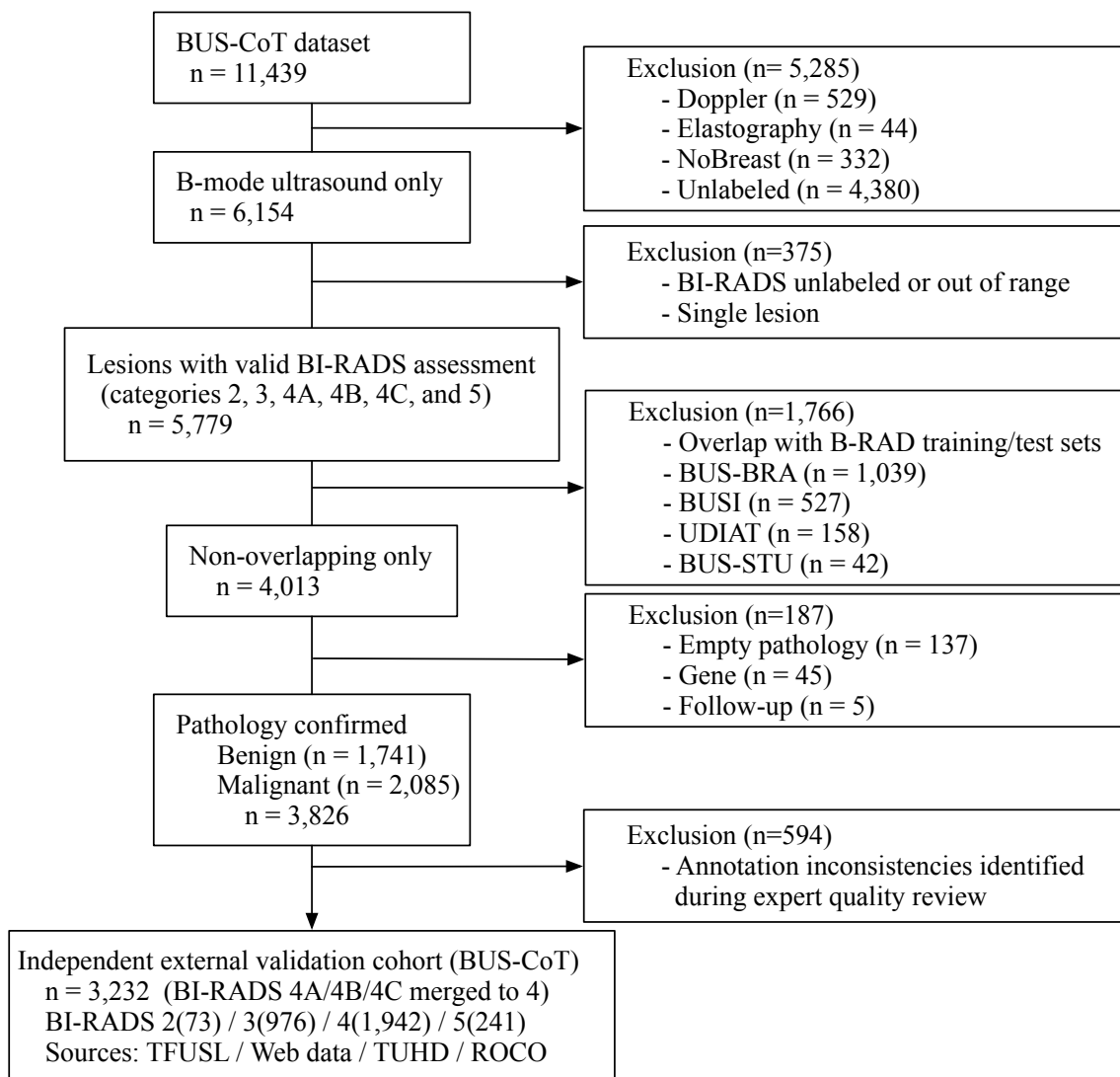


Supplementary Information

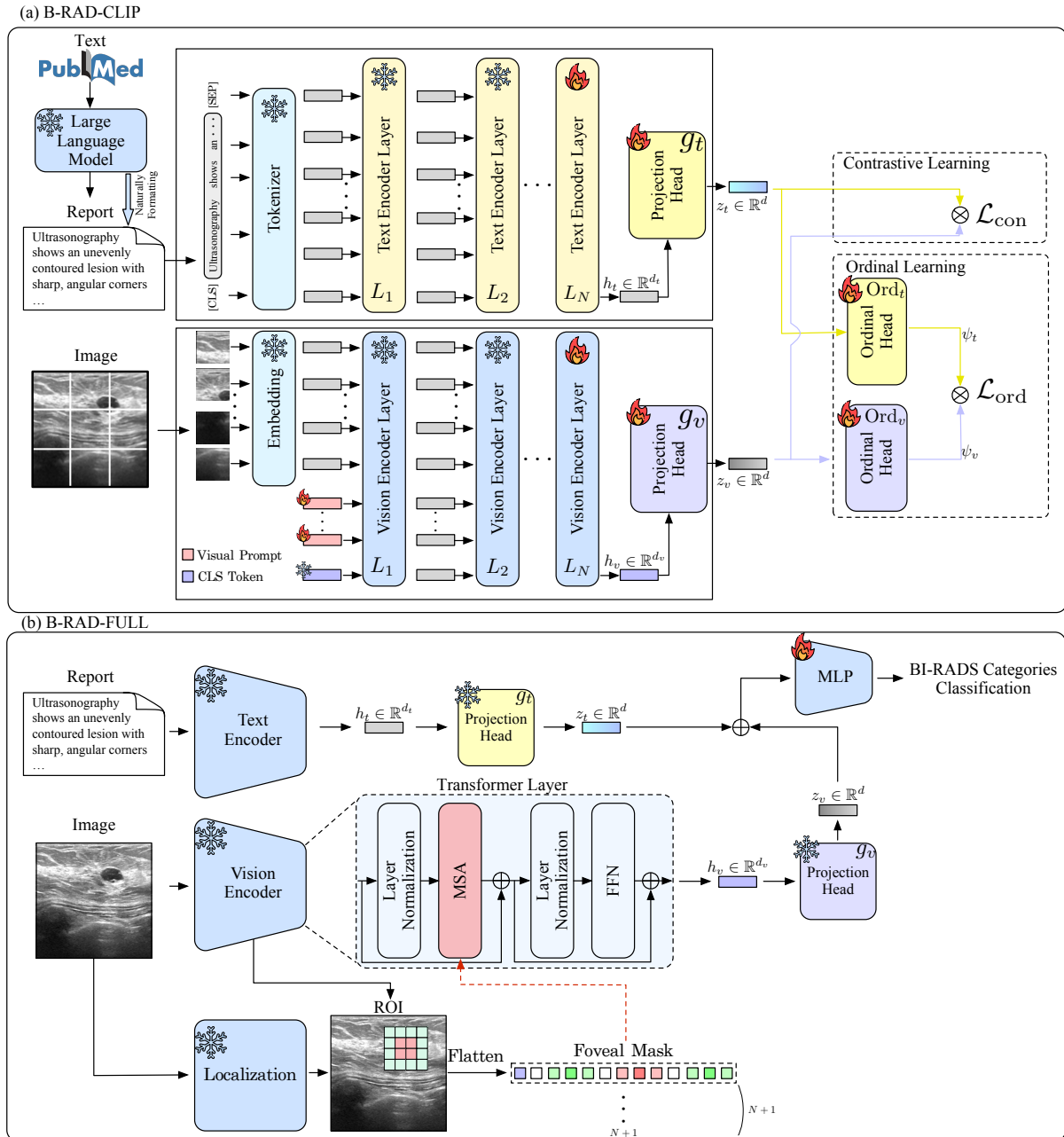
Interpretable multimodal retrieval augmented diagnosis for breast ultrasound
with multinational clinical validation and reader study

Supplementary Figure 1. Data curation flow diagram for the BUS-CoT external validation cohort.

