BreastGPT: A Multimodal Large Language Model for the Full Spectrum of Breast Cancer Clinical Routine

Anonymous Author(s)

Abstract

Breast cancer remains a leading cause of cancer-related mortality among women. Its clinical management requires multimodal reasoning across a clinical workflow that spans *screening*, *diagnosis* and *treatment planning*, where each stage involves distinct imaging modalities, task objectives, and reasoning patterns. However, constrained by data scarcity and model versatility, existing medical MLLMs are typically evaluated on isolated modalities or narrow task families, limiting their ability to support workflow-level clinical reasoning. In this work, we first introduce **BreastStage**, a workflow-aligned breast imaging instruction corpus comprising 1.86M instruction-following pairs curated from 17 sub-datasets across 5 imaging modalities and 136 task templates. Its held-out split, **BreastStage-Bench**, provides a comprehensive benchmark for evaluating multimodal reasoning across the breast cancer care continuum. Building on this corpus, we propose **BreastGPT**, a unified MLLM equipped with a dual-branch visual encoder and concept-preserving token compression to bridge the scale gap between standard radiology and gigapixel pathology. On BreastStage-Bench, BreastGPT achieves 75.66% closed-ended accuracy and 89.92% open-ended score, outperforming both general-purpose and medical-specific MLLMs across clinical stages and task formats. These results suggest that workflow-aligned data and cross-scale visual modeling are critical for clinically grounded medical MLLMs. All data, code, and model checkpoints are released at `https://anonymous.4open.science/r/BreastGPT`.

1 Introduction

Breast cancer remains one of the most prevalent and life-threatening malignancies among women worldwide [7, 20, 51, 35]. Its clinical management involves not just a single diagnostic decision, but a staged workflow spanning *screening*, *diagnosis*, and *treatment planning*. Each stage relies on distinct imaging evidence and pursues different clinical objectives: *Screening* prioritizes triage at the population level, where mammography, breast ultrasound (BUS) and chest CT are mainly used to detect early suspicious findings and estimate their risk [33]; *Diagnosis* then shifts toward fine-grained lesion characterization, where multimodal MRI is pivotal for evaluating morphology and probability of malignancy [6]; *Treatment planning* requires tumor biology and response assessment, for which gigapixel whole-slide images (WSIs) provide pathological subtyping and biomarker evidence [45]. This clinical continuum imposes a different requirement from conventional single-task image understanding: a clinically useful model must process heterogeneous modalities while preserving the stage-specific reasoning style expected at each point of care.

However, while recent advances in MLLMs have significantly expanded the scope of medical image understanding [43, 15, 67, 64, 68, 32, 3, 13, 40], recent efforts in breast imaging AI are mainly tailored for individual modalities or narrowly defined tasks. Specifically, mammography-oriented models [28, 74] achieve strong performance on mammograph interpretation; BUS-CoT [69] introduces structured reasoning for breast ultrasound images, and MOME [44] supports multimodal MRI analysis with performance comparable to experienced radiologists. Despite demonstrating promising modality-specific performance, these methods conventionally inherit a fragmented formulation of breast cancer

Submitted to 40th Conference on Neural Information Processing Systems (NeurIPS 2026). Do not distribute.

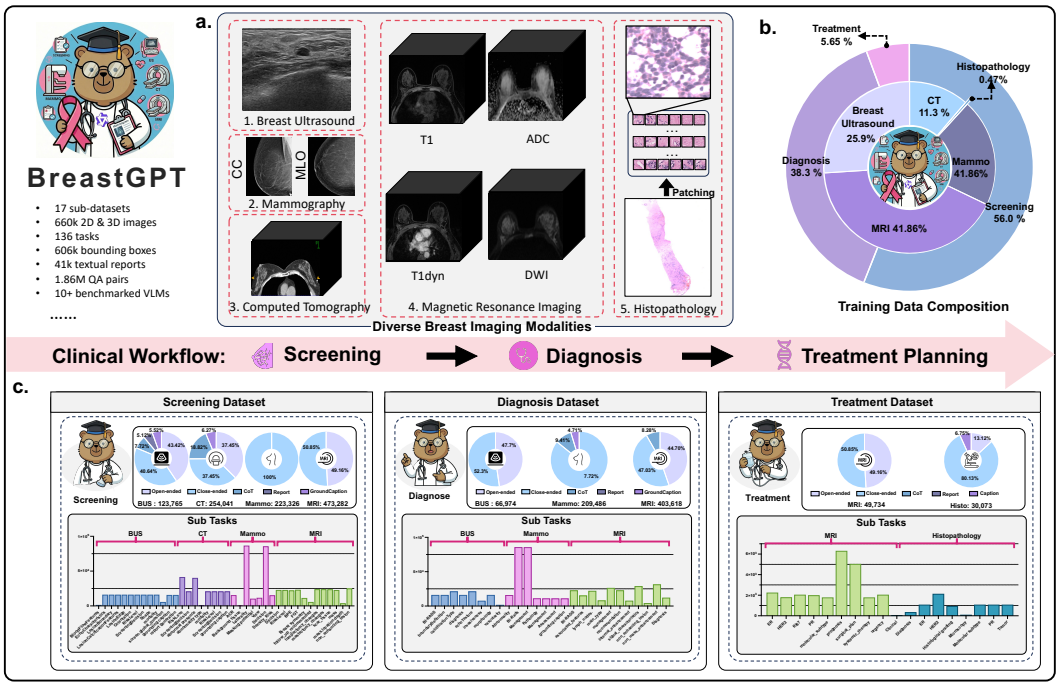


Figure 1: Overview of the **BreastStage** dataset aligned with the end-to-end breast cancer clinical workflow. The dataset is structured across three clinical stages: (1) **Screening** utilizes mammography and breast ultrasound (BUS) images for early lesion detection; (2) **Diagnosis** leverages MRI (main modality), BUS and Mammography for detailed tumor characterization; and (3) **Treatment planning** involves gigapixel whole-slide images (WSI) for pathological subtyping and therapeutic evaluation.

AI: datasets, models, and benchmarks are invariably organized around isolated imaging modalities and limited tasks for a single clinical stage. Consequently, 1) these single-modality datasets fail to comprehensively evaluate models' performance across the various stages within the practical clinical workflow; 2) there is a lack of a versatile model capable of processing different modalities uniformly and efficiently to address various tasks at different stages of real-world clinical workflows.

To this end, we first construct BreastStage (Fig. 1), a large-scale multimodal instruction corpus aligned with the breast cancer clinical workflow. BreastStage integrates **17** sub-datasets across **5** modalities (mammography, BUS, MRI, WSI, and CT), yielding **1.86M** instruction-following QA pairs over **136** task templates. These instructions cover attribute extraction, report generation, visual question answering, and image-grounded captioning, providing supervision over both radiological and pathological evidence. Based on the held-out split of BreastStage, we further build **BreastStage-Bench**, a comprehensive benchmark for workflow-level breast oncology reasoning developed with breast oncology experts. We assess multiple publicly available multimodal large language models, including both general-purpose and medical-specific variants, alongside advanced proprietary models. Extensive experiments demonstrate that BreastStage-Bench poses a significant challenge for current MLLMs: GPT-5.4 achieves only 54.00% closed-ended VQA accuracy and 53.58% open-ended VQA score on average, and existing medical MLLMs show no clear advantage over general-purpose models on these clinical workflow-aligned tasks.

Based on BrasetStage, we propose **BreastGPT**, a unified MLLM for breast cancer workflow reasoning. Built upon Qwen3-VL [9], BreastGPT handles all five imaging modalities within a single architecture and adopts stage-conditioned system prompts to switch between screening, diagnostic, and treatment-oriented reasoning behaviors. Meanwhile, to address the significant differences in image scale across modalities, BreastGPT adopts a dual-branch visual encoder with a resolution-aware gating module that automatically routes each input to an appropriate processing branch based on stage condition. Aligned with practical clinical workflow, BreastGPT demonstrates exceptional versatility and achieves remarkable performance across various tasks. Compared to existing general-purpose and medical-specific MLLMs, BreastGPT achieves performance gains of over 25%, 35%, and 40% for screening, diagnosis, and treatment planning, respectively. To summarize, our contributions are threefold:

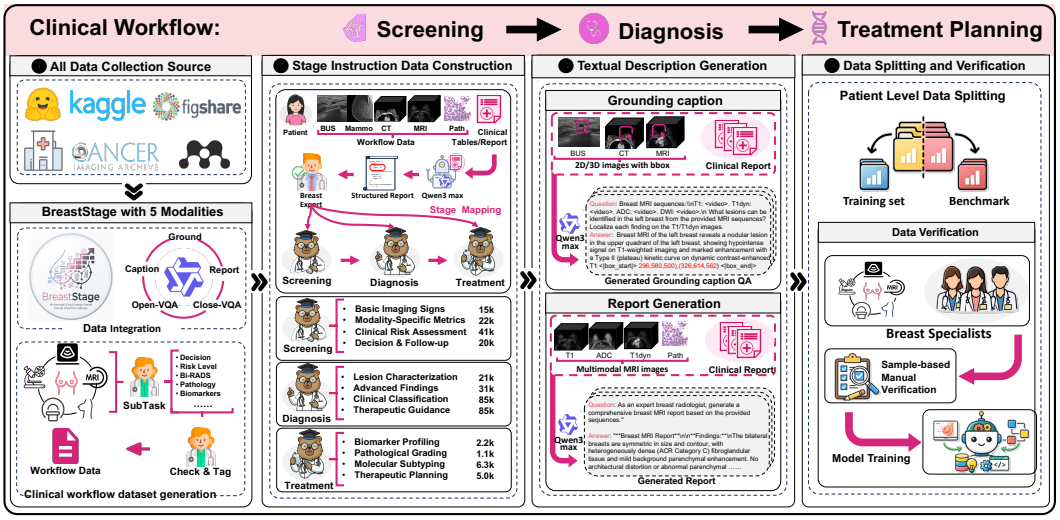


Figure 2: **BreastStage construction pipeline.** It involves curating diverse imaging modalities from public datasets and hospitals. The raw data is then cleaned, annotated, and checked by breast experts. Textual and instruction data are generated by prompting Qwen3-Max. Finally, the data is split into a training set and a benchmark, with professional breast experts assessing the training samples.

Workflow-level problem formulation and benchmark. We formulate breast cancer multimodal reasoning as a workflow-aligned problem spanning screening, diagnosis, and treatment planning. Based on this, we construct BreastStage, a large-scale instruction corpus with 1.86M QA pairs from 17 sub-datasets across 5 modalities, and BreastStage-Bench, a held-out benchmark for evaluating stage-specific clinical reasoning.

Unified cross-scale breast MLLM. We introduce BreastGPT, a unified multimodal LLM that handles mammography, ultrasound, CT, MRI, and gigapixel pathology within a single instruction-following framework, enabling consistent reasoning across the full breast cancer clinical routine.

Concept-preserving visual token compression. To bridge the extreme scale gap between standard radiology and WSIs, we propose a dual-branch visual encoder with a concept-based token selector that preserves clinically salient visual evidence under a fixed token budget, substantially improving both accuracy and inference efficiency.

2 BreastStage Dataset Curation

2.1 Scope and scale

BreastStage contains about 662,000 unique 2D and 3D images, 136 task templates, 606,226 records with bounding-box or mask annotations, and 1.86 million instruction-following pairs. Task coverage spans classification, visual grounding, open- and closed-ended question answering, captioning, and report generation, so the dataset measures both perception accuracy on narrow subtasks and the model's ability to follow diverse clinical instructions. Samples are distributed as 57.9% screening, 36.7% diagnosis, and 5.4% treatment, broadly tracking the relative prevalence and data availability of these stages in real breast care. Please check § C.1 for details.

2.2 Data construction pipeline

As illustrated in Fig. 2, the BreastStage curation pipeline has four steps:

i. Workflow Data Generation. We assemble a cohort of breast imaging studies that draws CT, BUS, mammography, and WSI from public repositories and adds multimodal MRI from two collaborating clinical institutions. For quality control, modality-specific visual specialist agents built on Qwen2.5-VL-72B [10] emit a structured `{validity, reason}` verdict for each image; samples flagged as low-quality are routed to a Low-Image Check in which a breast specialist confirms or restores the verdict, rather than being silently dropped. Lesion-level spatial annotations are then attached: BUS reuses the expert hand-drawn masks shipped with BUS-CoT [69], mammography reuses the EMBED-

released bounding boxes, CT volumes from CT-RATE [22] are passed through DRT-M3D [73] to obtain tumour segmentations, and MRI carries manual T1 and dynamic contrast-enhanced T1dyn annotations from a panel of 10 breast specialists. The resulting cohort is tagged with one of 136 specialist-designed sub-tasks (`task template`) and audited by hand. Full data provenance and per-modality counts are reported in Table 5 and the complete pipeline is detailed in § C.2.

ii. Stage Instruction Data Construction. All data is standardized, cleaned, and paired with its matching clinical tables or reports, then converted into instruction-following QA pairs: *1)* Chinese reports are first translated to English by Qwen3-Max acting as a radiologist agent; *2)* Each translated report is then parsed into a modality-specific and BI-RADS-aligned structured report whose schema is designed by breast experts following clinical reporting guidelines; *3)* For every `task template` from the previous stage, breast experts then write a closed-ended question whose answer is drawn from one or more leaf fields of the structured report, with options taken directly from the schema's enum values, so the resulting QA pairs cannot introduce facts beyond the structured record; *4)* Finally, each `task template` is mapped, with expert guidance, to one of the three clinical stages: screening, diagnosis, or treatment.

After the stage mapping, the resulting BreastStage instruction set is organized to match the cognitive demands of each clinical stage, with category-level counts summarized in Table 6. Screening tasks emphasize early detection and triage, covering modality / view / laterality recognition, breast-tissue characterization (FGT, BPE, density), risk and decision tasks, lesion presence and morphology, and post-surgical change tracking, for a total of 1,075,092 pairs. Diagnosis tasks emphasize tumour characterization and staging, including BI-RADS and pathology classification, lesion-level signal and kinetic characterization, and associated findings such as lymph-node involvement and tissue invasion, for a total of 680,409 pairs. Treatment tasks are grounded in histopathology and treatment-relevant radiology, including prognosis prediction, surgical planning, systemic therapy and biomarker assessment, and WSI-based pathology subtyping, for a total of 100,567 pairs. The hierarchical organization ensures that instruction tuning covers the distinct reasoning patterns required across the patient journey. Please check § C.2 for details.

iii. Textual Description Generation. The same structured records that feed VQA construction are also reused to synthesise two narrative task types: ground captions and report generation: A *ground caption* encodes 2D or 3D spatial coordinates of lesions together with their categories and anatomical locations, providing a multi-dimensional spatial observation for BUS, mammography, CT, and MRI. We construct ground captions by prompting Qwen3-Max to integrate the original clinical report with the available bounding-box annotations into a single visually grounded narrative. *Medical reports*, in contrast, condense the multimodal evidence into key findings, tissue characteristics, and diagnostic or therapeutic assessments. After consultation with senior breast oncologists, we organize each generated report into the four ACR-aligned sections *Findings*, *Impression*, *Final Assessment*, and *Management*. The report generation pipeline is iterative: Qwen3-Max first drafts a comprehensive report from the multimodal input, including multiparametric MRI sequences and histopathology WSIs; breast specialists then audit a sample, catalog the recurring reasoning errors and hallucinations, and we refine the modality-specific system prompts before regenerating.

iv. Data Splitting and Verification. To prevent data leakage while preserving task diversity, we enforce strict patient-level separation between the training set and the evaluation benchmark. We allow the same image to be reused across different task contexts—for instance a BUS image may carry both a BI-RADS label and a lesion-characterization task—but every record derived from a given patient is kept on a single side of the split. The split itself uses stratified sampling on the composite key of modality, task type, and pathology label, so the underlying task and class distributions stay consistent across train and test. Downstream evaluation therefore reflects clinical generalization rather than memorization of shared visual content.

