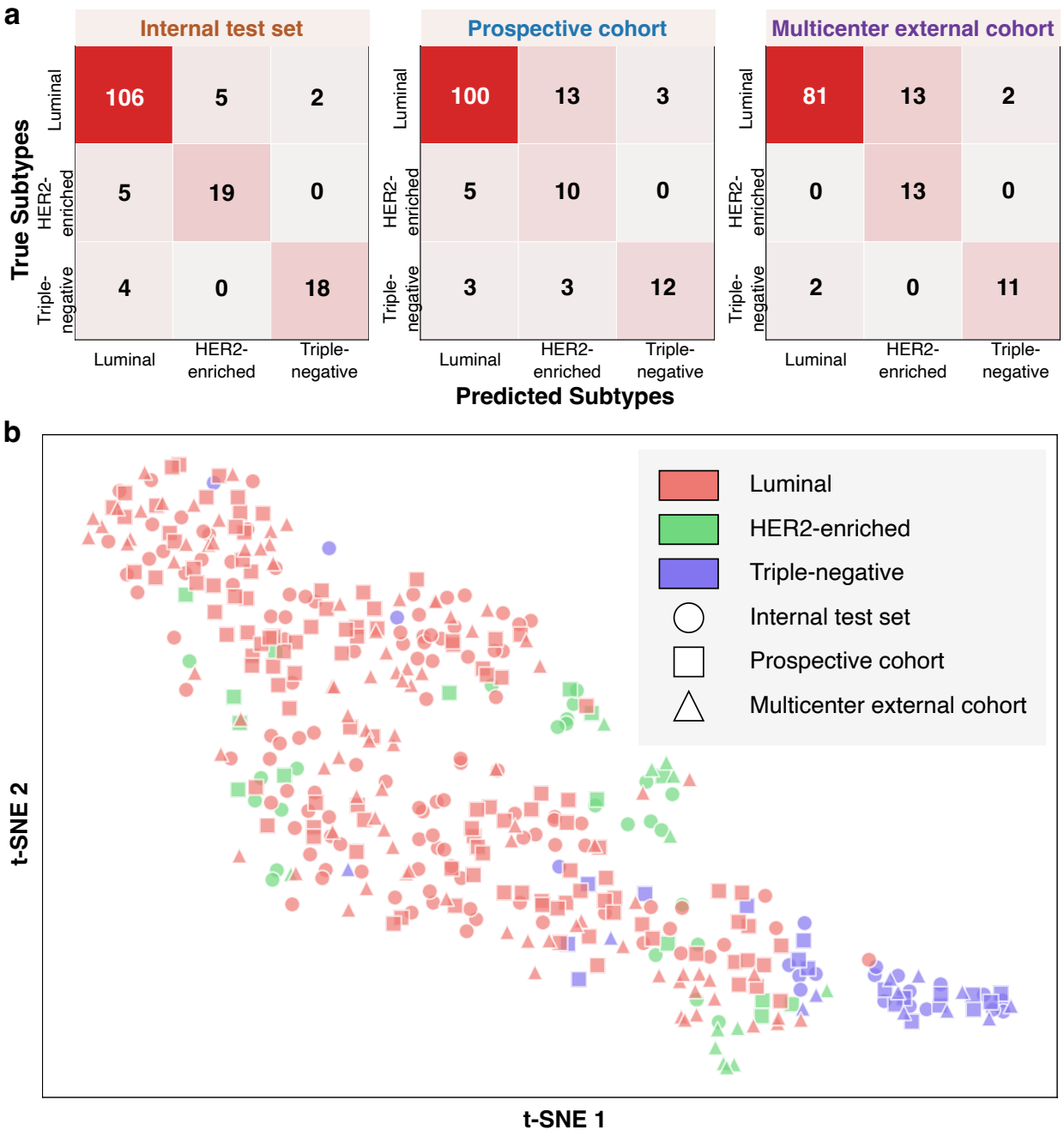
Example outputs from each model under both prompts are provided in Fig. SN7

