This MICCAI paper is the Open Access version, provided by the MICCAI Society. It is identical to the accepted version, except for the format and this watermark; the final published version is available on SpringerLink.

Supplementary Material

C. Indicator-Level Recognition Results

C.0 Clinical Indicators and Annotation Protocol

The private dataset includes patient-level annotations for 16 image-based clinical indicators. These indicators were annotated by senior physicians from tertiary hospitals using the available multiparametric ultrasound images for each patient. The annotations are used to construct synthetic CoT supervision and training targets, but they are not provided to the model as input during either training or inference. During both training and inference, the model receives the unified diagnostic prompt and the multiparametric ultrasound images as input.

Table 1. Clinical indicators and annotation targets.

Group	Indicator	Annotation target
Gland	Breast tissue composition	Fatty, heterogeneous echogenic, or fibroglandular tissue.
Gland	Architectural distortion	Presence or absence of distortion in the surrounding architecture.
Gland	Breast duct status	Presence or absence of ductal change.
Gland	Breast skin status	Skin thickening, skin retraction, both findings, or no skin change.
Gland	Breast edema	Presence or absence of breast edema.
Lymph node	Lymph-node shape	Oval, round, or irregular shape.
Lymph node	Lymph-node margin	Clear or unclear margin.
Lymph node	Lymph-node cortical thickening	No thickening, focal cortical thickening, or uniform concentric cortical thickening.
Lymph node	Lymph-node metastasis	No metastasis observed, suspicious metastasis, or high probability of metastasis.
Lesion	Lesion shape	Oval, round, or irregular shape.
Lesion	Lesion orientation	Parallel or non-parallel to the skin.
Lesion	Lesion margin	Clear or unclear, with unclear margins further characterized as indistinct, angular, microlobulated, or spiculated.
Lesion	Lesion echogenicity	Anechoic, isoechoic, hyperechoic, mixed, hypoechoic, or heterogeneous echogenicity.
Lesion	Posterior acoustic features	Enhancement, no change, shadowing, or mixed posterior acoustic change.
Lesion	Lesion calcification	Presence or absence of calcification, including intramass, extramass, intraductal, or coarse calcification when present.
Lesion	Lesion vascularity	Presence or absence of vascularity, with degree and distribution recorded when present.

The source experimental table reports recognition accuracy on the private internal test set. Rows 1-16 correspond to the 16 imaging indicators used by the prompt. Rows 17-18 are diagnostic endpoints and are included for completeness, but they are not included when computing the average recognition accuracy over the 16 imaging indicators.

C.1 Prompt and Reasoning Variants

Table 2. Prompt and reasoning variants. Values are percentages.

Indicator / endpoint	Base	L32-64	L64-128	L128-256	Ind.-wise	Region-wise	Malig.-only	GT text	GT text+img
1. Breast tissue composition	70.0	73.8	75.7	76.1	73.1	74.7	72.4	–	–
2. Architectural distortion	93.9	94.2	93.9	93.9	94.2	93.5	93.9	–	–
3. Breast duct status	90.9	91.6	91.9	91.9	90.6	91.6	91.6	–	–
4. Breast skin status	97.7	98.1	98.1	98.1	98.1	98.1	98.1	–	–
5. Breast edema	99.0	99.0	99.0	99.0	98.7	99.0	99.0	–	–
6. Lymph-node shape	93.5	94.5	95.2	93.5	94.2	93.9	94.2	–	–
7. Lymph-node margin	75.7	75.7	75.7	75.7	75.7	75.7	76.1	–	–
8. Lymph-node cortical thickening	83.8	85.8	86.4	85.1	85.1	84.5	84.1	–	–
9. Lymph-node metastasis	93.2	93.5	92.6	93.2	93.5	93.2	93.9	–	–
10. Lesion shape	82.2	80.9	81.6	78.6	78.0	81.9	80.6	–	–
11. Lesion orientation	89.3	90.0	87.1	91.9	90.9	89.6	87.4	–	–
12. Lesion margin	87.4	85.8	87.4	84.8	82.9	83.2	80.9	–	–
13. Lesion echogenicity	83.2	83.8	83.2	83.2	84.1	84.1	83.5	–	–
14. Posterior acoustic features	72.8	71.8	72.2	74.4	72.2	74.4	70.6	–	–
15. Lesion calcification	90.0	89.6	89.3	88.4	90.0	88.0	88.7	–	–
16. Lesion vascularity	94.5	94.5	94.5	93.9	94.9	94.5	95.2	–	–
17. BI-RADS category	57.3	60.2	58.9	62.1	59.9	61.8	60.2	63.4	67.0
18. Pathology result	92.2	92.9	93.2	94.8	92.9	94.1	93.5	92.6	93.5
Average (indicators 1–16)	87.32	87.66	87.74	87.61	87.26	87.49	86.89	–	–

C.2 GRPO Reward-Signal Variants

This table reports indicator-level recognition accuracy under different GRPO reward-signal settings. These experiments vary one reward signal at a time and should not be interpreted as the final GRPO setting that jointly optimizes all four reward components. All values are percentages.