For data reliability we run a multi-stage verification pipeline. Automated heuristic filters first remove hallucinated clinical terms, malformed bounding-box coordinates, instruction-answer conflicts, and near-duplicate QA pairs detected by MinHash similarity above 0.85. Three breast specialists then conduct an independent task-level audit on the held-out test partition (5 random samples per task per specialist), and any task with consistent flags triggers refinement of the generation prompts or filtering rules. **Notably**, the clinical reliability of all generated data has been audited by clinical experts. In § E.3, we report expert agreement on a stratified subset of BreastStage-Bench, including

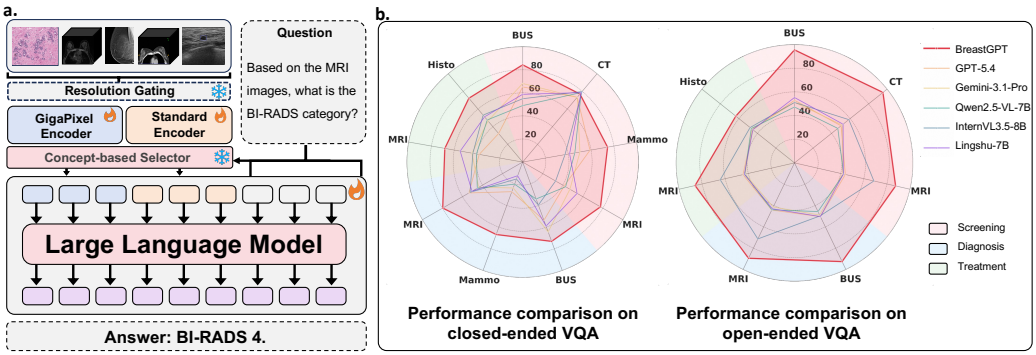


Figure 3: (a) **Architecture of BreastGPT.** A resolution gating mechanism routes standard radiological images to a ViT-based **Standard Branch**, and extreme-resolution WSIs to a specialized **GigaPixel Branch**. The WSI pipeline extracts patch features via a frozen CONCHv1.5 encoder, aggregates global context using LongNet, and compresses the sequence via a universal concept-based token selector before LLM injection. (b) **Performance Comparison.** BreastGPT demonstrates superior reasoning against state-of-the-art MLLMs across both closed- and open-ended clinical QA tasks.

task validity, answer correctness, and clinical consistency. The expert audit covers all modalities and task families, and disagreements are resolved by consensus review.

3 BreastGPT

BreastGPT is built upon Qwen3-VL [9], a vision-language model with native multi-image support, following the broader Qwen-VL lineage of resolution-aware multimodal modeling [8, 61, 10]. We preserve the LLM backbone architecture and adapt it to breast cancer tasks through stage-aware system prompts, which act as lightweight task routers at inference time. This design allows a single unified model to operate across the screening $\rightarrow$ diagnosis $\rightarrow$ treatment workflow without introducing task-specific heads. Compared with fragmented modality-specific systems, such a unified design shares multimodal representations across tasks while maintaining stage-aware specialization. As a result, BreastGPT supports diverse specialist-style reasoning across 136 task templates within a single instruction-following framework. To handle image scales ranging from standard radiological scans to gigapixel histopathology, we design a dual-branch visual processing architecture modulated by a *modality-aware resolution gating* mechanism. Inputs tagged as WSIs are routed to the GigaPixel branch, while all other radiological modalities (CT, MRI, BUS, mammography) flow through the Standard branch, as illustrated in Fig. 3 a.

3.1 Dual-Branch Visual Encoding

Standard Encoder for Radiology. Radiological modalities (CT, MRI, BUS, mammography) are encoded by the native Vision Transformer (ViT) of the Qwen3-VL backbone. This branch is optimized to capture macroscopic anatomical structures, tissue densities, and localized lesion textures that are critical for screening and diagnostic reasoning.

GigaPixel Encoder for Pathology. To integrate extreme-resolution histopathology without overwhelming the LLM context window, we introduce a specialized WSI pathology branch. Prior work on pathology foundation models and WSI transformers has shown that both spatially preserved local morphology and slide-level context are essential for reliable gigapixel image understanding [1, 43, 15, 16, 21, 11, 38, 65]. Pathology information must therefore be incorporated without collapsing tens of thousands of patch-level features into clinically uninformative representations. We separate WSI processing into two steps before compression:

- *Feature Extraction:* WSI tiles at 20$\times$ magnification are encoded by a frozen CONCH v1.5 foundation model [43], producing 512-dimensional patch embeddings $\{p_1, \ldots, p_N\}$, where N can reach tens of thousands. Using a frozen encoder retains robust pathology representations from large-scale pretraining while avoiding the computational instability of end-to-end retraining.
- *Context Aggregation:* A trainable LongNet dilated attention encoder [17] aggregates local cellular details and global architectural patterns across the full patch sequence, producing

contextualized representations $\{h_1, \ldots, h_N\}$. Because diagnostic cues arise from both local cell morphology and broader tissue organization, LongNet's exponentially growing dilation rates preserve long-range context without the $O(N^2)$ bottleneck of standard attention.

Full implementation details (hyperparameters, dilation schedule, projection dimensions) are in § D.2.

3.2 Universal Concept-Based Compression

A unified model spanning 3D radiological volumes and gigapixel pathology faces a severe token-budget mismatch, as a single WSI patch sequence routinely exceeds tens of thousands of tokens. Rather than resorting to modality-specific pooling or heuristic truncation, we adapt the MMTok multimodal coverage maximization framework [18] as a universal concept-based token selector, following recent work on efficient VLM token pruning and compact visual representations [52, 56, 55, 46, 49]. Given either redundant radiological ViT tokens or contextualized WSI embeddings $\mathbf{X} = \{x_i\}_{i=1}^{N}$, the selector distills the input into exactly k tokens by maximizing a concept-coverage objective:

$$\max_{S \subseteq \{1,\ldots,N\}, |S|=k} \underbrace{\frac{1}{M} \sum_{j=1}^{M} \max_{i \in S} \mathrm{sim}(t_j, x_i)}_{\text{text–vision coverage}} + \alpha \underbrace{\frac{1}{N} \sum_{j=1}^{N} \max_{i \in S} \mathrm{sim}(x_j, x_i)}_{\text{vision–vision coverage}}, \tag{1}$$

where $\{t_j\}_{j=1}^{m}$ are text tokens from the clinical instruction and α balances query relevance with visual representativeness. The first term anchors selection to prompt-relevant clinical concepts, while the second term prevents collapse onto redundant salient regions and encourages coverage of global tissue or anatomical structure. This retains visually distinct, clinically informative patterns, enabling BreastGPT to condense multi-scale visual evidence into a fixed token budget without sacrificing the diversity required for downstream reasoning. The complete calibration and greedy optimization procedure are provided in § D.3.

Training. BreastGPT is trained in two stages: a visual front-end warm-up that freezes the LLM and aligns the ViT and CONCH+LongNet branches with the multimodal token space, followed by end-to-end fine-tuning on all four task formats across every modality. The coverage token selector is training-free and applied identically at training and inference. Detailed optimizer settings, and per-stage trainable/frozen module lists are reported in § D.4.

4 Experiments

4.1 Experimental Setups

Baselines. We compare BreastGPT with three groups of representative vision–language models on BreastStage-Bench: *1) proprietary frontier models*: GPT-5.4, Claude-opus-4-6, Claude-sonnet-4-6, Gemini-3.1-Flash, and Gemini-3.1-Pro, queried through their official APIs under identical stage-specific system prompts; these models represent frontier model families commonly used in recent medical MLLM benchmarks [29, 58]; *2) open-source general-purpose VLMs* at comparable scale to BreastGPT: Qwen2.5-VL-Instruct (3B/7B) [10], Qwen3-VL-Instruct (4B/8B) [9], MiMo-VL-SFT [57], and InternVL3.5; *3) medical-specific VLMs*: Lingshu [67] and HuatuoGPT-V [13], with additional context from prior biomedical and medical VLMs [37, 26, 40]. All baselines are evaluated zero-shot on the same instruction-formatted prompts used for BreastGPT, ensuring that differences in performance reflect model capability rather than prompt engineering. Detailed per-baseline configuration is listed in § E.4.

BreastGPT variants. We report two variants that differ only in how the final 128 tokens are obtained to verify the effect of our training-free selection strategy: BreastGPT (cluster) uses the training-free greedy coverage selector, while BreastGPT (learn) replaces it with a learnable cross-attention retrieval head in the spirit of the Perceiver Resampler used in Flamingo [2], trained jointly in Stage 2. Both variants share the same backbone, branches, and training data.

Table 1: VQA performance on BreastStage-Bench. The left half reports **closed-ended accuracy (%)** and the right half reports **open-ended normalized score (%)** for the same set of models. Open-ended evaluation does not include Mammography. Avg is the mean across the task columns of the corresponding half. Best per column in **bold**; BreastGPT rows in cyan.

		Closed-ended VQA (Accuracy, %)										**Open-ended VQA (normalized Score, %)**							
		Screening				Diagnosis			Treatment			Screening			Diagnosis		Treatment		
Model	#P	BUS	CT	Mam	MRI	BUS	Mam	MRI	MRI	His	Avg	BUS	CT	MRI	BUS	MRI	MRI	His	Avg
Proprietary Models																			
GPT-5.4	–	64.89	78.55	68.51	41.43	53.46	55.26	53.50	38.10	32.28	54.00	53.43	49.34	59.85	50.42	59.20	58.07	44.73	53.58
Claude-opus-4-6	–	50.21	72.00	39.57	50.48	38.83	7.66	45.10	41.27	25.94	41.23	43.32	42.77	43.75	42.80	44.42	42.91	40.84	42.97
Claude-sonnet-4-6	–	65.53	68.00	37.02	40.00	54.79	19.67	55.94	42.86	36.30	46.68	43.63	43.12	44.31	43.00	44.98	42.62	41.06	43.25
Gemini-3.1-Flash	–	55.11	42.67	49.68	38.57	37.60	24.32	31.82	48.41	35.81	40.44	48.76	48.21	47.38	45.80	47.53	44.15	42.81	46.38
Gemini-3.1-Pro	–	68.09	73.33	50.21	47.14	64.36	23.87	43.88	44.44	46.53	51.32	51.19	51.46	41.60	48.94	41.39	42.80	45.77	46.16
Open-Source Models																			
Qwen2.5-VL	3B	46.81	51.39	33.51	45.71	30.85	12.76	51.75	44.44	51.52	40.97	49.52	48.39	49.36	47.36	50.54	47.04	44.80	48.14
Qwen2.5-VL	7B	49.15	**79.27**	44.47	44.76	34.31	14.41	46.85	37.30	47.62	44.24	46.20	46.86	49.47	44.99	49.67	44.70	43.98	46.55
Qwen3-VL	4B	52.13	75.39	41.91	49.05	35.37	21.17	48.08	43.65	38.61	45.04	45.20	55.49	47.71	44.93	51.46	48.98	41.26	47.86
Qwen3-VL	8B	57.87	78.55	39.68	51.43	48.14	25.83	47.90	34.92	47.02	47.93	44.94	48.53	44.85	44.34	46.24	44.15	41.21	44.89
MiMo-VL	7B	15.11	23.52	8.62	6.67	14.89	6.16	0.35	0.79	19.98	10.68	42.49	55.98	63.70	41.69	65.66	63.08	39.86	53.21
InternVL3.5	8B	51.91	77.70	35.85	34.76	42.55	16.22	52.27	44.44	52.98	45.41	50.74	58.05	56.56	48.80	58.89	55.94	46.50	53.64
Medical-Specific Models																			
Lingshu	7B	58.94	78.55	39.89	54.29	58.24	8.56	45.28	58.73	51.52	50.44	52.60	49.84	53.41	48.95	53.76	49.60	43.69	50.26
HuatuoGPT-V	7B	45.74	71.39	43.09	39.05	35.11	9.61	47.73	45.24	49.09	42.89	51.86	51.47	55.68	49.20	55.08	52.93	45.78	51.71
Qwen3-VL (SFT)	**8B**	72.34	76.73	65.74	78.10	71.01	59.76	75.87	53.97	60.41	68.21	94.43	**95.38**	94.57	91.70	95.28	88.36	57.96	88.24
BreastGPT (cluster)	**8B**	**86.81**	77.21	**75.00**	**82.86**	77.13	**68.32**	**81.12**	**61.11**	**71.38**	**75.66**	**95.97**	95.29	**95.48**	93.24	**95.72**	**89.93**	**63.80**	**89.92**
BreastGPT (learn)	**8B**	84.47	71.03	68.51	75.71	**78.46**	55.26	73.95	57.14	71.25	70.64	95.86	94.73	87.12	**93.57**	85.22	81.78	63.36	85.95

Note: BUS = Breast Ultrasound, CT = Computed Tomography, Mam = Mammography, His = Histopathology. Mammography is omitted from the right half because it has no open-ended evaluation set.

4.2 Evaluation Metrics

We employ a comprehensive suite of metrics tailored to the diverse task formats within BreastStage-Bench. For closed-ended tasks, we report standard classification accuracy (ACC). For open-ended reasoning and report generation tasks, we report BERTScore F1, BLEU, and ROUGE-1, alongside a weighted aggregate score (0.5 BERTScore F1, 0.25 BLEU, and 0.25 ROUGE-1). We assign a predominant weight to BERTScore F1 because it captures deep semantic equivalence rather than surface-level lexical overlap, a critical requirement for evaluating clinically constrained free-text responses. For visual grounding tasks, we report mean Intersection over Union (IoU); a sample without a ground-truth bounding box is credited as $\text{IoU} = 1$ when the model also abstains from emitting a box (true negative) and 0 when it hallucinates one (false positive). Please check § E.2 for details.

4.3 Evaluation Results

Overall observations. Across Tables 1 and 2, BreastGPT achieves the best score in every clinical stage and every task format. BreastGPT (cluster) reaches 75.66% closed-ended and 89.92% open-ended VQA on average—more than 20 points above the strongest non-SFT baselines in the main cohort (GPT-5.4 at 54.00% closed, InternVL3.5 at 53.64% open)—and raises the MRI report weighted score from 55.16% (GPT-5.4) to 67.67%. On 3D grounding (CT, MRI), BreastGPT is the only model that produces non-trivial volumetric bboxes (Table 13). Eight additional baselines (Grok-4.1-Fast [63], Gemma-4 [59], GLM-4.6V-Flash [60], LLaVA-OneVision-1.5 [5], HealthGPT [40], Hulu-Med [31], MedDr [26], RadFM [62]) are evaluated in Supplementary Tables 10 and 11; none narrows the gap to BreastGPT to under 25 points.

Proprietary frontier models underperform on breast workflows. Despite their general multimodal strength, Claude-opus-4-6, Claude-sonnet-4-6, and Gemini-3.1-Pro stay at 41–51% average on closed-ended VQA, and even the strongest proprietary model (GPT-5.4 at 54.00%) is more than 20 points below BreastGPT. Mammography-based diagnosis (Diag / Mammo) is especially brittle for the Claude family: Claude-opus-4-6 drops to 7.66% and Claude-sonnet-4-6 to 19.67%, while Gemini-3.1-Pro reaches 23.87%. This confirms that breast workflows require domain knowledge (BI-RADS lexicon, molecular subtyping conventions, multiparametric MRI interpretation) that frontier models have not internalized from general web-scale pretraining.

Table 2: Caption and Report Generation Performance (%). IoU = mean grounding IoU on the lesion-grounded captioning task, credited as 1.0 on samples with no ground-truth bbox where the model also refrains from emitting one (true negatives), 0 on hallucinated bboxes; "-" indicates the model produces no usable bbox in that modality. Wtd = weighted composite ($0.5\times$BERT $+$ $0.25\times$BLEU $+$ $0.25\times$R-1). IoU on the 3D modalities (CT, MRI) is reported in Supplementary Table 13. Best results in **bold**. BreastGPT rows highlighted in light blue.