Ground truth: HER2-enriched

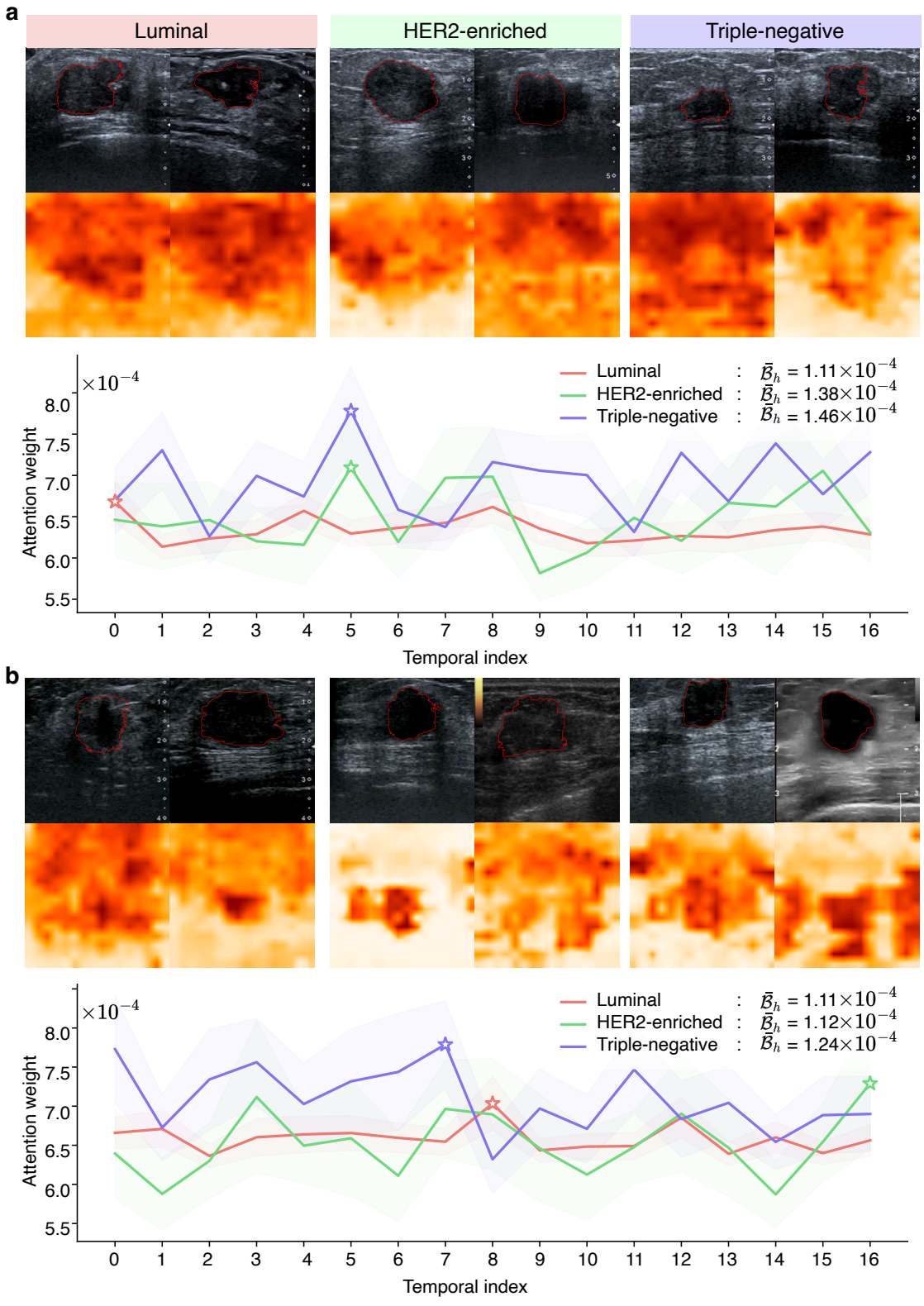


	B-mode US only	B-mode US + CEUS
Qwen3-VL 8B	Final subtype: **Triple-negative** - Irregular, poorly defined margins (not smooth or round) characteristic of aggressive subtypes like triple-negative - Hypoechoic lesion with potential posterior acoustic shadowing, a feature associated with malignancy in ultrasound - Absence of homogeneous echo pattern and well-circumscribed borders, which are more typical of luminal cancers	Final subtype: **Triple-negative** - B-mode: Hypoechoic mass with irregular, spiculated margins and heterogeneous internal echoes; posterior acoustic features suggestive of aggressive growth. - CEUS: Rapid wash-in (sharp rise in enhancement), high peak intensity, and rapid wash-out (quick decline post-peak), indicative of fast vascular permeability. - Integrated impression: Aggressive morphology (irregular margins, heterogeneity) combined with rapid wash-in/wash-out kinetics strongly aligns with triple-negative breast cancer.
Claude 4.5	Final subtype: **Triple-negative** - Oval to round hypoechoic mass with relatively circumscribed margins, which is characteristic of triple-negative breast cancers that often display a "benign-appearing" or pushing border morphology - Marked hypoechogenicity with relatively homogeneous internal echo pattern suggesting high cellularity and rapid growth typical of triple-negative tumors - Posterior acoustic enhancement (slight increased echogenicity deep to the lesion) indicating high cellular density without significant desmoplastic reaction, a feature more commonly associated with triple-negative subtype	Final subtype: **Triple-negative** - B-mode: The lesion demonstrates an oval-to-irregular shape with relatively circumscribed margins, hypoechoic echotexture with some internal heterogeneity, and no significant posterior acoustic shadowing. The morphology suggests a higher-grade, rapidly growing tumor pattern rather than the more spiculated appearance typical of luminal subtypes. - CEUS: The contrast enhancement sequence shows rapid, intense wash-in with heterogeneous internal enhancement and a suggestion of peripheral/rim enhancement pattern. There is relatively quick wash-out without prolonged persistent enhancement, consistent with aggressive tumor biology and increased vascular permeability. - Integrated impression: The combination of circumscribed-appearing margins on B-mode (often seen in high-grade tumors) with the rapid wash-in, heterogeneous enhancement, and early wash-out pattern on CEUS is most consistent with triple-negative breast cancer, which typically demonstrates aggressive vascular characteristics and less organized tumor architecture.
Hulu-Med	Final subtype: **Luminal** - Irregular, hypoechoic lesion with indistinct margins - Posterior acoustic shadowing suggestive of dense fibroglandular tissue - No clear signs of lymphovascular invasion or surrounding tissue distortion',	Final subtype: **Triple-negative** - B-mode: Irregular shape, spiculated margins, heterogeneous echotexture with posterior acoustic shadowing - CEUS: Rapid wash-in, non-homogeneous enhancement, early wash-out kinetics - Integrated impression: Aggressive morphology and rapid enhancement pattern are typical of triple-negative breast cancer\n- Uncertainty/limitations: CEUS sequence appears incomplete, potentially missing late phases which could provide additional diagnostic information'