Table 3. GRPO reward-signal variants.

Indicator / endpoint	BUThink	Gland	Lymph	Lesion	All Region	BI-RADS	Ben./mal.
1. Breast tissue composition	76.0	62.8	36.0	28.4	69.6	25.9	77.9	71.5
2. Architectural distortion	93.0	89.0	91.3	92.4	93.5	92.9	90.3	90.4
3. Breast duct status	92.0	89.0	91.3	92.4	93.5	92.9	91.4	88.4
4. Breast skin status	92.0	93.1	88.4	91.5	91.6	91.6	89.3	81.3
5. Breast edema	99.0	98.8	99.0	98.3	99.0	99.0	99.0	97.2
6. Lymph-node shape	93.9	88.0	92.0	92.7	93.9	94.5	81.2	85.3
7. Lymph-node margin	75.7	59.0	82.0	75.3	75.7	75.7	63.8	61.8
8. Lymph-node cortical thickening	85.8	63.0	86.1	81.0	80.0	80.3	75.4	77.4
9. Lymph-node metastasis	93.5	82.5	94.1	87.5	90.0	86.7	84.2	91.5
10. Lesion shape	80.9	69.0	72.4	74.6	71.2	66.3	75.8	81.3
11. Lesion orientation	90.3	84.2	87.8	88.4	87.7	86.1	84.2	92.4
12. Lesion margin	83.5	77.1	53.7	79.6	65.4	79.9	79.7	80.1
13. Lesion echogenicity	85.4	69.4	78.9	81.5	82.5	82.2	72.6	83.7
14. Posterior acoustic features	91.6	61.2	58.4	64.1	59.2	57.6	83.2	84.8
15. Lesion calcification	75.8	42.0	65.8	81.6	74.1	79.3	77.4	77.1
16. Lesion vascularity	94.8	68.3	79.3	87.8	86.1	74.4	91.8	93.5
17. BI-RADS category	63.4	43.0	52.0	51.3	40.0	47.6	66.4	54.3
18. Pathology result	93.5	78.1	83.6	81.7	88.4	79.3	91.2	96.2
Average (indicators 1–16)	87.70	74.78	78.53	81.07	82.06	79.08	82.33	83.61

D. Qualitative Reasoning Example

Figure 1 shows a representative qualitative example from the private dataset. BUThink first identifies image-based clinical indicators from the multiparametric ultrasound images, then organizes suspicious findings into a stepwise reasoning trace, and finally predicts the patient-level BI-RADS category together with the benign or malignant judgment.

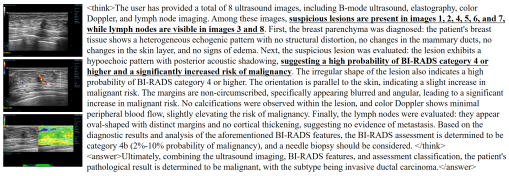


Fig. 1. Qualitative reasoning example on the private dataset.

E. MedThink-Bench-Style Reasoning Evaluation

Because the original MedThink-Bench is designed for text-based medical reasoning rather than image-based breast ultrasound diagnosis, we adapt its scoring-point evaluation protocol to our task. For each patient, the reference reasoning is decomposed into clinically meaningful scoring points, including coverage of the 16 image-based indicators, identification of major suspicious signs, BI-RADS assignment, and benign or malignant judgment. The generated `<think>` content is then evaluated against these scoring points. We evaluate three representative settings: w/o CoT, w/o GRPO, and the full BUThink baseline. The baseline corresponds to the region-wise reasoning setting with the final GRPO stage.

Table 4. MedThink-Bench-style reasoning coverage. All values are percentages, except average reasoning coverage and average covered scoring points.

Metric	w/o CoT	w/o GRPO	BUThink
Average reasoning coverage over all patients	0.58	0.76	0.84
Average covered scoring points	11.6 / 20.0	15.2 / 20.0	16.8 / 20.0
BI-RADS endpoint coverage	58.90	61.80	63.40
Benign/malignant endpoint coverage	86.11	94.17	96.25
16-indicator reasoning coverage	72.50	87.49	87.70

Average reasoning coverage is the mean proportion of covered scoring points across all patients. Average covered scoring points reports the mean number of covered scoring points over the mean total number of reference scoring points. BI-RADS endpoint coverage measures whether the generated reasoning supports the reference BI-RADS endpoint. Benign/malignant endpoint coverage measures whether the generated reasoning supports the reference benign or malignant endpoint. The 16-indicator reasoning coverage measures the proportion of predefined image-based indicators explicitly addressed in the reasoning trace.

F. Code and Prompt Availability

Prompts, key training code, CoT construction rules, GRPO reward implementations, and evaluation scripts are publicly available at `https://github.com/chenqixu925/BUThink-Multi-parametric-Breast-Ultrasound-MLLM`.