		Caption							Report
		BUS		CT	Mammo		Histo	MRI	MRI
Model	#P	IoU	Wtd	Wtd	IoU	Wtd	Wtd	Wtd	Wtd
Proprietary Models									
GPT-5.4	–	6.68	47.66	43.95	9.08	45.10	44.36	47.00	55.16
Claude-opus-4-6	–	11.41	46.99	43.68	12.06	47.04	44.94	47.75	48.56
Claude-sonnet-4-6	–	7.40	46.60	42.82	–	44.79	44.21	49.30	49.30
Gemini-3.1-Flash	–	56.20	51.48	46.01	7.52	49.29	44.89	51.45	51.03
Gemini-3.1-Pro	–	–	45.60	44.64	–	42.36	45.84	54.86	54.86
Open-Source Models									
Qwen2.5-VL	3B	2.67	46.84	46.88	1.38	48.36	45.71	50.69	49.59
Qwen2.5-VL	7B	6.38	46.79	47.31	1.47	48.11	45.07	51.95	50.71
Qwen3-VL	4B	33.64	46.68	41.66	4.29	46.19	45.38	37.66	49.46
Qwen3-VL	8B	29.35	46.97	47.91	4.06	47.09	46.03	49.57	50.49
MiMo-VL	7B	12.35	43.22	46.14	2.22	43.80	43.98	47.32	46.55
InternVL3.5	8B	0.35	46.97	47.69	0.17	45.72	45.77	49.28	49.12
Medical-Specific Models									
Lingshu	7B	5.00	48.43	49.45	0.55	49.14	46.86	51.84	51.14
HuatuoGPT-V	7B	0.21	47.46	48.92	0.47	48.84	46.24	51.46	51.00
Qwen3-VL (SFT)	**8B**	2.79	76.61	72.72	8.62	76.68	62.42	65.21	66.10
BreastGPT (cluster)	**8B**	**79.59**	**79.32**	73.16	**23.14**	**77.64**	66.78	**67.69**	**67.67**
BreastGPT (learn)	**8B**	–	79.16	**77.07**	–	76.54	**68.11**	–	66.04

Note: IoU = mean grounding IoU including true-negative credit; "-" indicates the model produces no usable bbox in that modality; "–" indicates the metric is not reported for that row. Wtd = weighted composite score. CT and MRI grounding IoU are reported in Supplementary Table 13. BUS = Breast Ultrasound, CT = Computed Tomography, Mammo = Mammography, Histo = Histopathology. Report/MRI = structured radiology report generation for multiparametric MRI.

Medical-specific VLMs offer no clear advantage. Lingshu and HuatuoGPT-V perform comparably to general-purpose 7–8B VLMs on closed-ended VQA (50.44% and 42.89%), and sometimes underperform them. This indicates that existing medical pretraining corpora do not transfer specifically to breast cancer workflow tasks, motivating the need for a workflow-aligned dataset like BreastStage.

Cluster vs. learn variant. BreastGPT (cluster), which uses the training-free greedy coverage selector, edges out BreastGPT (learn) on both closed-ended VQA (75.66% vs. 70.64%) and open-ended VQA (89.92% vs. 85.95%), despite having no additional learnable parameters for token selection. This confirms that the submodular coverage objective is a strong inductive bias for WSI and multimodal token compression, and that a learnable retriever does not trivially outperform a principled, training-free procedure in the breast workflow setting.

4.4 Ablation Studies

We conduct three ablation studies to verify the contribution of individual components.

Necessity of the GigaPixel branch. To isolate the GigaPixel branch, the Qwen3-VL (SFT) row in Tables 1 and 2 reports a controlled baseline sharing the same backbone (Qwen3-VL-8B), training data, and SFT recipe but processing WSIs through the standard ViT branch via 32 randomly sampled 512×512 patches per slide. Across the four radiology modalities, the gap to BreastGPT (cluster) is modest ($\sim$7 points on closed VQA), confirming that random patch sampling is not the bottleneck when the source image already fits a ViT. On histopathology, however, the gap widens sharply: the 32-patch baseline reaches only 60.41% closed-ended accuracy and 62.42 weighted caption score, versus 71.38% and 66.78 for BreastGPT (cluster). Since 32 tiles cover well under 1% of a typical breast WSI, subtype, grade, and treatment-response cues are routinely missed; the CONCHv1.5+LongNet pipeline together with the coverage selector recovers this signal by reasoning over a globally contextualised slide representation. A per-component ablation inside the GigaPixel branch is in Supplementary Table 16.

Visual token budget sensitivity. We sweep $k \in \{1, 8, 16, 32, 64, 128, 256, 512\}$ across five modalities and three task families (full results in Supplementary Tables 17 and 18). *Closed VQA* (Fig. 5a) saturates early: BUS-Diagnosis reaches 77.66% at k=64 and plateaus at 77.13% from k=128, while Mammo/MRI plateau by k=32. *Caption/report* (Fig. 5b) peaks near k=128: radiology gains marginally from k=128 to k=512 ($+0.06$ BUS, $+0.21$ Mammo) while histopathology actually degrades (67.63 at k=8 $\rightarrow$ 64.53 at k=512), suggesting additional WSI tokens dilute rather than add information. *Ground caption* (Fig. 5c) is more token-demanding—a single token cannot represent a

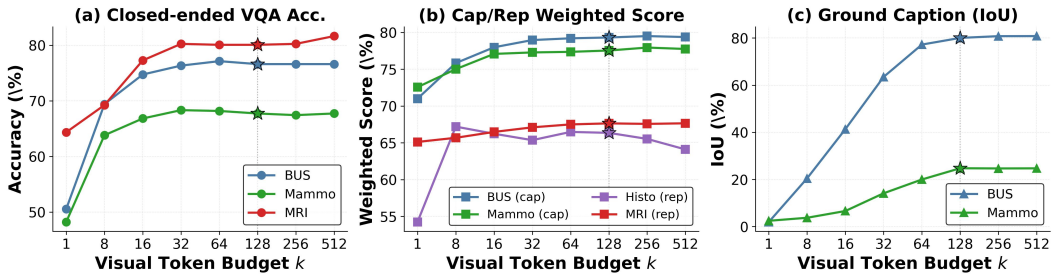


Figure 5: Visual token budget sweep on BreastStage-Bench. **(a)** Closed-ended VQA accuracy on diagnosis-stage tasks for BUS, mammography, and MRI. **(b)** Caption and report weighted score for BUS, mammography, histopathology report, and MRI report. **(c)** Ground caption IoU on BUS and mammography. Dotted vertical lines and stars mark the chosen operating point $k = 128$.

bounding region (IoU 1.50 BUS, 0.85 Mammo at k=1)—but k=128 already recovers >99% of the k=512 IoU. We therefore set k=128 as the unified budget: it sits on the saturation plateau across all task families and keeps radiology and pathology inputs at the same downstream cost.

GPU time. Fig. 4 measures inference latency on one representative histopathology WSI with 5,987 patches. Each k has two bars: the longer bar is the LLM prefill on the selected k tokens, and the shorter bar is the time spent inside the selector itself. At our chosen budget k=128, prefill takes 191.4 ms and the selector adds 9.3 ms, so the total inference latency is 200.6 ms. If we instead skip the selector and feed all 5,987 patch tokens directly to the LLM (the dashed line in the figure), total latency rises to 6.5 s, 33× slower than k=128 on the same slide, and peak memory grows from 16.97 GB to 17.95 GB. Going the other direction, raising k to 512 makes prefill grow to 605.7 ms but improves task quality by less than 1% over k=128. The selector is therefore the right tradeoff in this regime: 9.3 ms of extra compute removes roughly 6.3 s of LLM work that would otherwise dominate the pipeline. Full per-k numbers are in Supplementary Table 15.

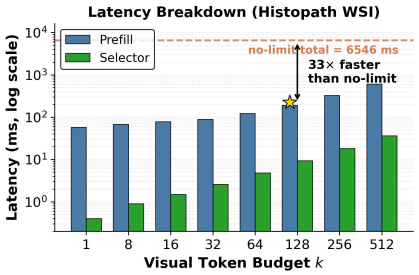


Figure 4: GPU latency on a histopathology WSI. Two bars per k give prefill and selector latency. The dashed line is the no-limit baseline that bypasses the selector; its gap to the chosen-k prefill bar is the latency saved by selection. Log-scaled y-axis.

More results. In § F, we further report per-modality performance (§ F.1) and inference efficiency (§ F.6). In § G, we present a more detailed ablation about WSI branch (§ G.1) and visual token budget sweep (§ G.2). A qualitative analysis and detailed case studies are presented in § H.

5 Conclusion

Breast cancer care is a tightly coupled clinical workflow spanning screening, diagnosis, and treatment, yet existing medical MLLMs are optimised for individual stages and struggle on cross-stage reasoning. We close this gap from both sides: **BreastStage**, a large-scale multimodal instruction corpus and benchmark organised around the screening–diagnosis–treatment pathway, and **BreastGPT**, a unified VLM that pairs stage-aware role prompting with a dual-branch visual encoder for cross-scale modelling from standard radiology to gigapixel pathology. On BreastStage-Bench, BreastGPT substantially closes the cross-stage gap left by both general-purpose and medical-specific MLLMs.

Limitation and Future Work. BreastStage covers all three stages but most samples are drawn from disjoint patient cohorts rather than longitudinal records of the same patient, so the corpus supports workflow-aligned supervision at the task/modality level without fully capturing patient-level temporal continuity. A held-out subset of BreastStage-Bench does contain full-workflow cases from the same patient, allowing partial probing of cross-stage reasoning. Scaling this to a per-patient cross-stage corpus is a field-level bottleneck (linked cross-department records, multi-year follow-up, IRB-approved longitudinal data) and is ongoing work.

References

[1] Faruk Ahmed, Andrew Sellergren, Lin Yang, Shawn Xu, Boris Babenko, Abbi Ward, Niels Olson, Arash Mohtashamian, Yossi Matias, Greg S Corrado, et al. Pathalign: A vision-language model for whole slide images in histopathology. *arXiv preprint arXiv:2406.19578*, 2024.

[2] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736, 2022.

[3] Rawan AlSaad, Alaa Abd-Alrazaq, Sabri Boughorbel, Arfan Ahmed, Max-Antoine Renault, Rafat Damseh, and Javaid Sheikh. Multimodal large language models in health care: applications, challenges, and future outlook. *Journal of medical Internet research*, 26:e59505, 2024.

[4] Asmaa S Alsolami, Wafaa Shalash, Wafaa Alsaggaf, Sawsan Ashoor, Haneen Refaat, and Mohammed Elmogy. King abdulaziz university breast cancer mammogram dataset (kau-bcmd). *Data*, 6(11):111, 2021.

[5] Xiang An, Yin Xie, Kaicheng Yang, Wenkang Zhang, Xiuwei Zhao, Zheng Cheng, Yirui Wang, Songcen Xu, Changrui Chen, Didi Zhu, et al. Llava-onevision-1.5: Fully open framework for democratized multimodal training. *arXiv preprint arXiv:2509.23661*, 2025.

[6] N Aristokli, I Polycarpou, SC Themistocleous, D Sophocleous, and I Mamais. Comparison of the diagnostic performance of magnetic resonance imaging (mri), ultrasound and mammography for detection of breast cancer based on tumor type, breast density and patient's history: A review. *Radiography*, 28(3):848–856, 2022.

[7] Melina Arnold, Eileen Morgan, Harriet Rumgay, Allini Mafra, Deependra Singh, Mathieu Laversanne, Jerome Vignat, Julie R Gralow, Fatima Cardoso, Sabine Siesling, et al. Current and future burden of breast cancer: Global statistics for 2020 and 2040. *The breast*, 66:15–23, 2022.

[8] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023.

[9] Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

[10] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, et al. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

[11] Zak Buzzard, Konstantin Hemker, Nikola Simidjievski, and Mateja Jamnik. Paths: A hierarchical transformer for efficient whole slide image analysis. *arXiv preprint arXiv:2411.18225*, 2024.

[12] Chris Carr, PhD FelipeKitamura, MD, George Partridge, inversion, Jayashree Kalpathy-Cramer, John Mongan, Katherine Andriole, Lavender, Maryam Vazirabad, Michelle Riopel, Robyn Ball, Sohier Dane, and Yan Chen. Rsna screening mammography breast cancer detection. `https://kaggle.com/competitions/rsna-breast-cancer-detection`, 2022. Kaggle.

[13] Junying Chen, Chi Gui, Ruyi Ouyang, Anningzhe Gao, Shunian Chen, Guiming Hardy Chen, Xidong Wang, Ruifei Zhang, Zhenyang Cai, Ke Ji, et al. Huatuogpt-vision, towards injecting medical visual knowledge into multimodal llms at scale. *arXiv preprint arXiv:2406.19280*, 2024.

[14] Pengcheng Chen, Jin Ye, Guoan Wang, Yanjun Li, Zhongying Deng, Wei Li, Tianbin Li, Haodong Duan, Ziyan Huang, Yanzhou Su, et al. Gmai-mmbench: A comprehensive multimodal evaluation benchmark towards general medical ai. *Advances in Neural Information Processing Systems*, 37:94327–94427, 2024.

[15] Richard J Chen, Tong Ding, Ming Y Lu, Drew FK Williamson, Guillaume Jaume, Bowen Chen, Andrew Zhang, Daniel Shao, Andrew H Song, Muhammad Shaban, et al. A general-purpose self-supervised model for computational pathology. *arXiv preprint arXiv:2308.15474*, 2023.

[16] Richard J Chen, Tong Ding, Ming Y Lu, Drew FK Williamson, Guillaume Jaume, Andrew H Song, Bowen Chen, Andrew Zhang, Daniel Shao, Muhammad Shaban, et al. Towards a general-purpose foundation model for computational pathology. *Nature medicine*, 30(3):850–862, 2024.

[17] Jiayu Ding, Shuming Ma, Li Dong, Xingxing Zhang, Shaohan Huang, Wenhui Wang, Nanning Zheng, and Furu Wei. Longnet: Scaling transformers to 1,000,000,000 tokens. *arXiv preprint arXiv:2307.02486*, 2023.

[18] Sixun Dong, Juhua Hu, Mian Zhang, Ming Yin, Yanjie Fu, and Qi Qian. MMTok: Multimodal coverage maximization for efficient inference of VLMs. In *The Fourteenth International Conference on Learning Representations*, 2026.

[19] Haodong Duan, Junming Yang, Yuxuan Qiao, Xinyu Fang, Lin Chen, Yuan Liu, Xiaoyi Dong, Yuhang Zang, Pan Zhang, Jiaqi Wang, et al. Vlmevalkit: An open-source toolkit for evaluating large multi-modality models. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 11198–11201, 2024.

[20] Omar Freihat, David Sipos, and Arpad Kovacs. Global burden and projections of breast cancer incidence and mortality to 2050: a comprehensive analysis of globocan data. *Frontiers in Public Health*, 13:1622954, 2025.

[21] Zhengrui Guo, Qichen Sun, Jiabo Ma, Lishuang Feng, Jinzhuo Wang, and Hao Chen. Context matters: Query-aware dynamic long sequence modeling of gigapixel images. *arXiv preprint arXiv:2501.18984*, 2025.

[22] Ibrahim Ethem Hamamci, Sezgin Er, Chenyu Wang, Furkan Almas, Ayse Gulnihan Simsek, Sevval Nil Esirgun, Irem Dogan, Omer Faruk Durugol, Benjamin Hou, Suprosanna Shit, et al. Generalist foundation models from a multimodal dataset for 3d computed tomography. *Nature Biomedical Engineering*, pages 1–19, 2026.

[23] Jing Hao, Yuxuan Fan, Yanpeng Sun, Kaixin Guo, Lizhuo Lin, Jinrong Yang, Qi Yong H Ai, Lun M Wong, Hao Tang, and Kuo Feng Hung. Towards better dental ai: A multimodal benchmark and instruction dataset for panoramic x-ray analysis. *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

[24] Jing Hao, Yuci Liang, Lizhuo Lin, Yuxuan Fan, Wenkai Zhou, Kaixin Guo, Zanting Ye, Yanpeng Sun, Xinyu Zhang, Yanqi Yang, et al. Oralgpt-omni: A versatile dental multimodal large language model. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

[25] Iryna Hartsock and Ghulam Rasool. Vision-language models for medical report generation and visual question answering: A review. *Frontiers in artificial intelligence*, 7:1430984, 2024.