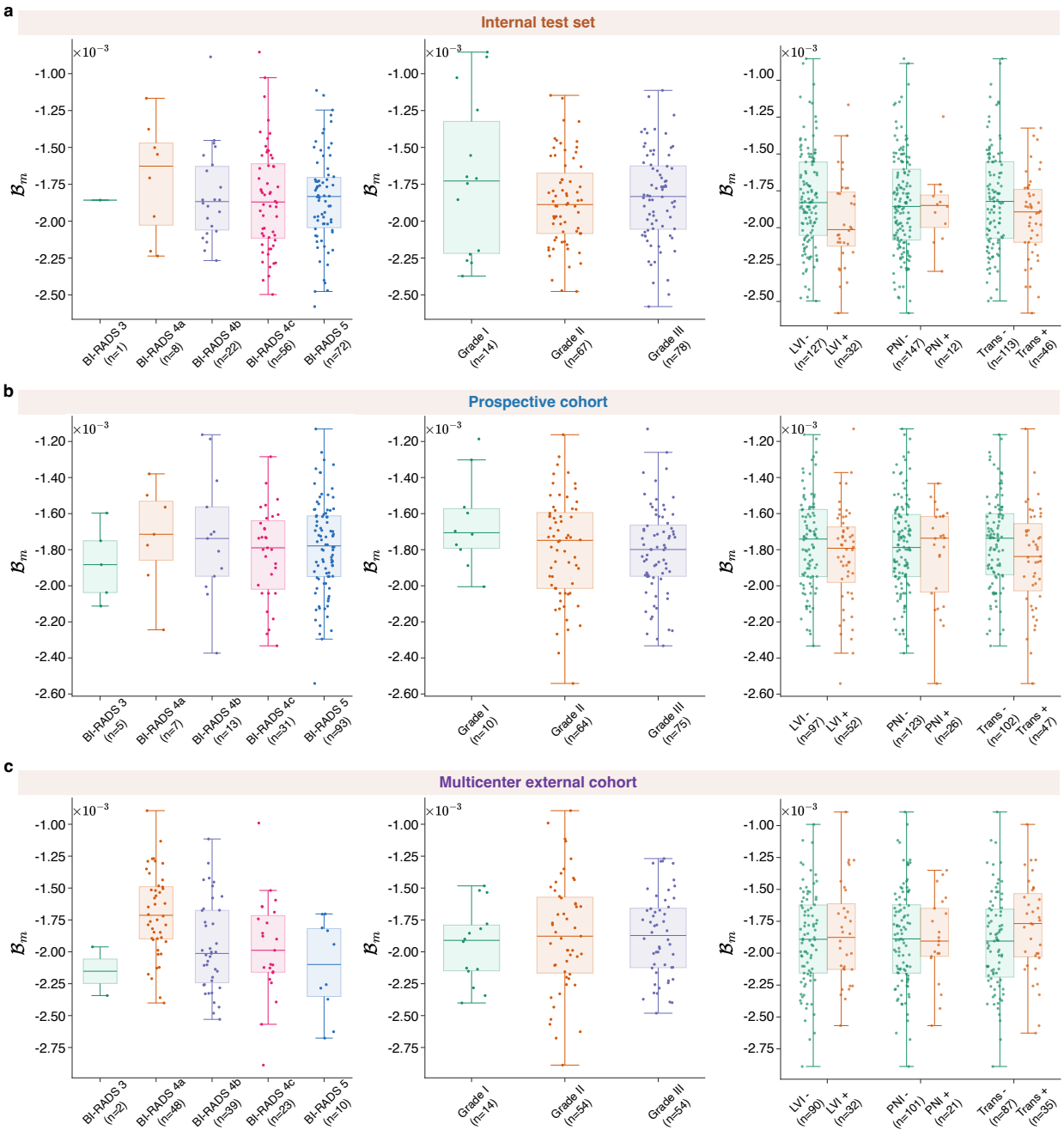
Figure SN7: VLM outputs for an HER2-enriched case. Left: B-mode US only. Right: B-mode US combined with CEUS.