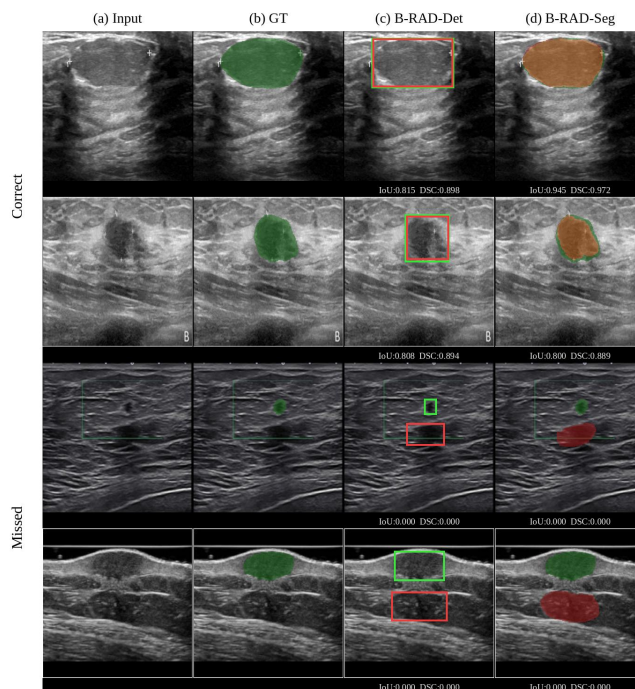
Starting from 11,439 records in the BUS-CoT dataset, sequential filtering retained B-mode ultrasound images with valid BI-RADS 2–5 assessment, single-lesion fields, confirmed pathology (Benign or Malignant), and no overlap with B-RAD training/test sets (BUS-BRA, BUSI, UDIAT, BUS-STU), yielding 3,826 images. An additional expert quality review by a board-certified radiologist excluded 594 cases with annotation inconsistencies, resulting in a final evaluation cohort of $n = 3,232$. BI-RADS subcategories 4A/4B/4C were merged into category 4.



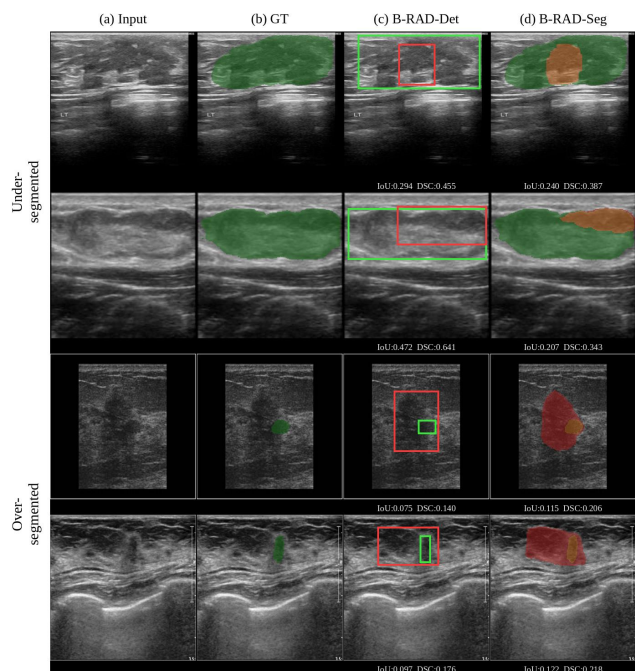
Supplementary Figure 2. Detailed architectural diagram of the B-RAD-CLIP pretraining (a) and B-RAD-FULL inference (b) architecture. (a) The dual-encoder design with Visual Prompt Tuning (VPT) on BiomedCLIP ViT and partial fine-tuning on BioLinkBERT-base, aligned via shared BI-RADS labels through contrastive and ordinal projection heads. (b) B-RAD-FULL inference with foveal attention injection, where the segmentation-derived ROI mask is applied to the Multi-Head Self-Attention (MSA) layers of the frozen vision encoder for morphology-aware classification.



Supplementary Figure 3. Correct and missed detection/segmentation examples on BUS-CoT. Columns: (a) input, (b) GT mask, (c) B-RAD-Det (green = GT box, red = predicted), (d) B-RAD-Seg overlay (green = GT, orange = predicted). Top: correct cases (IoU at least 0.5). Bottom: missed cases (IoU = 0).



Supplementary Figure 4. Under-segmented and over-segmented examples on BUS-CoT. Layout follows Supplementary Figure 3. Top: under-segmented (prediction contained within GT). Bottom: over-segmented (prediction extends beyond GT boundary).



Supplementary Figure 5. Prompt template for MedGemma-27B text corpus standardization. PubMed-derived text is converted to natural language via BI-RADS synonym substitution, then augmented with LLM-generated sonographic detail.

Prompt template for text corpus standardization

System:

You are a breast-ultrasound reporting assistant.

User:

Write ONE concise sentence (20 words or fewer) that adds a sonographic detail not already stated. Follow BI-RADS terminology when possible.