[26] Sunan He, Yuxiang Nie, Zhixuan Chen, Zhiyuan Cai, Hongmei Wang, Shu Yang, and Hao Chen. Meddr: Diagnosis-guided bootstrapping for large-scale medical vision-language learning. *arXiv e-prints*, pages arXiv–2404, 2024.

[27] Yutao Hu, Tianbin Li, Quanfeng Lu, Wenqi Shao, Junjun He, Yu Qiao, and Ping Luo. Omnimed-vqa: A new large-scale comprehensive evaluation benchmark for medical lvlm. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22170–22183, 2024.

[28] Fuxiang Huang, Jiayi Zhu, Yunfang Yu, Yu Xie, Yuan Guo, Qingcong Kong, Mingxiang Wu, Xinrui Jiang, Shu Yang, Jiabo Ma, et al. A versatile foundation model for ai-enabled mammogram interpretation. *arXiv preprint arXiv:2509.20271*, 2025.

[29] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

[30] Jiwoong J Jeong, Brianna L Vey, Ananth Bhimireddy, Thomas Kim, Thiago Santos, Ramon Correa, Raman Dutt, Marina Mosunjac, Gabriela Oprea-Ilies, Geoffrey Smith, et al. The emory breast imaging dataset (embed): A racially diverse, granular dataset of 3.4 million screening and diagnostic mammographic images. *Radiology: Artificial Intelligence*, 5(1):e220047, 2023.

[31] Songtao Jiang, Yuan Wang, Sibo Song, Tianxiang Hu, Chenyi Zhou, Bin Pu, Yan Zhang, Zhibo Yang, Yang Feng, Joey Tianyi Zhou, et al. Hulu-med: A transparent generalist model towards holistic medical vision-language understanding. *arXiv preprint arXiv:2510.08668*, 2025.

[32] Béria Chingnabé Kalpélbé, Angel Gabriel Adaambiik, and Wei Peng. Vision language models in medicine. *arXiv preprint arXiv:2503.01863*, 2025.

[33] Laskarina Katsika, Eirini Boureka, Ioannis Kalogiannidis, Ioannis Tsakiridis, Ilias Tirodimos, Konstantinos Lallas, Zoi Tsimtsiou, and Themistoklis Dagklis. Screening for breast cancer: a comparative review of guidelines. *Life*, 14(6):777, 2024.

[34] Rana Khaled, Maha Helal, Omar Alfarghaly, Omnia Mokhtar, Abeer Elkorany, Hebatalla El Kassas, and Aly Fahmy. Categorized contrast enhanced mammography dataset for diagnostic and artificial intelligence research. *Scientific data*, 9(1):122, 2022.

[35] Joanne Kim, Andrew Harper, Valerie McCormack, Hyuna Sung, Nehmat Houssami, Eileen Morgan, Miriam Mutebi, Gail Garvey, Isabelle Soerjomataram, and Miranda M Fidler-Benaoudia. Global patterns and trends in breast cancer incidence and mortality across 185 countries. *Nature medicine*, 31(4):1154–1162, 2025.

[36] Rebecca Sawyer Lee, Francisco Gimenez, Assaf Hoogi, Kanae Kawai Miyake, Mia Gorovoy, and Daniel L Rubin. A curated mammography data set for use in computer-aided detection and diagnosis research. *Scientific data*, 4(1):170177, 2017.

[37] Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36:28541–28564, 2023.

[38] Honglin Li, Yunlong Zhang, Pingyi Chen, Zhongyi Shui, Chenglu Zhu, and Lin Yang. Rethinking transformer for long contextual histopathology whole slide image analysis. *Advances in Neural Information Processing Systems*, 37:101498–101528, 2024.

[39] Haoneng Lin, Cheng Xu, and Jing Qin. Taming vision-language models for medical image analysis: A comprehensive review. *arXiv preprint arXiv:2506.18378*, 2025.

[40] Tianwei Lin, Wenqiao Zhang, Sijing Li, Yuqian Yuan, Binhe Yu, Haoyuan Li, Wanggui He, Hao Jiang, Mengze Li, Xiaohui Song, et al. Healthgpt: A medical large vision-language model for unifying comprehension and generation via heterogeneous knowledge adaptation. *arXiv preprint arXiv:2502.09838*, 2025.

[41] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pages 216–233. Springer, 2024.

[42] Kosmia Loizidou, Galateia Skouroumouni, Costas Pitris, and Christos Nikolaou. Digital subtraction of temporally sequential mammograms for improved detection and classification of microcalcifications. *European radiology experimental*, 5(1):40, 2021.

[43] Ming Y Lu, Bowen Chen, Drew FK Williamson, Richard J Chen, Ivy Liang, Tong Ding, Guillaume Jaume, Igor Odintsov, Long Phi Le, Georg Gerber, et al. A visual-language foundation model for computational pathology. *Nature medicine*, 30(3):863–874, 2024.

[44] Luyang Luo, Mingxiang Wu, Mei Li, Yi Xin, Qiong Wang, Varut Vardhanabhuti, Winnie CW Chu, Zhenhui Li, Juan Zhou, Pranav Rajpurkar, et al. A large model for non-invasive and personalized management of breast cancer from multiparametric mri. *Nature communications*, 16(1):3647, 2025.

[45] Christine McCaffrey, Chowdhury Jahangir, Clodagh Murphy, Caoimbhe Burke, William M Gallagher, and Arman Rahman. Artificial intelligence in digital histopathology for predicting patient prognosis and treatment efficacy in breast cancer. *Expert review of molecular diagnostics*, 24(5):363–377, 2024.

[46] Yu Meng, Kaiyuan Li, Chenran Huang, Chen Gao, Xinlei Chen, Yong Li, and Xiaoping Zhang. Plphp: Per-layer per-head vision token pruning for efficient large vision-language models. *arXiv preprint arXiv:2502.14504*, 2025.

[47] Inês C Moreira, Igor Amaral, Inês Domingues, António Cardoso, Maria Joao Cardoso, and Jaime S Cardoso. Inbreast: toward a full-field digital mammographic database. *Academic radiology*, 19(2):236–248, 2012.

[48] Hieu T Nguyen, Ha Q Nguyen, Hieu H Pham, Khanh Lam, Linh T Le, Minh Dao, and Van Vu. Vindr-mammo: A large-scale benchmark dataset for computer-aided diagnosis in full-field digital mammography. *Scientific Data*, 10(1):277, 2023.

[49] Yasmine Omri, Parth Shroff, and Thierry Tambe. Token sequence compression for efficient multimodal computing. *arXiv preprint arXiv:2504.17892*, 2025.

[50] Parita Oza, Urvi Oza, Rajiv Oza, Paawan Sharma, Samir Patel, Pankaj Kumar, and Bakul Gohel. Digital mammography dataset for breast cancer diagnosis research (dmid) with breast mass segmentation analysis. *Biomedical Engineering Letters*, 14(2):317–330, 2024.

[51] Rui Sha, Xiang-meng Kong, Xin-yu Li, and Ya-bing Wang. Global burden of breast cancer and attributable risk factors in 204 countries and territories, from 1990 to 2021: results from the global burden of disease study 2021. *Biomarker Research*, 12(1):87, 2024.

[52] Yuzhang Shang, Mu Cai, Bingxin Xu, Yong Jae Lee, and Yan Yan. Llava-prumerge: Adaptive token reduction for efficient large multimodal models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22857–22867, 2025.

[53] Moein Sorkhei, Yue Liu, Hossein Azizpour, Edward Azavedo, Karin Dembrower, Dimitra Ntoula, Athanasios Zouzos, Fredrik Strand, and Kevin Smith. Csaw-m: An ordinal classification dataset for benchmarking mammographic masking of cancer. *arXiv preprint arXiv:2112.01330*, 2021.

[54] John Suckling. The mammographic images analysis society digital mammogram database. In *Exerpta Medica. International Congress Series, 1994*, volume 1069, pages 375–378, 1994.

[55] Yizheng Sun, Yanze Xin, Hao Li, Jingyuan Sun, Chenghua Lin, and Riza Theresa Batista-Navarro. Lvpruning: An effective yet simple language-guided vision token pruning approach for multi-modal large language models. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 4299–4308, 2025.

[56] Hao Tang and Chengchao Shen. Learning compact vision tokens for efficient large multimodal models. *arXiv*, 2025.

[57] Core Team, Zihao Yue, Zhenru Lin, Yifan Song, Weikun Wang, et al. Mimo-vl technical report, 2025.

[58] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

[59] Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*, 2024.

[60] V Team, Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, et al. Glm-4.5v and glm-4.1v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning, 2025.

[61] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

[62] Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Hui Hui, Yanfeng Wang, and Weidi Xie. Towards generalist foundation model for radiology by leveraging web-scale 2d&3d medical data. *Nature Communications*, 16(1):7866, 2025.

[63] xAI. Grok 4 fast model card. `https://data.x.ai/2025-09-19-grok-4-fast-model-card.pdf`, 2025.

[64] Hanguang Xiao, Feizhong Zhou, Xingyue Liu, Tianqi Liu, Zhipeng Li, Xin Liu, and Xiaoxuan Huang. A comprehensive survey of large language models and multimodal large language models in medicine. *Information Fusion*, 117:102888, 2025.

[65] Conghao Xiong, Hao Chen, and Joseph JY Sung. A survey of pathology foundation model: Progress and future directions. *arXiv preprint arXiv:2504.04045*, 2025.

[66] Hanwen Xu, Naoto Usuyama, Jaspreet Bagga, Sheng Zhang, Rajesh Rao, Tristan Naumann, Cliff Wong, Zelalem Gero, Javier González, Yu Gu, et al. A whole-slide foundation model for digital pathology from real-world data. *Nature*, 630(8015):181–188, 2024.

[67] Weiwen Xu, Hou Pong Chan, Long Li, Mahani Aljunied, Ruifeng Yuan, Jianyu Wang, Chenghao Xiao, Guizhen Chen, Chaoqun Liu, Zhaodonghui Li, et al. Lingshu: A generalist foundation model for unified multimodal medical understanding and reasoning. *arXiv preprint arXiv:2506.07044*, 2025.

[68] Jiarui Ye and Hao Tang. Multimodal large language models for medicine: A comprehensive survey. *arXiv preprint arXiv:2504.21051*, 2025.

[69] Haojun Yu, Youcheng Li, Zihan Niu, Nan Zhang, Xuantong Gong, Huan Li, Zhiying Zou, Haifeng Qi, Zhenxiao Cao, Zijie Lan, Xingjian Yuan, Jiating He, Haokai Zhang, Shengtao Zhang, Zicheng Wang, Dong Wang, Ziwei Zhao, Congying Chen, Yong Wang, Wangyan Qin, and Qingli Zhu. A chain-of-thought reasoning breast ultrasound dataset covering all histopathology categories. *Scientific Data*, 13:236, 2026.

[70] Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*, 2023.

[71] Weihao Yu, Zhengyuan Yang, Lingfeng Ren, Linjie Li, Jianfeng Wang, Kevin Lin, Chung-Ching Lin, Zicheng Liu, Lijuan Wang, and Xinchao Wang. Mm-vet v2: A challenging benchmark to evaluate large multimodal models for integrated capabilities. *arXiv preprint arXiv:2408.00765*, 2024.

[72] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9556–9567, 2024.

[73] Jiaheng Zhou, Wei Fang, Luyuan Xie, Yanfeng Zhou, Lianyan Xu, Minfeng Xu, Ge Yang, and Yuxing Tang. Dual-res tandem mamba-3d: Bilateral breast lesion detection and classification on non-contrast chest CT. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026.

[74] Jiayi Zhu, Fuxiang Huang, Qiong Luo, and Hao Chen. A benchmark for breast cancer screening and diagnosis in mammogram visual question answering. *Nature Communications*, 16(1):11683, 2025.

BreastGPT: A Multimodal Large Language Model for the Full Spectrum of Breast Cancer Clinical Routine

Supplementary Materials

Contents

A Supplementary Overview

This supplementary material provides the technical and procedural details that support the main claims of the paper. We first describe the construction of **BreastStage**, including data provenance, modality coverage, instruction generation, quality control, and task taxonomy. We then provide implementation details for **BreastGPT**, with particular emphasis on the GigaPixel WSI branch and the coverage-maximizing visual token selector that enables a unified token budget across standard radiology images and gigapixel pathology slides. Finally, we describe the BreastStage-Bench evaluation protocol, additional quantitative analyses, qualitative examples, limitations, ethics, and release plans.

This supplementary is intended to make three aspects of the paper auditable: (i) how heterogeneous breast imaging data are mapped into a stage-aware clinical workflow, (ii) how BreastGPT processes image scales that differ by several orders of magnitude, and (iii) how benchmark results, ablations, and error analyses are produced.

A.1 LLM Usage Statement

In this work, we primarily employ LLMs in two aspects: (i) data construction, where Qwen2.5-VL-72B and Qwen3-Max drive the orchestration pipeline described in Section C.2; and (ii) manuscript polishing, where we use LLMs to improve grammar and clarity of the writing without altering the technical claims.

A.2 Ethics Approval and Privacy

BreastStage combines two kinds of sources with different ethical pathways. The BUS, mammography, CT, and histopathology subsets come entirely from public datasets that were de-identified by their original publishers and released under their respective data-use terms; we use them within those terms and do not perform any additional re-identification of individual samples. The MRI subset is the only institutional cohort: it was acquired from two collaborating hospitals [anonymous for review] under IRB approval, with patient consent obtained where required by the contributing institutions.

De-identification of the institutional MRI cohort. Because the MRI cohort is the only data source that originates inside hospitals, our de-identification protocol is scoped specifically to it. Before MRI data leave either source institution, identifying information is removed at three levels:

- **DICOM header level.** Patient names, dates of birth, medical-record numbers, hospital identifiers, referring-physician fields, and exam dates are stripped; absolute timestamps are converted to relative offsets so longitudinal ordering is preserved without disclosing calendar dates.
- **Image content level.** The MRI volumes are reconstructed within a breast-coil field of view that does not include the patient's face, so no additional image-side scrubbing is required.
- **Free-text level.** The accompanying Chinese radiology and pathology reports are passed through the bilingual parsers in Figure 8, which emit a strictly schema-bounded English JSON so that no free-text PHI (referring physician name, hospital department, etc.) propagates into the released instruction-following pairs.

Access control. Access to identifiable raw MRI data is restricted to authorised research personnel under signed data-use agreements. The public mirror of BreastStage will release only the de-identified, instruction-formatted records and the derived bounding-box / mask annotations; raw DICOMs from the institutional MRI cohort are not redistributed. The other four subsets (BUS, mammography, CT, histopathology) are accessed via their original public-dataset distribution channels.

Risks acknowledged. Because the model is trained on clinical imaging data, performance may vary across institutions, scanners, demographic groups, and disease subtypes; we recommend that any downstream use perform site-specific validation before clinical deployment. The benchmark is intended to encourage transparent comparison of breast cancer workflow models, not to replace clinician judgement.

A.3 Additional Limitations

Dataset. While BreastStage covers five imaging modalities, it does not include molecular imaging (PET / CT-PET) or emerging modalities such as photoacoustic imaging. The current workflow also focuses on imaging-centred decision support; longitudinal treatment histories, genomic assays, laboratory results, and medication records are only partially available depending on the source cohort. Future work should expand BreastStage toward longitudinal multimodal patient records and evaluate whether workflow-aligned MLLMs can support temporal reasoning over treatment response, recurrence risk, and survivorship care.

Clinical deployment. BreastGPT is a research prototype and has not undergone clinical validation or regulatory approval. Deployment in a real reading room would additionally require prospective clinical trials, FDA 510(k) clearance or equivalent regulatory approval, integration with existing PACS / RIS systems, and continuous monitoring for model drift and performance degradation. BreastGPT should therefore be interpreted as a foundation model for research and benchmarking, not as an autonomous diagnostic system.