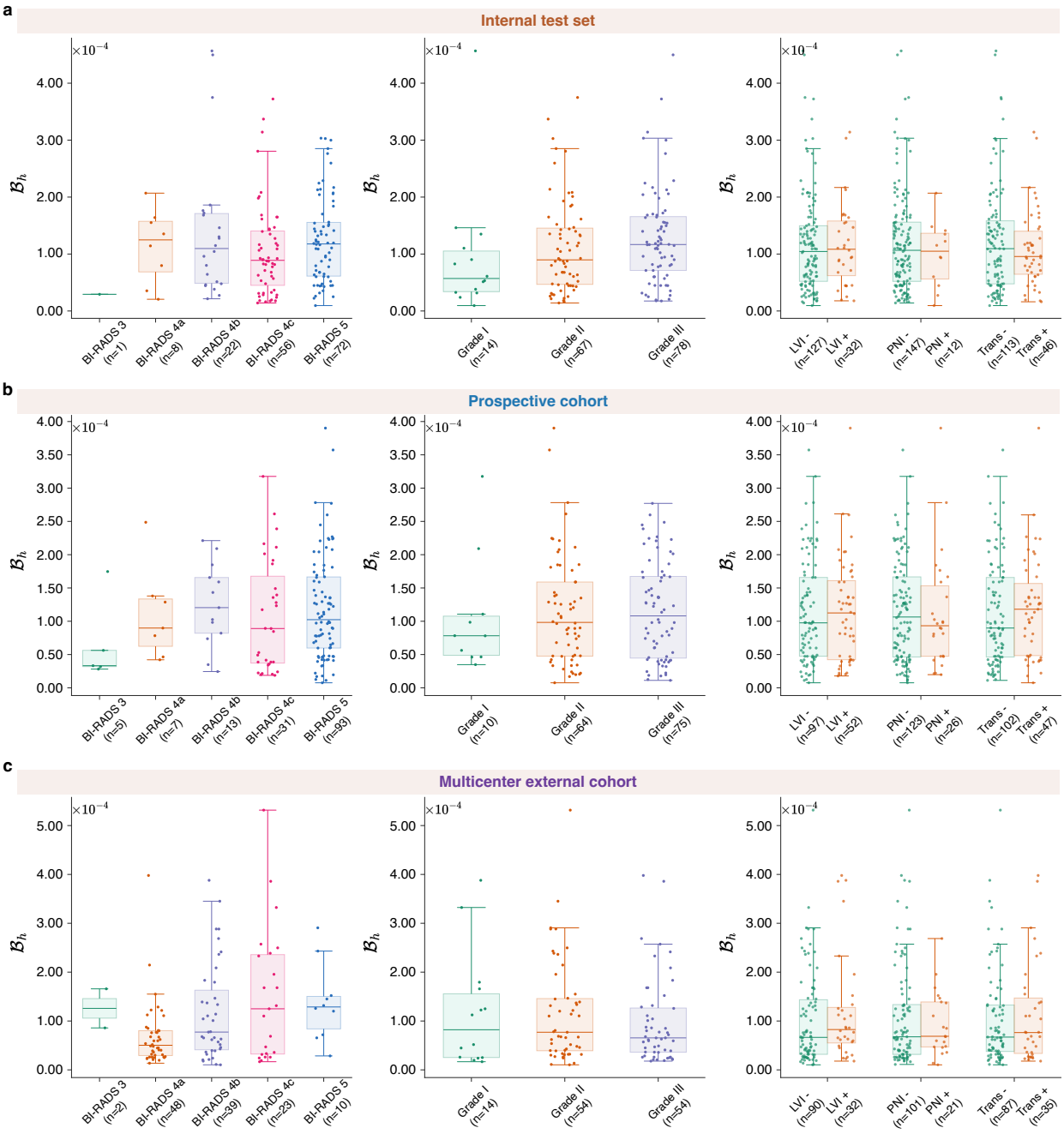
Supplementary Figure 1: **Confusion matrices and t-SNE visualizations of MUST-Sub predictions for the internal, prospective, and multicenter external validation cohorts**. Confusion matrices are based on assigning each case to the class with the highest predicted probability (i.e., `argmax` rule).



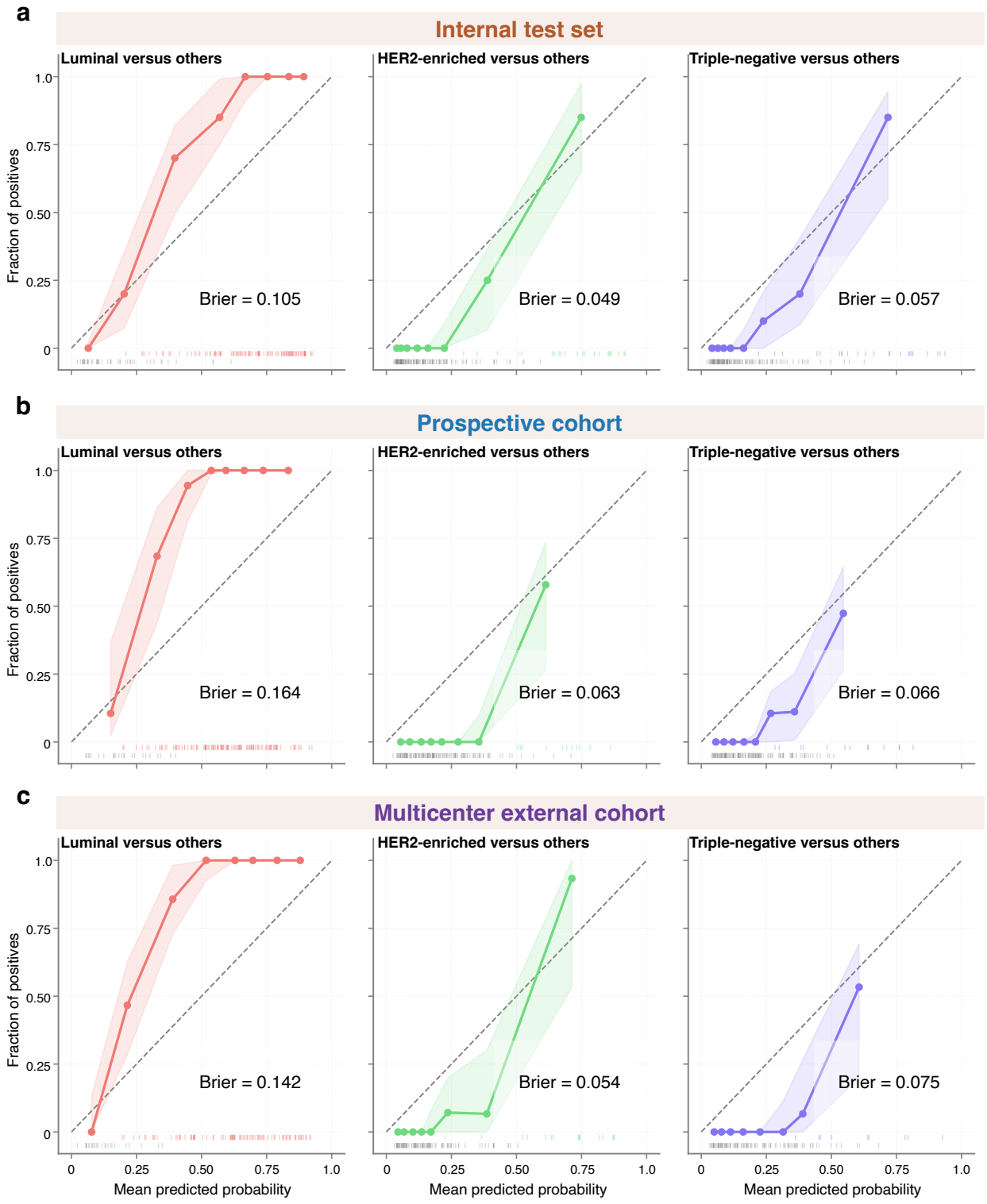
Supplementary Figure 2: **Validation of spatiotemporal biomarkers in the (a) prospective cohort and (b) multi-center external cohort**.