BI-RADS: {birads}

Natural report:
“{natural_text}”

Source excerpt:
“{free_text}”

Hints: architectural distortion, duct changes, skin thickening, skin retraction, edema, internal vascularity, rim vessels, calcifications, elasticity grade, lesion size

Extra sentence (end with a period):

Pre-processing step (rule-based, applied before LLM call):
Raw structured BI-RADS annotations are converted to natural language via deterministic synonym substitution following ACR BI-RADS Atlas lexicon conventions:

oval to egg-shaped	round to ball-shaped
irregular to neither round nor oval	spiculated to star-shaped edges
circumscribed to smooth-edged	indistinct to fuzzy-bordered
hypoechoic to dark-gray	hyperechoic to bright-echo
anechoic to echo-free	heterogeneous to mixed-echo
shadowing to casting an acoustic shadow	enhancement to bright through-transmission
skin thickening to thickened skin	skin retraction to pulled-in skin
not parallel to taller than wide	parallel to wider than tall

Supplementary Figure 6. Prompt template for MedGemma-4B pseudo-report generation at inference time, covering the BI-RADS lexicon descriptors with plain-language substitutions.

Prompt template for pseudo-report generation

System:

You are an expert breast radiologist.

User:

You are an expert breast radiologist writing for a general audience. Describe this breast ultrasound image in exactly ONE flowing, natural sentence.

IMPORTANT RULES:

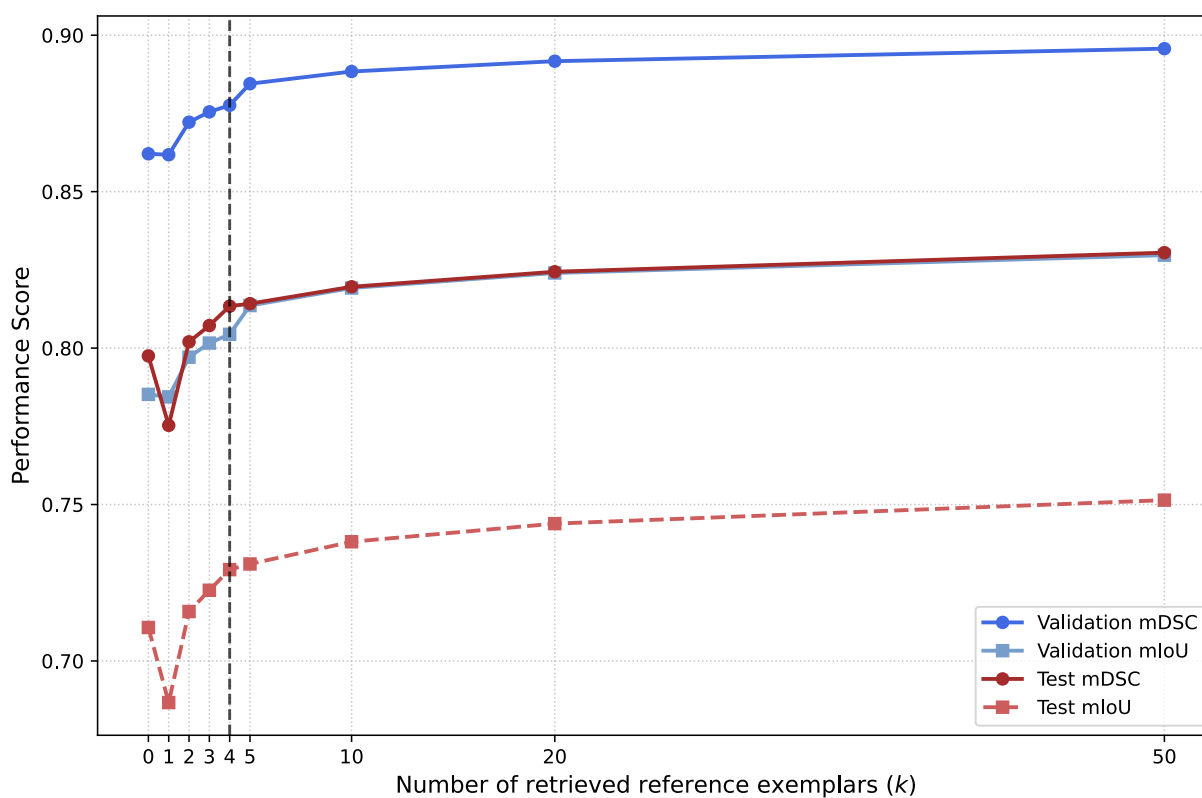
- Do NOT use markdown formatting (no bold, no asterisks, no **).
- Use plain, easy-to-understand English instead of technical BI-RADS jargon.
- Omit any feature that is absent or normal; do NOT write ‘none’.
- Cover: shape, border quality, orientation, echogenicity, posterior acoustics, calcifications (if any), and associated findings (if any).