A.4 Compute Resources and Environmental Impact

Model training. BreastGPT training was carried out on 32 NVIDIA H100 GPUs (4 nodes $\times$ 8 GPUs) with DeepSpeed ZeRO-2, FlashAttention, and bfloat16 precision. End-to-end wall-clock training time across the two stages was 3 days, 7 hours, 19 minutes, 53 seconds, corresponding to approximately 2,530 H100-hours.

Dataset construction and evaluation. Visual quality control with Qwen2.5-VL-72B and Qwen3-Max-driven instruction generation are run through external APIs and shared infrastructure; per-call API expenditure was not separately tracked. Specialist annotation of MRI tumours (10 board-certified breast radiologists) and the 3-specialist Bench audit were conducted by clinical collaborators rather than billed compute.

A.5 Code and Data Availability

Code. An anonymised snapshot of the BreastGPT codebase, including training scripts, evaluation pipelines, and the BreastStage data-construction tooling, is available for review at `https://anonymous.4open.science/r/BreastGPT`. The full code release with model checkpoints will be moved to a non-anonymised public repository upon publication.

Data. The BreastStage dataset and BreastStage-Bench evaluation suite will be made publicly available under a Creative Commons BY-NC 4.0 license at `https://[ANONYMOUS]/breaststage`. For components derived from institutional clinical data, release will follow the corresponding IRB protocol, de-identification requirements, and data-use agreements. When raw image release is restricted, we will provide metadata, task definitions, evaluation code, and access instructions.

Model weights. Pre-trained BreastGPT model weights will be released on Hugging Face: `https://huggingface.co/[ANONYMOUS]/BreastGPT`.

A.6 Acknowledgments

We thank the breast oncology experts from the contributing institutions [anonymous for review] for their guidance during dataset construction and clinical validation. We acknowledge the computational resources provided by [anonymous for review] and the funding from [anonymous for review].

B Related Work

Recent medical MLLMs have extended general-purpose vision-language modeling into clinical image understanding, report generation, visual question answering, and interactive decision support [68, 32, 3, 64, 39, 25]. Representative systems such as LLaVA-Med, HuatuoGPT-Vision, MedDr, HealthGPT, and Lingshu show that biomedical instruction tuning can improve clinical-domain perception and

dialogue [37, 13, 26, 40, 67]. Meanwhile, broad evaluation suites such as OmniMedVQA, GMAI-MMBench, MMMU, MMBench, and VLMEvalKit expose the gap between generic multimodal ability and reliable medical reasoning [27, 14, 72, 41, 19]. Domain-specific studies in dentistry further show that clinically meaningful evaluation often requires benchmarks organized around specialty-specific anatomy, modalities, reporting conventions, and treatment workflows [23, 24]. BreastGPT follows this specialty-grounded direction for breast oncology, where the challenge is not only domain-specific perception but also workflow-aware reasoning across screening, diagnosis, and treatment.

Breast imaging AI has progressed rapidly, yet most prior work remains modality-specific. In mammography, VersaMammo builds a large multi-institutional mammogram foundation model and evaluates detection, segmentation, classification, retrieval, and VQA tasks [28], while MammoVQA standardizes mammogram VQA across 15 public datasets and shows that both general-purpose and medical-specific LVLMs remain unreliable for mammogram interpretation [74]. For breast ultrasound, BUS-CoT provides chain-of-thought reasoning annotations across all 99 WHO histopathology categories, connecting observations, imaging features, BI-RADS assessment, pathology labels, and reasoning traces [69]. For MRI, MOME integrates multiparametric breast MRI through a mixture-of-modality-experts design and supports malignancy identification, biopsy recommendation, triple-negative breast cancer classification, and neoadjuvant chemotherapy response prediction [44]. These works show the value of breast-specific modeling, but they do not by themselves provide a single model or benchmark spanning the full clinical pathway.

Computational pathology has similarly moved from patch-level representation learning toward slide-level and vision-language foundation models. Recent pathology foundation models, including UNI, CONCH, and large-scale whole-slide encoders, learn transferable morphology representations for diagnostic, prognostic, retrieval, and report-generation tasks [16, 43, 15, 66]. WSI vision-language and long-sequence modeling efforts such as PathAlign, query-aware gigapixel modeling, hierarchical WSI transformers, and long-context histopathology transformers further bridge patch-level embeddings with slide-level reasoning and language interaction [1, 21, 11, 38, 65]. These studies highlight a central difficulty for BreastGPT: pathology slides contain sparse, distributed, and scale-dependent evidence, while breast radiology images use much smaller visual inputs and different clinical semantics.

Efficient visual token selection is therefore essential for unified breast workflow modeling. Recent methods reduce MLLM inference cost through token pruning, learned token selection, adaptive compression, query-aware retrieval, and layer-wise token reduction [18, 52, 56, 55, 46, 49]. MMTok formulates token selection as multimodal coverage maximization, balancing text-vision alignment with vision-vision diversity [18]. BreastGPT adapts this principle to clinical workflow modeling: for radiology images, the selector removes redundant visual tokens while preserving prompt-relevant lesion and anatomical information; for WSIs, it compresses LongNet-contextualized patch embeddings into a fixed-size representation that retains both query-relevant pathology regions and global tissue coverage.

Overall, existing work establishes strong foundations for medical MLLMs, breast imaging models, pathology foundation models, and efficient token compression, but these directions remain largely separate. BreastGPT integrates them into a unified breast cancer workflow model, and BreastStage-Bench evaluates whether such integration can support stage-aware reasoning across heterogeneous modalities and clinical roles.

C BreastStage Curation Details

C.1 Data Sources and Modality Coverage

As described in Section 2 of the main paper, BreastStage integrates 17 sub-datasets across 5 imaging modalities. The goal of this curation is not merely to aggregate public resources, but to reorganize heterogeneous imaging evidence around the clinical progression of breast cancer care: screening, diagnosis, and treatment. This section details the source datasets, selection criteria, quality control procedures, and annotation protocols used to construct each modality subset.

CT Imaging. Computed tomography volumes are sourced from the CT-RATE dataset [22], which contains 25,692 non-contrast chest CT volumes from 21,304 unique patients. Since breast cancer

predominantly affects women, we first filter the dataset to retain only female patients (sex metadata field equal to "F"). We then deploy Qwen2.5-VL-72B [10] as an automated quality inspector, assigning confidence scores (0–10 scale) based on image quality, anatomical coverage, and the presence of breast tissue in the field of view; volumes scoring below the threshold are excluded. For breast cancer screening and diagnosis, we further pass each volume through DRT-M3D [73] to generate automated tumor segmentations and risk assessments, identifying suspicious breast lesions, calcifications, and lymph node involvement visible in the chest CT field of view. After quality filtering and clinical relevance screening, the CT subset of BreastStage contains 20,546 reconstructed volumes from female patients (Table 5), covering both screening-stage opportunistic detection and diagnosis-stage tumor characterization tasks.

Breast Ultrasound (BUS). BUS images are sourced from the **BUS-CoT** dataset [69], the first chain-of-thought reasoning breast ultrasound dataset that covers *all* 99 histopathology categories defined by the WHO classification of breast tumors. BUS-CoT contains 11,439 images of 10,019 lesions from 4,838 patients across 18 different ultrasound device types, including B-mode US, Doppler US, and Elastography. Images are collected from open-access publications, publicly available case studies (Radiopaedia, PubMed), and open-access biopsy-confirmed repositories, with each case annotated through a rigorous five-level protocol by six senior breast ultrasound radiologists with 8–26 years of clinical experience: (1) *Observation* (lesion presence, calcification presence), (2) *Feature* (boundary, edge, echo characteristics, calcification feature), (3) *Diagnosis* (BI-RADS score), (4) *Pathology* (benign/malignant + WHO histopathology subtype), and (5) *Chain-of-Thought reasoning* explicitly linking imaging features to the final pathological label. Table 3 summarizes the dataset's pathology distribution.

Table 3: Pathology distribution of the BUS-CoT source dataset [69], used as the BUS data source for BreastStage. Numbers in parentheses indicate per-subtype lesion counts.

	Benign (4,856)	**Malignant (4,814)**	**Others (349)**
Categories	Fibroadenoma (1,047)	Invasive ductal carcinoma (896)	Others (349)
	Phyllodes tumour (74)	Invasive lobular carcinoma (75)	
	Intraductal papilloma (72)	Ductal carcinoma in situ (72)	
	Atypical ductal hyperplasia (62)	Mucinous carcinoma (64)	
	Radial scar (56)	Paget disease of the breast (57)	
	Other benign (3,545)	Other malignant (3,650)	

After quality filtering and mapping to our clinical workflow taxonomy (screening-stage BI-RADS triage and diagnosis-stage lesion characterization), the BUS subset of BreastStage contains 10,405 unique image files and 190,730 instruction-following pairs (Table 5), with expert-verified BI-RADS scores, lesion bounding boxes, and chain-of-thought rationales.

Mammography. Digital mammography is sourced from a subset of **MammoVQA** [74], which originally unifies 15 public mammogram datasets. We adopt the 11 datasets that are licensed for derivative works and have stage-aware annotations compatible with our taxonomy (BMCD, CBIS-DDSM, CDD-CESM, CSAW-M, DMID, EMBED, INbreast, KAU-BCMD, MIAS, RSNA, VinDr-Mammo); we then re-organise their labels around the screening / diagnosis / treatment continuum and add new task templates on top of MammoVQA. Table 4 lists every adopted dataset together with its sample count and the task templates it supports. The union of these 11 sources covers the 9 clinically validated MammoVQA task templates—View (CC/MLO), Laterality, BI-RADS, Pathology, Masking potential, Background tissue, Subtlety, Density, and Abnormality detection (calcification, architectural distortion, mass, asymmetry, etc.).

Table 4: Internal mammography source datasets aggregated by MammoVQA [74] and adopted as the BreastStage mammography source. "Type" follows the MammoVQA labelling: *breast* = per-image labels, *finding* = per-bounding-box labels, *exam* = per-examination labels.

Dataset	Type	Size	Supported Tasks
BMCD [42]	breast	400	Density, BI-RADS, Laterality
CBIS-DDSM [36]	breast	3,103	Density, BI-RADS, Laterality, View
	finding	6,464	Abnormality, Pathology, Subtlety
CDD-CESM [34]	breast	1,003	Density, BI-RADS, Laterality, Pathology
DMID [50]	breast	510	Abnormality, Background tissue, Laterality, Pathology, View
	finding	868	Abnormality, Pathology
INbreast [47]	breast	410	Density, Abnormality, BI-RADS, Laterality, View
MIAS [54]	breast	322	Abnormality, Background tissue, Pathology
	finding	234	Abnormality, Pathology
CSAW-M [53]	breast	10,020	Laterality, Masking potential
KAU-BCMD [4]	breast	2,370	BI-RADS
VinDr-Mammo [48]	breast	20,000	Density, Abnormality, BI-RADS, Laterality, View
	finding	4,505	Abnormality
RSNA [12]	breast	54,705	Density, BI-RADS, Laterality, View
EMBED [30]	exam	72,518	Density (exam), BI-RADS (exam)
Internal total	—	**177,432**	—

For BreastStage we use these 11 datasets with patient-level stratification preserved across train and test splits. After mapping the original classification labels to our screening / diagnosis / treatment stage taxonomy and synthesising stage-specific QA pairs, the mammography subset comprises 592,470 unique 2D images and 453,956 instruction-following pairs.

MRI. Multiparametric MRI sequences (T1-weighted, T2-weighted, DWI, DCE-MRI / T1dyn) are acquired from two collaborating clinical institutions [anonymous for review] under IRB approval. Board-certified breast radiologists at the contributing institutions provide expert annotations for tumor segmentation, molecular subtype prediction, and neoadjuvant chemotherapy response assessment. The MRI subset comprises 36,124 unique multiparametric sequence volumes (T1, T1dyn, T2W, DWI, ADC) across screening, diagnosis, and treatment-stage tasks, yielding 926,521 instruction-following pairs.

Whole Slide Imaging (WSI). Histopathology WSIs are sourced from three public repositories—**BCNB**, **TCGA-BRCA**, and **TCGA-HISTAI**—covering invasive ductal carcinoma, invasive lobular carcinoma, ductal carcinoma in situ, and benign / atypical lesions across histological grades I–III. Slides are scanned at $20\times$ or $40\times$ magnification with native resolutions ranging from $80{,}000\times80{,}000$ to $120{,}000\times120{,}000$ pixels. The WSI subset contains 2,510 slides with pathologist-verified molecular subtype labels (ER/PR/HER2 status, Ki-67 index, Nottingham grade), yielding 30,073 instruction-following pairs (captioning, closed- and open-ended VQA, plus prognosis / surgical-planning / systemic-therapy treatment-stage tasks).

Aggregate Statistics. Table 5 summarizes the per-modality data sources adopted by BreastStage. These modality-level counts correspond to the curated image pools before conversion into instruction-following examples. The final instruction corpus expands these image pools into multiple task-specific QA pairs, captions, grounding targets, and report-generation instances according to the stage-aware taxonomy described below.

Table 5: Summary of BreastStage data sources by modality. "Unique 2D / 3D images" counts the distinct image (2D radiology, mammography, histopathology slide) and 3D-volume (CT, MRI multiparametric sequence) files used by the model, after filtering each modality to its retained sub-datasets. Each unique image typically backs many QA pairs.

Modality	Primary Source	Selection Criterion	Unique 2D / 3D Images	Instruction Pairs
CT	CT-RATE [22]	Female-breast volumes, post-QC	20,546 volumes	254,041
BUS	BUS-CoT [69]	All 99 histopathology cats	10,405 images	190,730
Mammo	MammoVQA-derived 11 sets [74]: BMCD, CBIS-DDSM, CDD-CESM, CSAW-M, DMID, EMBED, INbreast, KAU-BCMD, MIAS, RSNA, VinDr	Patient-level stratified, stage-specific re-organisation with new tasks added on top of MammoVQA	592,470 images	454,590
MRI	In-house multi-institution multiparametric MRI	Multiparametric studies with complete sequences	36,124 volumes	926,634
WSI	BCNB / TCGA-BRCA / TCGA-HISTAI	Subtype-labeled, 20–40× slides	2,510 slides	30,073
Total	**17 sub-datasets**	—	**662,055 images**	**1,856,068**

C.2 Instruction Data Generation Pipeline

This section expands the four-stage curation pipeline introduced in Section 2 of the main paper (*Workflow Data Generation → Stage Instruction Data Construction → Textual Description Generation → Data Splitting and Verification*, illustrated in Figure 2). The four stages are powered by a Data Orchestration Engine that uses two LLMs in complementary roles: **Qwen2.5-VL-72B** [10] is used wherever a decision requires looking at the image (modality-specific quality control, visual-attribute extraction), and **Qwen3-Max** is used for every text-only transformation (Chinese-to-English report parsing, structured-record-to-open-ended-VQA rewriting, ground-caption synthesis, report generation). Both LLMs are queried with strict JSON output schemas so downstream stages consume structured outputs without bespoke parsing. Per-modality QC prompts and the Chinese-to-English parsers appear in Figure 8, and the stage-aware persona library together with the system-prompt assembly rule appear in Figure 9. The remainder of this section walks through the four stages in one-to-one correspondence with the main-paper subsections.

Workflow Data Generation. The cohort consists of public BUS, mammography, histopathology, and CT data plus an in-house multi-institution MRI dataset acquired from two collaborating hospitals [anonymous for review] under IRB approval; identifying information is removed at the source institutions before any data leaves the hospitals (Section A.2). On top of the raw studies, this stage produces three artefacts that the rest of the pipeline relies on: per-image quality scores, lesion-level bounding-box / mask annotations, and stage-aware task labels.