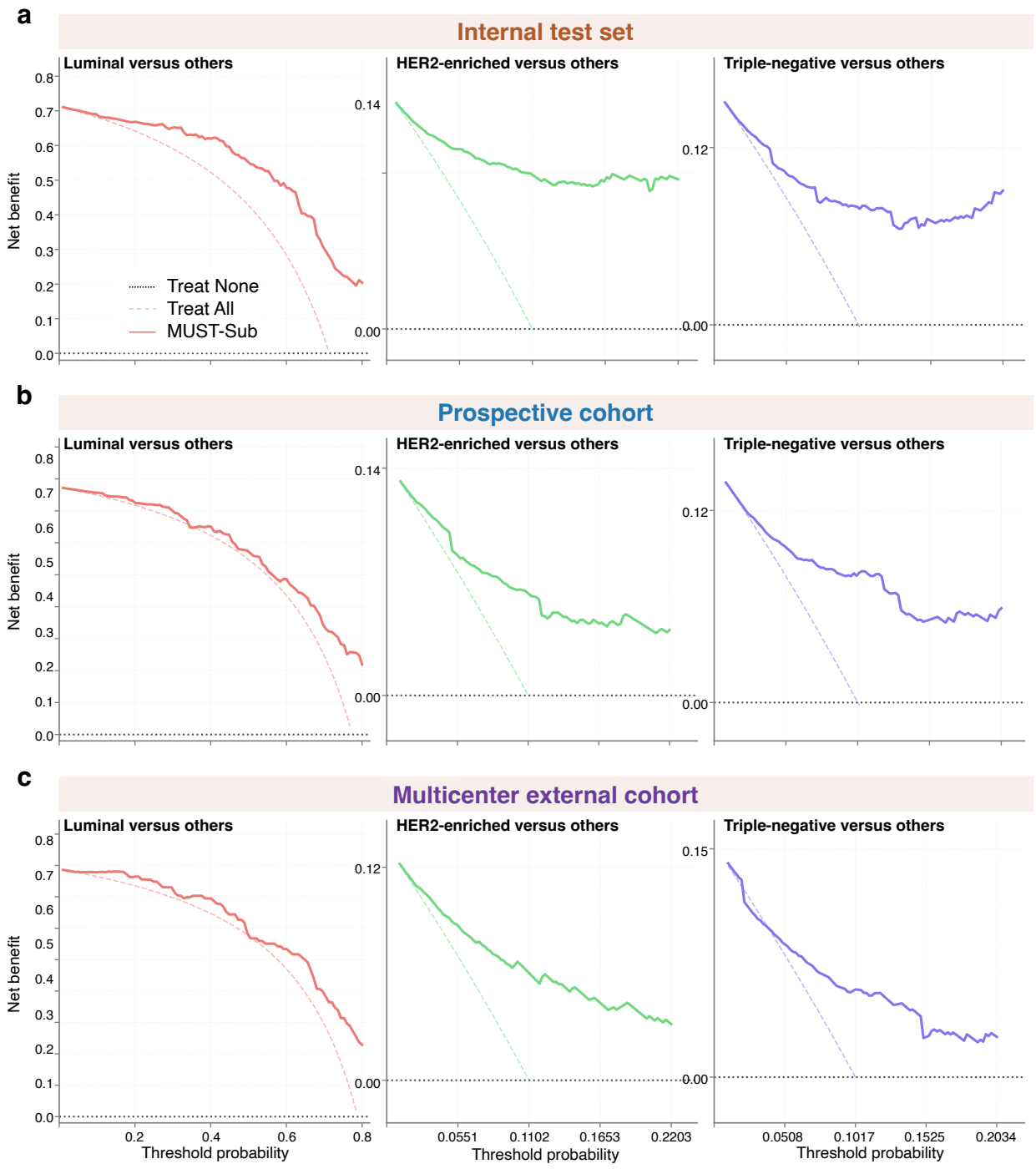
Supplementary Figure 3: Associations between the spatial biomarker and clinicopathologic factors, including BI-RADS category, pathological grade, lymphovascular invasion (LVI), perineural invasion (PNI), and lymph node metastasis, evaluated in the **(a)** internal, **(b)** prospective, and **(c)** external validation cohorts. Group comparisons were performed using the Kruskal–Wallis test and Mann–Whitney U test. No statistically significant differences were observed among subgroups (all $p > 0.05$). BI-RADS, Breast Imaging Reporting and Data System; LVI, lymphovascular invasion; PNI, perineural invasion.