Plain-language substitutions to follow:

oval to egg-shaped | round to ball-shaped | irregular to neither round nor oval
circumscribed to smooth, sharply defined borders
spiculated to star-shaped edges | indistinct to fuzzy borders
hypoechoic to darker than surrounding tissue | isoechoic to matches surrounding brightness
anechoic to completely echo-free | heterogeneous to mixed echogenicity
not parallel to taller than wide | parallel to wider than tall
shadowing to casting an acoustic shadow behind it
enhancement to bright through-transmission behind it

Now describe the image in one sentence, ending with a period.

Supplementary Figure 7. Detection performance as a function of the number of retrieved reference exemplars (k). Validation and test mDSC and mIoU are shown for $k = 0$ to 50. Performance drops slightly at $k = 1$, recovers at $k = 2$, and plateaus beyond $k = 4$ (dashed line), which was used in all subsequent experiments. The x-axis uses non-uniform spacing to accommodate the sampled values of k .



Supplementary Table 1. Per-fold classification performance of B-RAD-FULL on the BrEaST external test cohort (n = 252). Per-fold metrics are computed independently for each fold's trained model, whereas the main manuscript tables report the 5-fold ensemble. AUROC is computed as weighted one-vs-rest (OvR).

Fold	Accuracy	Weighted F1	AUROC (4-class, weighted OvR)
0	0.9048	0.9035	0.9908
1	0.9325	0.9307	0.9915
2	0.9484	0.9477	0.9943
3	0.9286	0.9275	0.9907
4	0.9286	0.9284	0.9920
Mean $\pm$ SD	0.9286 $\pm$ 0.0156	0.9276 $\pm$ 0.0158	0.9919 $\pm$ 0.0015

Supplementary Table 2. Per-fold classification performance of B-RAD-FULL on the BUS-CoT external cohort (n = 3,232). Per-fold metrics are computed independently for each fold's trained model, whereas the main manuscript tables report the 5-fold ensemble. AUROC is computed as weighted one-vs-rest (OvR).

Fold	Accuracy	Weighted F1	AUROC (4-class, weighted OvR)
0	0.6540	0.6350	0.7280
1	0.6650	0.6480	0.7400
2	0.6580	0.6370	0.7290
3	0.6530	0.6320	0.7340
4	0.6520	0.6310	0.7310
Mean $\pm$ SD	0.6564 $\pm$ 0.0054	0.6366 $\pm$ 0.0066	0.7324 $\pm$ 0.0049

Supplementary Table 3. Per-fold classification performance of B-RAD-FULL on the institutional cohort (n = 289). Per-fold metrics are computed independently for each fold's trained model, whereas the main manuscript tables report the 5-fold ensemble. AUROC is computed as weighted one-vs-rest (OvR).

Fold	Accuracy	Weighted F1	AUROC (4-class, weighted OvR)
0	0.6505	0.6439	0.8093
1	0.5917	0.5850	0.7949
2	0.6159	0.6173	0.7998
3	0.5986	0.6242	0.8065
4	0.5952	0.6132	0.8013
Mean ± SD	0.6104 ± 0.0243	0.6167 ± 0.0213	0.8024 ± 0.0057

Supplementary Table 4. Stage-wise inference latency of the B-RAD-FULL pipeline. All measurements were conducted on a single NVIDIA RTX 3090 with batch size 1. Total latency represents end-to-end processing time from image input to final BI-RADS prediction.

Pipeline Stage	Time per case (s)	Notes
Pseudo-report generation (MedGemma-4B)	1.83	Single image input, max 64 tokens, bf16
B-RAD-CLIP retrieval (5-fold ensemble)	0.01	Embedding + cosine similarity
B-RAD-Det detection (k=4 shot)	0.18	4 reference-based predictions
B-RAD-Seg segmentation (UltraSAM)	0.23	Box-prompted, 1024×1024
B-RAD-FULL (Classification w/ Foveal attention)	0.02	Training-free
Total	2.27	Full pipeline