Modality-specific quality control. The visual specialist agents from Section 2 are implemented as Qwen2.5-VL-72B prompted with modality-specific quality-control rules (Fig. 8). The BUS selector rejects convex / sector-probe images outside the breast-ultrasound depth range; the mammography selector strictly rejects chest X-rays and applies PGMI positioning checks plus Eklund implant-displaced views; the breast MRI selector requires bilateral-coil framing or sagittal / axial single-breast views and verifies fat-saturation; the CT selector requires breast-inclusive FOV. Histopathology WSIs are sourced from a curated public collection where modality and magnification are already standardised, so they bypass this LLM-driven QC step. The selector emits a JSON `{validity, reason}`; samples flagged as low-quality are not silently dropped but routed to a *Low-Image Check* step in which a breast specialist reviews the image and the LLM's stated reason, and either confirms the rejection or restores the sample.

Bounding-box and mask generation. The four imaging modalities (BUS, mammography, CT, MRI) carry lesion-level spatial annotations, with the source differing by modality. BUS uses the expert hand-drawn lesion masks released with the BUS-CoT dataset [69]. Mammography (EMBED subset) ships with official lesion bounding boxes; we re-format the released `[ymin, xmin, ymax, xmax]` tuples into the `[xmin, ymin, xmax, ymax]` convention used throughout BreastStage, and no mask step is performed on this modality. CT lesion masks come from DRT-M3D [73], which also emits a per-volume cancer-risk score that we retain in the record metadata for the screening-stage CT tasks. MRI lesion masks are drawn manually on T1 and T1dyn sequences by 10 board-certified breast

specialists at the contributing hospitals. For the mask-based modalities (BUS, CT, MRI), bounding boxes are derived by connected-component analysis: components with fewer than 50 voxels are discarded as noise, and the tight axis-aligned envelope of every retained component is emitted as a 4-tuple `[xmin, ymin, xmax, ymax]` for BUS or a 6-tuple `[xmin, ymin, zmin, xmax, ymax, zmax]` for the 3D volumes (CT, MRI). Histopathology WSIs do not carry pixel-level masks or bounding boxes in BreastStage; they are routed only to the closed/open VQA and report-generation tasks, not to the ground-caption task.

Expert-driven sub-task labelling. After QC and mask generation, the per-dataset clinical tables and reports are reviewed by breast specialists who tag each record with one of **136 sub-task labels** reflecting BI-RADS items, mass / calcification descriptors, biomarker fields, and treatment-stage decisions. Each label is a `task template`: a high-level sub-task such as "mass margin description" or "HER2 status", typically spanning several leaf fields of the structured report defined in the next stage. Per-modality counts after this stage are reported in Supplementary Table 5, and the resulting 136-task taxonomy in Table 6.

Stage Instruction Data Construction. This stage converts each curated study into a schema-bounded structured record, so that no free-text generation step downstream is allowed to introduce clinical content beyond what this record already encodes. The flow follows the four steps introduced in the main paper. (i) **Report translation.** Chinese radiology and pathology reports are translated to English by Qwen3-Max acting as a radiologist agent, using the bilingual prompts in Figure 8. (ii) **Structured report extraction.** Every translated report is then parsed by the radiologist agent into an expert-designed BI-RADS-aligned structured report whose schema is written per modality by breast specialists following clinical reporting guidelines. For example, the MRI structured report (Figure 7) covers per-side breast-level findings (FGT, BPE, post-surgical changes, lymph nodes, BI-RADS, management) plus per-lesion descriptors for masses, non-mass enhancement, and non-enhancing lesions. (iii) **Question bank derivation.** For each `task template` from Stage 1, breast specialists then write a closed-ended question that draws its answer from one or more leaf fields of the structured report, with options taken directly from the schema's enum values; this constrains BreastStage to questions that the structured record already answers. The resulting tuples carry the form `<Task, Question, options, Answer, Images>`. (iv) **Task–Stage mapping.** Specialists finally map each `task template` to one of the three clinical stages (*screening* / *diagnosis* / *treatment*), which is also the key used for task-key balanced sampling on BreastStage-Bench; the per-stage record counts of the final task taxonomy are reported in Table 6.

From the structured records, BreastStage derives two parallel VQA partitions, both grounded in the same schema-bounded source. The *closed-ended* partition packs the schema's enum values as an option list into the question text, so each tuple takes the form `<Task, Question+options, Answer, Images>` with the answer being one of the options. The *open-ended* partition takes the same structured tuple, drops the option list, and asks Qwen3-Max to rewrite the categorical answer into a fluent clinical description, yielding `<Task, Question, Answer, Images>` where the answer is free-form text but the underlying clinical fact is unchanged; the rewriting prompt is reproduced in Fig. 6. Because both partitions are pinned to the same structured record, the LLM cannot fabricate clinical content beyond the schema, and the answer cannot leak through the question text into a text-only shortcut — a property reflected in the $k{=}1$ collapse of BreastGPT's closed-ended accuracy in Fig. 5a, where the model degrades sharply once the visual evidence is removed.

Open-VQA Generator

You are a medical imaging data assistant specialising in breast `{modality}`. Your task is to rewrite a structured attribute into a natural-language Question–Answer pair suitable for a Breast `{modality}` VQA dataset.

Rules.

1. The `Question` *must* be open-ended and descriptive: do *not* ask yes/no questions, do *not* include options or binary choices.
2. The `Answer` *must* be a complete sentence that is strictly based on the provided `Answer` value, and does *not* introduce any new medical interpretation, diagnosis, or inference.
3. Do *not* introduce information not explicitly contained in the input.
4. Do *not* repeat or list the provided options in the output.
5. Use varied sentence structures across different samples.
6. Use precise anatomical and imaging terminology appropriate for breast `{modality}`.
7. Output *only* valid JSON in the specified format. No extra text.

Input (JSON): `{"Question": "...", "options": ["...","...","..."], "Answer": "..."}`.
Output (JSON): `{"Question": "...", "Answer": "..."}`.

Grounded Caption Generator

You are an **expert Radiologist Assistant** producing a **Grounded Caption** for breast `{modality}`. Strictly separate *visual attributes* (shape, margin, echo / signal / density, posterior or kinetic features) from *clinical context* (BI-RADS, risk, diagnosis, management).

Step 1 – Reference phrase. For each finding, write a noun phrase that names *only what is visible inside the bounding box*: include side, view (when applicable), abnormality type, and the relevant visual descriptors. Exclude BI-RADS, risk, diagnosis, and management.

Step 2 – Caption skeleton. Compose a professional radiology sentence with placeholder tokens `<ref-object><bbox>` immediately after each lesion phrase; inject the excluded clinical context (BI-RADS, risk, recommendation) *around* the placeholders, never inside the reference phrase.

Step 3 – Strict alignment. Emit `ref[]`, `bbox[]` as flat lists in caption order; `ref[i]` corresponds to `bbox[i]` and to the i-th `<ref-object><bbox>` token. Boxes are 2D `[xmin,ymin,xmax,ymax]` for BUS / mammography / histopathology and 3D `[xmin,ymin,zmin,xmax,ymax,zmax]` for CT / MRI.

Output JSON: `{caption, objects:{ref, bbox}}`.

Report Generator

You are a **board-certified breast radiologist**. Generate a formal breast `{modality}` report from the structured `patient_data`, compliant with the **ACR BI-RADS Lexicon** for that modality and reflecting real-world dictation style.

Hard safety rules. (1) No fabrication of laterality, lesion size, clock-face position, or distance to nipple unless explicitly present in `patient_data`; if missing, refer generically ("the lesion"). (2) No histological / molecular confirmation: replace "X is invasive carcinoma" with probabilistic phrasing ("findings are suspicious for X"). (3) Bounding boxes are internal; never mention coordinates, "bbox", or any data-structure terminology in the prose.

Sections (fixed order). *Findings*: one passive-voice paragraph per lesion, ordered by descending BI-RADS, integrating only descriptors present in `patient_data`. *Impression*: probabilistic synthesis. *Final Assessment*: the highest BI-RADS category among findings. *Management*: ACR-aligned recommendation (Cat 2 – routine; Cat 3 – short-interval follow-up; Cat 4–5 – biopsy / tissue diagnosis). Surgical / systemic therapy decisions are out of scope.

Output: four-section markdown with the section names above.

Figure 6: Generator prompts that produce the BreastStage instruction-following partitions from each structured record: the open-ended VQA generator (top), the grounded caption generator (middle), and the report generator (bottom). All three are instantiated per modality with `{modality}` $\in$ {BUS, mammography, CT, MRI, histopathology}.

MRI Structured Report Template

```
{
  "breast_level": {
    "FGT": {"left": "C", "right": "C"}, // "A","B","C","D"
    "BPE": {"left": "Moderate", "right": "Mild"}, // "Minimal","Mild","Moderate","Marked"
    "breast_symmetry": true,
    "post_surgical_changes": {
      "left": {"scar_tissue": false, "lumpectomy_changes": false, "mastectomy_changes": false},
      "right": {"scar_tissue": false, "lumpectomy_changes": false, "mastectomy_changes": false}
    },
    "associated_features": {
      "left": {
        "nipple_retraction": null, "nipple_invasion": null,
        "skin_retraction": null, "skin_thickening": null,
        "skin_invasion": {"direct_invasion": null, "inflammatory_cancer": null},
        "axillary_adenopathy": null, "pectoralis_muscle_invasion": null,
        "chest_wall_invasion": null, "architectural_distortion": null,
        "ductal_dilation": null // "mild","marked",null
      },
      "right": { ... same schema as "left" ... }
    },
    "lymph_nodes": {
      "axillary_left": null, "axillary_right": null, // "normal","abnormal"
      "internal_mammary_left": null, "internal_mammary_right": null
    },
    "BI-RADS": {"left": null, "right": null}, // "0","1","2","3","4A","4B","4C","5","6"
    "management":{"follow_up_recommended": null, "biopsy_recommended": null}
  },
  "mass": [
    {
      "id": 1, "side": "left",
      "location": {"quadrant": "UOQ", // "UIQ","UOQ","LIQ","LOQ","central",null
                   "depth": "middle"}, // "anterior","middle","posterior",null
      "size_mm": {"long": 12, "short": 9, "depth": 8},
      "signal_characteristics": {
        "T1W": "hyperintense", "T2W": "hyperintense", // "hyperintense","isointense","hypointense",
              null
        "DWI_restriction": "present", // "present","absent",null
        "ADC": "hyperintense",
        "kinetics": {"initial": "Fast", // "Slow","Medium","Fast",null
                     "delayed": "Wash-out"} // "Persistent","Plateau","Wash-out",null
      },
      "morphology": {"shape": "oval", // "round","oval","lobular","irregular",null
                     "margin": "circumscribed"}, // "circumscribed","irregular","spiculated",null
      "internal_enhancement": {
        "pattern": "heterogeneous", // "homogeneous","heterogeneous","rim_enhancement","
              dark_internal_septations",null
        "associated_vascularity": {"adjacent_vessel_sign": null,
                                   "increased_peritumoral_vascularity": null}
      }
    }
  ],
  "non_mass_enhancement": [
    {
      "lesion_id": 1, "side": "left",
      "location": {"quadrant": null, "depth": null},
      "size_mm": {"long": null, "short": null, "depth": null},
      "distribution": "Segmental", // "Focal","Linear","Regional","Segmental","Multiple regions","
            Diffuse",null
      "internal_enhancement_pattern": "Clustered_ring", // "Homogeneous","Heterogeneous","Clustered","
            Clustered_ring",null
      "signal_characteristics": { ... same as "mass" ... }
    }
  ],
  "non_enhancing_lesion": [
    {
      "lesion_id": 1, "side": "left",
      "location": {"quadrant": null, "depth": null},
      "size_mm": {"long": null, "short": null, "depth": null},
      "signal_characteristics": { ... same as "mass" ... },
      "lesion_type_detail": "ductal_precontrast_high_signal_on_T1W"
                            // "cyst","non_enhancing_mass","architectural_distortion",null
    }
  ]
}
```

Figure 7: Expert-designed BI-RADS-aligned structured template used for the MRI cohort. Every leaf field is either a numeric measurement, a Boolean flag, or one of a fixed enum list (shown as a comment). Reports are first translated to English by Qwen3-Max and then populated into this template under its enum and type constraints; specialists then derive a question per `task template` whose answer is drawn from one or more leaf fields. Schema-conformant population guarantees that the resulting QA pairs cannot introduce facts that the source report does not already record.

Textual Description Generation. The same structured records that feed VQA construction are also reused to synthesise the two narrative task formats, ground caption and report generation, with the modality-parameterised prompts in Figure 6. Both formats remain pinned to the structured record, so Qwen3-Max is responsible only for the surface form and never for the underlying clinical fact.

Ground caption. Ground captions are produced only for the four imaging modalities that carry bounding boxes (BUS, mammography, CT, MRI); histopathology, which has no boxes, is excluded. Each caption is generated by binding every visible lesion to its box and weaving the resulting visually grounded phrases into a clinical narrative; the surrounding context drawn from the structured record (BI-RADS, risk level, recommendation) is allowed to appear around the boxes but never inside the lesion phrase, so a reader can recover, from the caption alone, which descriptors apply to which spatial region. The modality-specific vocabularies differ in the obvious places (posterior acoustic features for BUS, ACR density for mammography, FGT, BPE and dynamic kinetic curves for MRI), but the visual-vs-context separation is identical across modalities.

Report generation. The report generator targets all five modalities, with one prompt variant per modality so that the output respects the relevant ACR BI-RADS lexicon. Every report is forced into the same four sections (Findings, Impression, Final Assessment, Management) so that downstream evaluation can score them section by section. Three constraints prevent the failure modes typical of pure-LLM report generators: spatial details that are not in the structured input (laterality, lesion size, clock-face position, distance to nipple) cannot appear, source pathology labels must remain probabilistic in the prose ("findings are suspicious for invasive carcinoma" rather than "this is invasive carcinoma"), and any reference to internal field names or coordinates is forbidden so the report reads as clinical dictation. Drafts then flow into the specialist-audit loop in Stage 4 below, where modality-specific prompts are revised and failing records regenerated whenever a per-task flag rate exceeds threshold.

Data Splitting and Verification. Records are split between training and BreastStage-Bench at patient-id granularity; because the source datasets cover disjoint patient populations, no cross-dataset patient deduplication is required. The split itself uses a single global stratified sampler over the composite (modality, task type, pathology label) key, so the resulting train and test partitions match on the underlying task and class distributions. BreastStage-Bench equals the test partition, drawn with task-key balanced sampling so that every task template in Table 6 is represented. Verification has two layers. Automated heuristics first remove records with malformed bounding-box coordinates, invalid option labels, instruction–answer conflicts, hallucinated clinical terms, and near-duplicate QA pairs (MinHash > 0.85). Three breast specialists then conduct an independent task-level audit on Bench: each specialist reviews 5 random samples per task and flags clinically inconsistent or stage-mismatched cases. Any task whose flag rate is non-negligible across specialists triggers re-running the relevant LLM stage with the failing prompts revised.

BUS Quality Selector

You are a **Breast Ultrasound Quality Specialist**. Your goal is to validate if the image is likely a **Breast/Axilla Ultrasound**. Evaluate the following:

1. Modality & Probe Geometry (BREAST FOCUS):

- *Linear Probe (Rectangular Field):* high probability of valid Breast US.
- *Convex/Sector Probe (Fan-shaped Field):* **WARNING** – typical for Abdomen/Liver. Mark Valid only if the text explicitly confirms a deep breast mass.
- *Modality:* confirm Ultrasound (speckle texture).

2. Image Quality & Settings:

- *Depth:* $<$ 4–6 cm; reject if deep abdominal structures (liver / kidney) are visible.
- *Focus zone:* aligned with the mammary layer.
- *Gain:* fat / glandular interface distinguishable.

3. Text–Image Alignment: flag if the text mentions "Liver", "Gallbladder", or "Thyroid".

Output JSON: {validity: Valid/Invalid, reason}.

Mammography Quality Selector

You are a **Mammography Quality Control Expert**. Validate if the image is a diagnostic **Mammogram (MG)**. Evaluate the following:

1. Modality Distinction (CRITICAL):

- Confirm Mammogram: soft-tissue breast against black air background.
- **STRICTLY REJECT** Chest X-Ray (CXR): if you see ribs, lungs, clavicles, or the spine, mark Invalid.