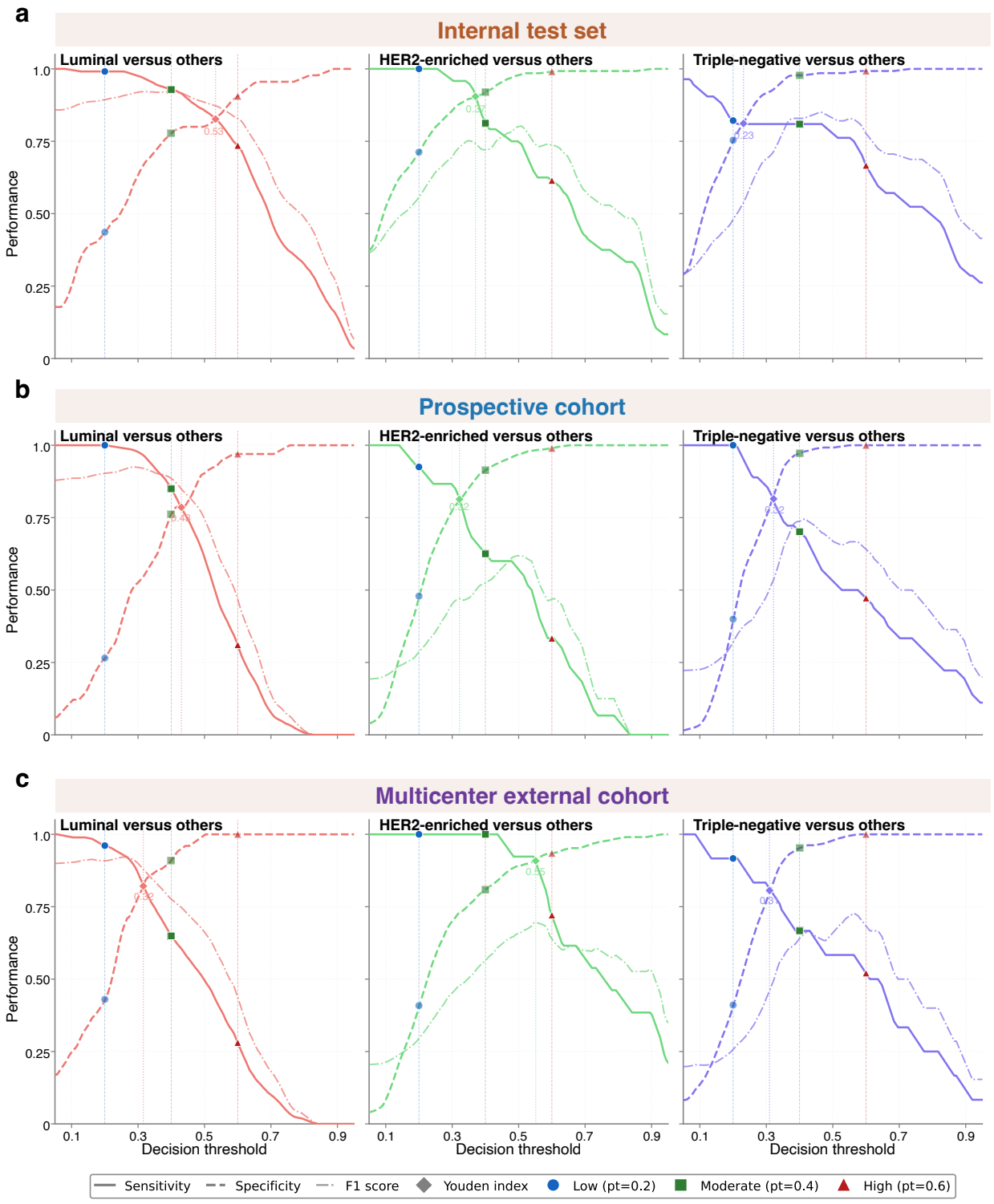
Supplementary Figure 4: **Associations between the temporal biomarker and clinicopathologic factors**, including BI-RADS category, pathological grade, lymphovascular invasion (LVI), perineural invasion (PNI), and lymph node metastasis, evaluated in the **(a)** internal, **(b)** prospective, and **(c)** external validation cohorts. Group comparisons were performed using the Kruskal–Wallis test and Mann–Whitney U test. No statistically significant differences were observed among subgroups (all $p > 0.05$). BI-RADS, Breast Imaging Reporting and Data System; LVI, lymphovascular invasion; PNI, perineural invasion.



Supplementary Figure 5: **Calibration curves for MUST-Sub on the (a) internal, (b) prospective, and (c) external cohorts**. Each column shows one molecular subtype (Luminal, HER2-enriched, triple-negative). Shaded regions indicate 95% confidence intervals derived from 1,000 bootstrap iterations.



Supplementary Figure 6: **Decision curve analysis for MUST-Sub across (a) internal cohort, (b) prospective cohort, and (c) external cohort**. Solid colored lines show the net benefit of using MUST-Sub for each subtype; dashed and dotted lines show "treat all" and "treat none" strategies, respectively.



Supplementary Figure 7: **Threshold-dependent performance trade-offs of MUST-Sub across (a) internal cohort, (b) prospective cohort, and (c) external cohort.**

Supplementary Table 1: **Macro-averaged performance of multiple methods for molecular subtype classification on the multicenter external cohort**. PPV, positive predictive value; NPV, negative predictive value; AUC, area under the receiver operating characteristic curve; w/o, without.

	Sensitivity (95% CI)	**Specificity (95% CI)**	**PPV (95% CI)**	**NPV (95% CI)**	**AUC (95% CI)**
Input with B-mode US only					
Qwen3-VL 8B	0.318 (0.000-0.911)	0.663 (0.040-1.000)	0.284 (0.000-0.744)	0.618 (0.212-1.000)	-
Claude 4.5	0.397 (0.000-0.878)	0.732 (0.315-1.000)	0.348 (0.000-0.910)	0.702 (0.275-1.000)	-
Hulu-Med 4B	0.298 (0.000-0.607)	0.645 (0.275-1.000)	0.282 (0.000-0.749)	0.652 (0.204-1.000)	-
ResNet	0.478 (0.048-0.908)	0.765 (0.428-1.000)	0.506 (0.168-0.845)	0.785 (0.529-1.000)	0.796 (0.695-0.844)
MedViT	0.520 (0.240-0.800)	0.783 (0.581-0.984)	0.480 (0.092-0.867)	0.759 (0.435-1.000)	0.798 (0.670-0.837)
MedMamba	0.584 (0.265-0.903)	0.829 (0.619-1.000)	0.593 (0.295-0.891)	0.834 (0.644-1.000)	0.822 (0.742-0.877)
Input with B-mode US and CEUS					
Qwen3-VL 8B	0.337 (0.000-0.913)	0.660 (0.049-1.000)	0.312 (0.086-0.538)	0.636 (0.213-1.000)	-
Claude 4.5	0.292 (0.000-0.835)	0.631 (0.051-1.000)	0.175 (0.000-0.429)	0.600 (0.187-1.000)	-
Hulu-Med 4B	0.367 (0.000-0.921)	0.706 (0.298-1.000)	0.360 (0.000-0.938)	0.696 (0.247-1.000)	-
MUST-Sub w/o pretraining	0.764 (0.608-0.920)	0.870 (0.745-0.996)	0.623 (0.244-1.000)	0.796 (0.443-1.000)	0.873 (0.782-0.946)
MUST-Sub	0.897 (0.814-0.958)	0.928 (0.884-0.964)	0.774 (0.673-0.863)	0.866 (0.824-0.904)	0.922 (0.825-0.964)

Supplementary Table 2: **Stepwise ablation study of MUST-Sub components (Macro-AUC)**.

Model variant	**Internal**	**Prospective**	**External**
Input with B-mode US only			
CNN (ResNet-18)	0.821	0.810	0.796
ViT (MedViT)	0.793	0.778	0.798
VAE + MLP	0.832	0.819	0.808
VAE + Spatial decoder without pretraining	0.815	0.793	0.781
VAE + Spatial decoder with pretraining	0.859	0.841	0.849
Input with B-mode US and CEUS			
VAE + Spatiotemporal decoder without pretraining	0.924	0.842	0.873
VAE + Spatiotemporal decoder with pretraining (**MUST-Sub**)	**0.936**	**0.898**	**0.922**

Supplementary Table 3: **Patient characteristics across training, validation, and test cohorts**.

	Train ($N = 547$)	**Internal Val** ($N = 77$)	**Test** ($N = 159$)	**Prospective Test** ($N = 149$)	**External Test** ($N = 122$)
Molecular subtypes ($p = 0.970$)					
Luminal	391 (71.5%)	55 (74.4%)	113 (71.1%)	116 (77.9%)	96 (78.7%)
HER2-enriched	81 (14.8%)	11 (14.3%)	24 (15.1%)	15 (10.1%)	13 (10.7%)
Triple-negative	75 (13.7%)	11 (14.3%)	22 (13.8%)	18 (12.1%)	13 (10.7%)
Lesion location ($p = 0.854$)					
Left	283 (51.7%)	43 (55.8%)	86 (54.1%)	84 (56.4%)	65 (53.3%)
Right	264 (48.3%)	34 (44.2%)	73 (45.9%)	65 (43.6%)	57 (46.7%)
Age (years) ($p = \mathbf{0.044}$)					
Age ≤50	156 (28.5%)	29 (37.7%)	52 (32.7%)	39 (26.2%)	49 (40.2%)
Age >50	391 (71.5%)	48 (62.3%)	107 (67.3%)	110 (73.8%)	73 (59.8%)
Pathologic size (mm) ($p = 0.297$)					
≤20	286 (52.3%)	39 (50.6%)	86 (54.1%)	70 (47.0%)	41 (33.6%)
>20	260 (47.5%)	37 (48.1%)	73 (45.9%)	73 (49.0%)	58 (47.5%)
NA	1 (0.2%)	1 (1.3%)	0 (0.0%)	6 (4.0%)	23 (18.9%)
Histologic grade ($p = 0.842$)					
Grade I	47 (8.6%)	9 (11.7%)	14 (8.8%)	12 (8.1%)	14 (11.5%)
Grade II	252 (46.1%)	34 (44.2%)	63 (39.6%)	65 (43.6%)	53 (43.4%)
Grade III	248 (45.3%)	34 (44.2%)	82 (51.6%)	72 (48.3%)	55 (45.1%)
PR status ($p = 0.896$)					
Positive	338 (61.8%)	50 (64.9%)	101 (63.5%)	94 (63.1%)	81 (66.4%)
Negative	209 (38.2%)	27 (35.1%)	58 (36.5%)	55 (36.9%)	41 (33.6%)
HER2 status ($p = 0.138$)					
Positive	122 (22.3%)	22 (28.6%)	39 (24.5%)	43 (28.9%)	39 (32.0%)
Negative	425 (77.7%)	55 (71.4%)	120 (75.5%)	106 (71.1%)	83 (68.0%)
Ki-67 index (%) ($p = 0.303$)					
≤14	124 (22.7%)	23 (29.9%)	36 (22.6%)	26 (17.4%)	30 (24.6%)
>14	423 (77.3%)	54 (70.1%)	123 (77.4%)	123 (82.6%)	92 (75.4%)
Lymphovascular invasion ($p = 0.061$)					
Present	141 (25.8%)	24 (31.2%)	33 (20.8%)	52 (34.9%)	32 (26.2%)
Absent	406 (74.2%)	53 (68.8%)	126 (79.2%)	97 (65.1%)	90 (73.8%)
Perineural invasion ($p = \mathbf{0.002}$)					
Present	60 (11.0%)	3 (3.9%)	12 (7.5%)	26 (17.4%)	22 (18.0%)
Absent	487 (89.0%)	74 (96.1%)	147 (92.5%)	123 (82.6%)	100 (82.0%)
Lymph-node metastasis ($p = 0.800$)					
0	379 (69.3%)	46 (59.7%)	112 (70.4%)	102 (68.5%)	86 (70.5%)
1-3	106 (19.4%)	21 (27.3%)	30 (18.9%)	31 (20.8%)	26 (21.3%)
>3	62 (11.3%)	10 (13.0%)	17 (10.7%)	16 (10.7%)	10 (8.2%)