2. Breast-Specific Positioning (PGMI Criteria):

- *Pectoralis muscle:* for MLO views, must extend to the nipple line.
- *Nipple profile:* in profile (not superimposed) unless retracted.
- *Skin folds:* reject severe folds that mimic abnormalities.
- *Implants:* identify "Implant Displaced (Eklund)" views vs. standard views.

3. Exposure & Contrast: adipose (dark gray) vs. fibroglandular tissue (white) must be distinguishable; reject motion blur at calcifications.

4. Text–Image Alignment: text describes mammographic findings (asymmetry, distortion, calcification).

Output JSON: {validity, reason}.

Breast MRI Quality Selector

You are a **Breast MRI Physics Specialist**. Validate if the image is a diagnostic **Breast MRI** sequence. Evaluate the following:

1. Coil & Field of View (BREAST FOCUS):

- *Breast coil geometry:* bilateral breasts hanging in the coil, or a dedicated sagittal/axial single-breast view.
- *Exclude:* Brain MRI, Spine MRI, or Abdominal MRI where the breast is incidental and compressed.
- *FOV:* must include the axillary tail region.

2. Sequence & Physics:

- *Fat Saturation:* **CRITICAL** for breast – fat must be suppressed (dark) except in T1 non-fat-sat anatomy scans.
- *Cardiac motion artefacts:* check the left-breast phase-encoding direction; severe pulsation rendering the medial breast unreadable $\rightarrow$ Invalid.
- *Silicone select:* for implant cases, confirm silicone-specific sequences are used.

3. Text–Image Alignment: text describes "enhancement kinetics", "BPE", or "washout"; flag mismatch if it describes CT density.

Output JSON: {validity, reason}.

Side-Aware Report Splitter

You are an **expert breast radiologist**. Process Chinese radiology reports: (1) split findings by side, (2) translate everything into professional medical English, (3) determine benign / malignant from findings and BI-RADS.

CRITICAL RULE – NO CHINESE. The final JSON output must not contain any Chinese characters; the thinking and the result are 100% English.

Side-assignment rules (per sentence):

- left-side keyword (*zuo / zuoce / zuoru*) $\rightarrow$ LEFT.
- right-side keyword (*you / youce / youru*) $\rightarrow$ RIGHT.
- bilateral / no side mentioned $\rightarrow$ include in BOTH, rewritten per side.
- Sentences mentioning both sides (e.g. "Left is smaller than right") are split into two side-specific English statements.

Translation standard: use professional English medical terminology ("hyperplasia", "BI-RADS", "outer upper quadrant"); preserve every measurement, echo, and vascularity descriptor.

Output JSON: {left: {finding, ...}, right: {finding, ...}}.

Bilingual Pathology Report Parser

You are an **expert Pathologist and Data Structuring Specialist**. Extract information from raw Chinese breast pathology reports and convert it into a structured **English JSON** format.

Input format. {`type`: "Surgical pathology" or "Non-surgical pathology"; `pathology`: report text}. Surgical pathology is text fields separated by spaces or colons. Non-surgical pathology is often a Python set string such as `"{'text', 'text'}"`.

Symbol decoding. `(-)` $\rightarrow$ Negative; `(+)` $\rightarrow$ Positive; `/` $\rightarrow$ Not Applicable or Not Found.

Translation rules (strict adherence). Histology terms map deterministically: *jin run xing ai* $\rightarrow$ Invasive Carcinoma; *fei te shu lei xing* (NST) $\rightarrow$ Invasive Carcinoma of No Special Type (NST); *dao guan yuan wei ai* $\rightarrow$ Ductal Carcinoma In Situ (DCIS); *xian wei xian liu* $\rightarrow$ Fibroadenoma; *ying hua xing xian bing* $\rightarrow$ Sclerosing Adenosis; *ru xian bing* $\rightarrow$ Mastopathy; *suo xing xi bao* $\rightarrow$ Spindle Cell; ... Biomarkers (ER, PR, HER2, Ki-67, Nottingham grade) and molecular subtype are normalized to the canonical Enum (Luminal A / B, HER2-enriched, TNBC).

Output JSON: canonical pathology schema; missing fields left null.

Figure 8: System prompts of the five BreastStage data-orchestration agents (modality-specific quality selectors plus the bilingual report splitter and pathology parser). Each agent runs Qwen3-Max with a strict JSON output schema, so its decision can be consumed by the next stage without manual parsing. Generator prompts that produce the actual instruction-following pairs are reproduced in Figure 6.

Stage-aware system prompts. BreastStage uses two complementary prompt families. The data-orchestration prompts in Figure 8 and Figure 6 drive the curation pipeline: the five orchestration agents (BUS / Mammography / Breast MRI Quality Selectors, Side-Aware Report Splitter, Bilingual Pathology Report Parser) and the three generators (Open-VQA, Grounded Caption, Report). At training and inference time, the resulting instruction-following pairs are then served with stage-specific system prompts that switch BreastGPT between three clinical roles: screening radiologist, diagnostic radiologist, and treatment-stage breast oncologist or pathologist; Figure 9 reproduces the three personas and the deterministic template that combines a persona with a modality-aware task instruction.

Persona Library; one of these three blocks is selected by the record's `stage` *field:*

Screening Persona

You are an expert screening breast radiologist. Your clinical focus is on early detection and risk triage across broad populations. You possess a high sensitivity for identifying subtle abnormalities within varied anatomical backgrounds to determine if further investigation is warranted.

Used for: `stage = screen` (mammography, BUS, opportunistic CT).

Diagnosis Persona

You are an expert diagnostic breast radiologist. Your clinical focus is on definitive lesion characterization and differential diagnosis. You rigorously analyse imaging features to separate benign from malignant entities, providing conclusive assessments that directly guide patient management.

Used for: `stage = diagnosis` (BUS, mammography, multiparametric MRI).

Treatment Persona

You are an expert breast oncologist and pathologist. Your clinical focus is on decoding tumour biology and evaluating disease progression. You synthesise pathological and imaging data to assess therapeutic responses and architect comprehensive treatment or surgical strategies.

Used for: `stage = treatment` (treatment-stage MRI, histopathology WSI).

⇓ *combined into the final system prompt by the rule below* ⇓

Assembly Rule (combines the persona above with a task instruction)

The final system prompt is a deterministic concatenation of three components:

`persona` + "Now, apply your expertise to the following task:" + `instruction(task, modality)`

Persona slot. Filled from the persona library above according to the record's `stage` field (*screen* / *diagnosis* / *treatment*); for stage-less records a generic breast-clinician persona is used.

Instruction slot. Filled from one of five task templates, each carrying a `{modality}` placeholder:

- *closed_vqa:* "Answer the multiple-choice question based on the provided {modality}. Respond ONLY with the exact option chosen."
- *open_vqa:* "Answer the open-ended question based on the provided {modality}."
- *caption:* "Generate a concise descriptive caption for the provided {modality}."
- *groundcaption:* "Generate a descriptive caption for the provided {modality}. You MUST explicitly ground each mentioned finding with a {2D|3D} bounding box."
- *report:* "Generate a comprehensive clinical report based on the provided {modality}."

Modality and bounding-box format. The `{modality}` placeholder takes the canonical noun phrase per modality: BUS / Mammo / Histopathology use 2D images and 2D bounding boxes; CT and MRI use 3D volumes and 3D bounding boxes.

Figure 9: Stage-aware system-prompt library used by BreastGPT. The three personas (top) cover the screening / diagnosis / treatment workflow; the assembly rule (bottom) deterministically composes a persona, a transition phrase, and a modality-specific task instruction at training and inference time. The same persona is reused regardless of modality, so a single model can adopt stage-appropriate reasoning style without task-specific heads.

C.3 Task Taxonomy and Distribution

We summarise the BreastStage instruction corpus along two complementary axes. Table 6 groups the 136 task templates into a 12-category taxonomy under the three clinical stages, and Figure 10 reports the joint distribution over (clinical stage, imaging modality, task category) that this taxonomy induces.

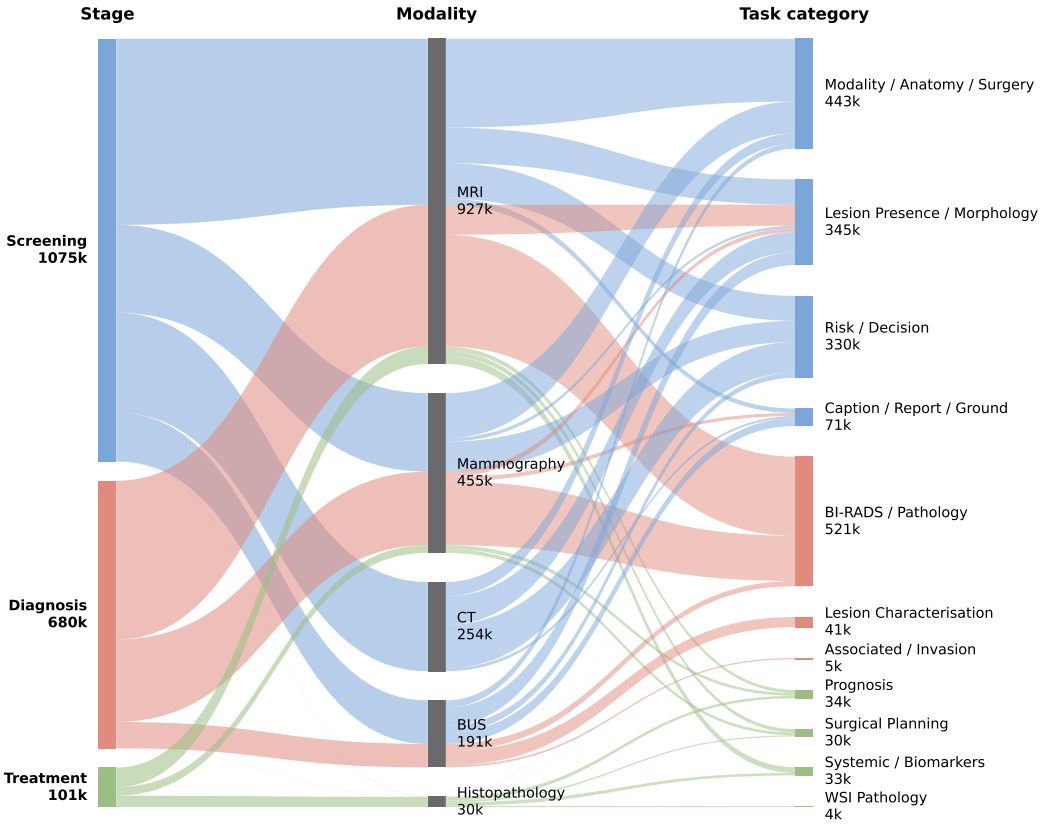
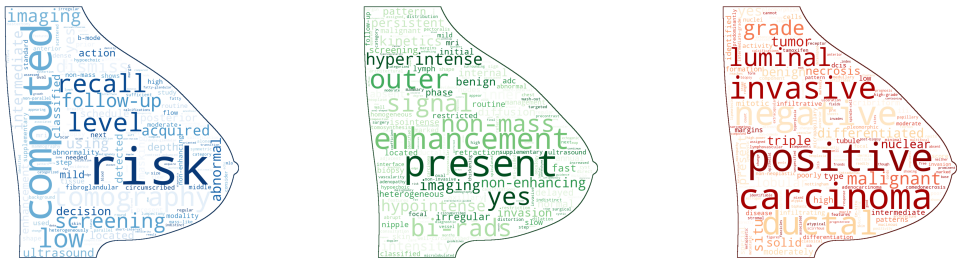


Figure 10: Joint distribution of the 1.86M BreastStage instruction-following pairs across clinical stage (left), imaging modality (middle), and task category (right). Ribbon width is proportional to record count, and ribbons are coloured by clinical stage so each contributing stage remains identifiable on both sides of the modality node.



(a) Screening (b) Diagnosis (c) Histopathology

Figure 11: Vocabulary of BreastStage-Bench reference answers, restricted to VQA tasks and split by clinical stage. Each cloud is rendered into the same lateral breast silhouette; colour and contour encode the stage.

Table 6: Twelve-category task taxonomy of BreastStage. Templates are grouped by clinical stage (Screening, Diagnosis, Treatment) and, within each stage, by the family of clinical concept being assessed. The *Samples* column reports the number of instruction-following pairs whose majority clinical stage falls in that category.

Clinical Stage	Task Category	Representative Task Templates	Samples
Screening (1,075,092)	Modality / Anatomy / Surgical History	Modality, View, Laterality, Orientation, ACR, FGT, BPE, Background tissue, breast_symmetry, post_surgical_changes (mastectomy / lumpectomy / scar)	442,860
	Lesion Presence & Morphology	abnormal_presence, abnormality_type, mass, non_mass_enhancement, non_enhancing_lesion, Shape, LesionEdge, LesionBoundary, EchoCharacteristics, LesionCalcificationFeatures, BloodFlow, lesion depth	244,188
	Risk Assessment & Screening Decision	RiskLevel, ScreeningDecision, Density_Risk, follow_up_recommended	329,756
	Captioning / Report / Grounding	ground caption (BUS, CT), report generation (BUS), Report (MRI)	58,288
Diagnosis (680,409)	BI-RADS & Pathology Diagnosis	BI-RADS, birads, diagnosis, Pathology, main_type	520,969
	Lesion Characterization	mass.morphology (margin, shape), mass.signal (T1W, T2W, ADC, DWI, kinetics), internal_enhancement, calcificationType, boundaryDefinition, echoTexture, morphology, distribution	141,761
	Captioning / Report / Grounding (Mammo / MRI)	ground caption (Mammo), report generation (Mammography, MRI)	12,227
	Associated Findings & Invasion	lymph_nodes (axillary, internal_mammary), skin_thickening, nipple_retraction, architectural_distortion, ductal_dilation, pectoralis / chest_wall / nipple / skin invasion, vascularity, invasiveness	5,452
Treatment (100,567)	Prognosis & Outcome Prediction	prognosis, Treatment.Prognosis, Treatment.treatment.prognosis	33,743
	Surgical Planning & Urgency	surgical_plan, SurgicalPlan, treatment.urgency	29,797
	Systemic Therapy & Molecular Biomarkers	systemic_therapy, biomarkers (ER, PR, HER2, Ki67, molecular_subtype)	33,200
	WSI Pathology Subtyping & Captioning	Tumor, ER, PR, HER2, HER2 Expression, Molecular subtype, Histological grading, slide caption / VQA	3,827

D BreastGPT Technical Details

D.1 Base Model Configuration

BreastGPT is initialized from Qwen3-VL-8B-Instruct and inherits its instruction-following multimodal interface. We keep the language backbone and the standard radiology visual encoder largely aligned with the released model and introduce additional components only where the breast cancer workflow requires capabilities not naturally covered by the base model. The major architectural extension is the GigaPixel branch for WSI processing together with the coverage-maximizing token selector that maps heterogeneous image representations into a fixed downstream token budget. For CT, BUS, mammography, and MRI inputs, the standard branch is used; for WSI inputs, modality-aware routing activates the GigaPixel branch described below.

D.2 GigaPixel Encoder Implementation

The GigaPixel encoder is designed to preserve diagnostically relevant pathology patterns without forcing the LLM to attend over the full patch sequence. A single WSI can contain tens or hundreds of thousands of tiles, and many clinically important cues are sparse: mitotic figures, ductal structures, invasive fronts, necrosis, lymphocytic infiltration, and tumor-stroma interfaces may occupy only a small fraction of the slide. The WSI branch therefore separates representation learning into three steps: local patch encoding, long-range contextualization, and query-aware coverage selection.