NA, not available. Pathologic size was not recorded for 23 patients in the external cohort due to differences in clinical documentation practices across institutions.

Supplementary Table 4: **Pairwise DeLong test p-values comparing MUST-Sub against baseline models across validation cohorts**. Bold indicates $p < 0.05$ (uncorrected); the Bonferroni-corrected threshold for macro-average comparisons is $\alpha = 0.0056$.

Comparison	**Internal (n=159)**	**Prospective (n=149)**	**External (n=122)**
MUST-Sub vs. ResNet-18			
Luminal vs others	$p = \mathbf{0.013}$	$p = \mathbf{0.039}$	$p = 0.184$
HER2-enriched vs others	$p = \mathbf{0.002}$	$p = \mathbf{0.034}$	$p < \mathbf{0.001}$
Triple-negative vs others	$p = \mathbf{0.026}$	$p = \mathbf{0.018}$	$p = 0.225$
Macro-average	$p < \mathbf{0.001}$	$p = \mathbf{0.033}$	$p = \mathbf{0.003}$
MUST-Sub vs. MedViT			
Luminal vs others	$p < \mathbf{0.001}$	$p = \mathbf{0.024}$	$p = 0.267$
HER2-enriched vs others	$p = \mathbf{0.001}$	$p = \mathbf{0.038}$	$p < \mathbf{0.001}$
Triple-negative vs others	$p = \mathbf{0.023}$	$p = \mathbf{0.002}$	$p = 0.252$
Macro-average	$p < \mathbf{0.001}$	$p = \mathbf{0.005}$	$p = \mathbf{0.001}$
MUST-Sub vs. MedMamba			
Luminal vs others	$p = \mathbf{0.032}$	$p = \mathbf{0.022}$	$p = 0.391$
HER2-enriched vs others	$p < \mathbf{0.001}$	$p = \mathbf{0.039}$	$p = \mathbf{0.004}$
Triple-negative vs others	$p = 0.148$	$p = \mathbf{0.023}$	$p = 0.251$
Macro-average	$p = \mathbf{0.004}$	$p = \mathbf{0.031}$	$p = \mathbf{0.023}$

References

[1] Ashish Vaswani et al. Attention is all you need. In *NIPS*, volume 30, 2017.

[2] Haoge Deng et al. Autoregressive video generation without vector quantization. In *ICLR*, 2024.

[3] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

[4] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *ICCV*, pages 1501–1510, 2017.

[5] Stefan Elfwing, Eiji Uchibe, and Kenji Doya. Sigmoid-weighted linear units for neural network function approximation in reinforcement learning. *Neural Networks*, 107:3–11, 2018.

[6] Tianhong Li, Yonglong Tian, He Li, Mingyang Deng, and Kaiming He. Autoregressive image generation without vector quantization. In *NIPS*, pages 56424–56445, 2024.

[7] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, pages 586–595, 2018.

[8] Xudong Mao et al. Least squares generative adversarial networks. In *ICCV*, pages 2794–2802, 2017.

[9] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *ICML*, pages 8162–8171. PMLR, 2021.

[10] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NIPS*, volume 33, pages 6840–6851, 2020.

[11] Huiwen Chang et al. Muse: Text-to-image generation via masked generative transformers. In *ICML*, pages 4055–4075, 2023.

[12] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *ICCV*, pages 2980–2988, 2017.

[13] Adam Paszke et al. Pytorch: An imperative style, high-performance deep learning library. In *NIPS*, volume 32, 2019.

[14] Diederik P Kingma. Adam: A method for stochastic optimization. In *ICLR*, page 13, 2014.

[15] Rich Caruana, Steve Lawrence, and C Giles. Overfitting in neural nets: Backpropagation, conjugate gradient, and early stopping. In *NIPS*, page 381–387, 2014.

[16] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NIPS*, 2021.

[17] Mikołaj Bińkowski, Danica J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying MMD GANs. *arXiv preprint arXiv:1801.01401*, 2018.

[18] Shuai Bai et al. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025.

[19] Songtao Jiang et al. Hulu-Med: A transparent generalist model towards holistic medical vision-language understanding. *arXiv preprint arXiv:2510.08668*, 2025.