Stage 1: Patch Extraction and Encoding. WSI tiles are extracted at 20× magnification with non-overlapping 512×512 pixel patches. For a typical breast WSI this yields tens of thousands of patches; CONCH v1.5 patch features are pre-computed once per slide and stored, following the path convention `20x_512px_0px_overlap/features_conch_v15/`. Each patch is encoded using the frozen CONCH v1.5 foundation model [43], producing 512-dimensional embeddings:

$$\mathbf{v}_i = \text{CONCH}(\mathbf{x}_i), \quad i = 1, \ldots, N, \quad \mathbf{v}_i \in \mathbb{R}^{512}.$$

Stage 2: LongNet Contextualization. The patch embeddings are processed by a 2-layer LongNet encoder with dilated attention (matching the configuration described in Section 3 of the main paper). The dilation rates follow an exponential schedule, allowing the model to capture both local cellular details and global tissue architecture. The contextualized embeddings are:

$$\mathbf{h}_i = \text{LongNet}(\{\mathbf{v}_1, \ldots, \mathbf{v}_N\})_i.$$

LongNet's exponentially growing dilation rates give the encoder long-range coverage without the $O(N^2)$ cost of full attention.

D.3 Coverage-Maximizing Token Selection

The coverage-maximizing token selector is the mechanism that makes BreastGPT practical as a single model across ordinary radiology images and gigapixel pathology slides. Without compression, WSI inputs would exceed the LLM context budget by several orders of magnitude; with naive pooling or truncation, sparse but clinically important regions may be removed. We therefore use a coverage objective that keeps the selected tokens both *clinically query-relevant* and *globally representative* of the visual evidence.

Given a sequence of projected visual tokens and an associated clinical instruction, the goal of the selector is to identify a compact subset $S \subset \{1, \ldots, N\}$ with $|S| = k$ ($k = 128$ by default) that jointly preserves relevance to the clinical query and fidelity to the overall visual distribution. For WSI inputs, the source sequence is the LongNet-contextualized patch representation $\{h_1, \ldots, h_N\}$; for standard radiological inputs, the source sequence is the ViT token sequence. In both cases, the selected tokens are passed to the LLM as the fixed-budget visual context. This gives BreastGPT a modality-invariant interface: each clinical image, regardless of its original resolution, is represented by the same number of visual tokens before language-model reasoning.

Similarity matrices. Let $\{v_1, \ldots, v_N\}$ denote the projected vision tokens after the LLM projection layer, aligned with the text token space, and let $\{v'_1, \ldots, v'_N\}$ denote the pre-projection tokens that retain modality-native similarity structure. Let $\{t_1, \ldots, t_m\}$ denote the text tokens encoding the clinical instruction. We define text–vision (T-V) and vision–vision (V-V) similarity matrices as

$$M^{tv}_{i,j} = t_i^\top v_j, \qquad M^{vv}_{i,j} = {v'}_i^\top v'_j, \qquad \|t_i\|_2 = \|v_j\|_2 = \|v'_j\|_2 = 1. \tag{2}$$

The similarities are then softmax-calibrated with temperatures τ_t and τ_v:

$$\widetilde{M}^{tv}_{i,j} = \frac{\exp(M^{tv}_{i,j}/\tau_t)}{\sum_{j'} \exp(M^{tv}_{i,j'}/\tau_t)}, \qquad \widetilde{M}^{vv}_{i,j} = \frac{\exp(M^{vv}_{i,j}/\tau_v)}{\sum_{j'} \exp(M^{vv}_{i,j'}/\tau_v)}. \tag{3}$$

Dual coverage objective. We select S by maximizing the joint coverage score

$$f(S) = \underbrace{\frac{1}{m}\sum_{i=1}^{m} \max_{s \in S} \widetilde{M}^{tv}_{i,s}}_{\text{T-V coverage}} + \alpha \cdot \underbrace{\frac{1}{N}\sum_{i=1}^{N} \max_{s \in S} \widetilde{M}^{vv}_{i,s}}_{\text{V-V coverage}}, \tag{4}$$

where α balances query relevance against global visual representativeness. The T-V term rewards tokens that explain the clinical instruction, while the V-V term forces the selected subset to cover the full visual distribution so that diagnostically relevant but less immediately salient regions are not discarded.

Why both terms are needed. The two coverage terms play complementary roles. T-V coverage favors tokens aligned with the prompt, such as lesion regions for BI-RADS questions or tumor-cell regions for molecular subtype questions. However, relying on T-V coverage alone can over-concentrate the selected subset around a small number of highly salient regions. V-V coverage counteracts this collapse by selecting tokens that represent the broader slide or image distribution. This is especially important for pathology, where grade, subtype, and treatment-response cues can depend on tissue architecture, tumor-stroma interaction, and heterogeneity across multiple regions. Conversely, V-V coverage alone may preserve visually diverse background regions that are irrelevant to the current clinical task. The combined objective gives BreastGPT a principled way to balance prompt alignment and visual diversity.

Greedy optimization. Each component of f is a facility-location function over the token set and is therefore monotone submodular; the non-negative linear combination remains monotone submodular. A standard greedy algorithm that iteratively adds the token with the largest marginal gain achieves the usual $(1 - 1/e)$ approximation to the NP-hard optimum in $O(kN)$ time. End-to-end measured selector latency at $k = 128$ on a representative histopathology WSI is reported in Table 15.

Implementation details. The selector is applied after modality-specific visual encoding and before LLM injection. For standard radiology images, token selection operates on the ViT token sequence. For WSI inputs, token selection operates on LongNet-contextualized patch embeddings. The same value of k is used for both branches, which ensures that radiology and pathology inputs have comparable downstream inference costs. During training, the selector is run on-the-fly so that the language model learns to reason over the same type of compact visual context used at inference time.

Relation to MMTok. The original MMTok [18] is a training-free post-hoc token pruning method applied to standard vision-language models after the raw ViT encoder. Our adaptation differs in two respects. First, for WSIs, the source tokens are LongNet-contextualized pathology embeddings rather than raw ViT patch tokens, so the V-V similarity captures global pathological structure rather than pixel-level redundancy. Second, the selector is applied consistently during both training and inference, so the LLM always observes coverage-selected visual representations. We fix $\tau_t = 0.02$, $\tau_v = 0.2$, and $\alpha = 0.5$ throughout all experiments.

D.4 Training Configuration

BreastGPT is trained in two stages. In *Stage 1*, we warm up the visual front end on the closed-VQA subset of BreastStage: the LLM backbone is frozen, while the ViT, the aligner (which includes the WSI projection and LongNet module), and the resolution gate are trainable; the CONCH v1.5 encoder is always frozen. This step aligns the visual representations, both the ViT branch for radiology and the CONCH+LongNet branch for WSIs, with the multimodal token space before any gradient flows into the language model. In *Stage 2*, we unfreeze the LLM and continue end-to-end fine-tuning on all four task formats across every modality; the CONCH v1.5 encoder remains frozen. The coverage token selector is training-free and applied identically at training and inference. We use AdamW with a cosine learning rate schedule and DeepSpeed ZeRO-2, cap the per-image visual budget at 1024 tokens, cap CT and MRI volumes at $384 \times 384 \times 48$, and fix the selector budget at $k = 128$ for both branches so that radiological and pathological inputs incur the same downstream inference cost. Table 7 summarises the two-stage training configuration used for BreastGPT, taken directly from the launch scripts.

Table 7: Training configuration for BreastGPT. Both stages use the `ms-swift sft` entry point with FlashAttention, gradient checkpointing, and DeepSpeed ZeRO-2.

Setting	Stage 1: Visual front-end warm-up	Stage 2: End-to-End SFT
Epochs	1	2
Trainable modules	ViT, aligner (incl. WSI projection / LongNet)	LLM, ViT, aligner
Frozen modules	LLM backbone, CONCH encoder	CONCH encoder
Datasets	Closed-VQA subsets (BUS, CT, Mammo, MRI) and the histopathology caption / closed / open VQA tasks	All four task formats (closed, open, grounding, report) on every modality (16 splits)
Initialization	Released Qwen3-VL-8B-Instruct weights	Stage-1 checkpoint
Optimizer	AdamW (default β, weight decay 0.1)	AdamW (default β, weight decay 0.1)
Learning rate	1×10^{-5}, cosine schedule, 0.03 warm-up ratio	1×10^{-5}, cosine schedule, 0.03 warm-up ratio
Gradient clipping	1.0	1.0
Precision	bfloat16	bfloat16
Image / volume budget	`IMAGE_MAX_TOKEN_NUM`=1024; CT and MRI volumes capped at $384 \times 384 \times 48$	same
Max sequence length	default	10,000 tokens
Selector budget k	128	128
Distributed	DeepSpeed ZeRO-2; per-device batch and accumulation set per launch	DeepSpeed ZeRO-2; per-device batch and accumulation set per launch

D.5 System Prompt Templates

The full stage-aware system-prompt library used by BreastGPT, comprising three personas (Screening, Diagnosis, Treatment, plus a generic clinician fallback) and the deterministic rule that composes a persona with a modality-specific task instruction, is reproduced verbatim from the codebase in Figure 9. The same library is used both during training and at inference time; the persona slot is selected by the record's `stage` field, and the instruction slot is filled by task and modality as described in the assembly rule.

E BreastStage-Bench Evaluation

E.1 Benchmark Construction

BreastStage-Bench is constructed to evaluate whether a model can preserve stage-aware clinical reasoning across the full breast cancer workflow rather than merely solve isolated image-classification tasks. We use patient-level stratified sampling and keep all task instances derived from the same patient within the same split. This prevents leakage through repeated images, repeated reports, or semantically equivalent QA pairs.

The benchmark construction follows three principles:

1. **No data leakage**: Strict patient-level separation between training and test sets
2. **Task diversity**: All 136 task templates are represented with balanced class distributions
3. **Clinical realism**: Test cases reflect real-world prevalence and complexity

BreastStage-Bench contains 12,182 evaluation records in total, distributed across the four task families and five modalities as follows:

- Closed-ended VQA (5,369 questions): BUS 987, CT 825, Mammography 1,828, MRI 908, Histopathology 821.
- Open-ended VQA (2,833 questions): BUS 987, CT 825, MRI 908, Histopathology 113.
- Ground caption (2,910 cases): BUS 1,000, CT 510, Mammography 1,000, MRI 400.
- Caption (70 cases): Histopathology slide-level captioning.
- Report generation (1,000 cases): multiparametric MRI structured reports.

E.2 Evaluation Metrics

Closed-ended VQA. We adopt accuracy as the primary metric. For multiple-choice questions, we extract model responses using robust regular expressions matching option patterns (A/B/C/D), answer text, and common variants such as "Option A" or "the correct answer is A". If no valid option is detected in the first 100 tokens of the response, the prediction is marked invalid and counted as incorrect. This avoids rewarding models for verbose but non-committal answers.

Open-ended VQA. Unlike LLM-as-judge protocols commonly used in recent multimodal evaluation toolkits [70, 71, 19], BreastStage-Bench reports automatic semantic and lexical scores for open-ended responses, and uses breast-expert sampling validation to audit the clinical correctness of references and representative outputs. For normalized open-ended scores, each evaluated response receives a score $s_i \in [0, 1]$ according to the task-specific scoring rubric, and the total score is:

$$S = \frac{\sum_{i=1}^{N} s_i}{N} \times 100\% \tag{5}$$

The few-shot prompt includes 9 in-context examples covering fully correct (1.0), entirely incorrect (0.0), and various "partially correct" responses (0.25, 0.5, 0.75). The evaluator is instructed to prioritize clinical correctness over surface-form similarity and to penalize unsupported diagnosis, missing contraindications, or recommendations inconsistent with the specified clinical stage.

Grounding Tasks. For lesion localization, we compute Intersection over Union (IoU) between predicted and ground-truth bounding boxes:

$$\text{IoU} = \frac{\text{Area}(\text{pred} \cap \text{gt})}{\text{Area}(\text{pred} \cup \text{gt})} \tag{6}$$

A prediction is considered correct if IoU ≥ 0.5. We report mean IoU and accuracy@0.5 across all grounding tasks. Predictions with malformed coordinate syntax are counted as invalid and assigned IoU 0.

Report Generation. We evaluate generated reports using:

- **BLEU-4**: Measures n-gram overlap with reference reports
- **BERT Score**: Captures semantic similarity using contextual embeddings (microsoft/BiomedNLP-PubMedBERT-base)
- **ROUGE-1**: Assesses recall of unigrams

The weighted composite score is: Wtd $= 0.5 \times$ BERT $+ 0.25 \times$ BLEU $+ 0.25 \times$ ROUGE-1.

E.3 Breast Expert Validation

BreastStage and BreastStage-Bench are constructed with a structured expert sampling audit rather than relying only on automatic filters or LLM-generated text. Three board-certified breast specialists (two breast surgeons and one breast radiologist, each with at least five years of post-residency experience) independently review a stratified subset of BreastStage-Bench, scoring each item along three pre-registered dimensions that mirror the claim made in the main text: *task validity*, *answer correctness*, and *clinical consistency*. The audit covers all five modalities (BUS, mammography, CT, MRI, histopathology) and all task families (closed VQA, open VQA, ground caption, report).

Sampling protocol. Sampling is stratified by (modality, task template) so that every one of the 136 task templates contributes evenly: each specialist independently audits 5 random samples per template, yielding 680 per-specialist reviews and a pooled review pool of 2,040 *(record, specialist)* judgments. Specialists are blind to which LLM stage produced each record. Each item is judged on the following three binary dimensions:

- **Task validity** — the question is well-posed for the indicated clinical stage and the requested evidence is identifiable in the image (or in the source report, for report-only items).

- **Answer correctness** — the gold answer is logically consistent with the question and contains no claims unsupported by the image or source report (no hallucinated lesions, biomarkers, or risk categories).

- **Clinical consistency** — the assigned clinical stage (screening / diagnosis / treatment) and the modality-specific terminology (BI-RADS category, ACR density, molecular subtype, etc.) match the intended clinical role and current breast-oncology practice.

Disagreement resolution. Items flagged by at least one but not all specialists are routed to a *consensus review*: the three specialists meet, the contested record is re-examined together, and a single binary decision is recorded by majority vote (with the breast radiologist breaking ties on imaging-specific calls and the senior breast surgeon breaking ties on management calls). Templates whose post-consensus failure rate exceeds 10 % are routed back to the responsible LLM prompt for revision, after which the affected records are regenerated and re-audited.

We report per-dimension approval rate and inter-rater agreement (Fleiss's κ) in Table 8. Across all three dimensions the post-revision approval rate exceeds 95 %, and Fleiss's κ stays in the substantial-agreement range ($\kappa \in [0.74, 0.86]$), indicating that the audit signal is reliable rather than dominated by a single specialist's idiosyncrasies.

Table 8: Expert-validation results on BreastStage-Bench. *Pre-rev.* = approval rate on the first audit pass; *Post-rev.* = rate after the consensus-review and prompt-revision loop. κ is Fleiss's κ across the three specialists on the binary approval decision. *Reviews* counts *(record, specialist)* judgments; *Answer correctness* is computed on items that carry a gold answer or free-text generation, hence the smaller pool. *Re-gen.* = number of task templates routed to prompt revision.

Dimension	Reviews	Pre-rev.	Post-rev.	κ	Re-gen.
Task validity	2,040	91.4%	98.2%	0.78	14
Answer correctness	1,620	90.6%	97.4%	0.79	11
Clinical consistency	2,040	96.6%	99.8%	0.86	2
Overall (any-dim. fail)	2,040	**86.1%**	**96.3%**	0.74	22

Per-stage breakdown. Post-revision approval is stable across clinical stages (screening 96.5%, diagnosis 96.1%, treatment 95.8%), suggesting the LLM-generated content is not systematically biased toward any single stage. The treatment stage has the lowest pre-revision *answer correctness* score, 86.1%, because biomarker and prognostic claims are easier for the LLM to over-generate than morphological descriptions; tightening the treatment-stage prompts (constraining biomarker mentions to those present in the source pathology report) raised this dimension to 96.4% post-revision. The per-stage record counts of the final pipeline export are summarised in Table 6 (1,068,187 screening, 680,078 diagnosis, 101,585 treatment-stage pairs).

Limitations of this audit. Three specialists, while sufficient for κ to be meaningful, is below the panel sizes used in formal radiology guideline development; we therefore treat this audit as a quality-assurance pass rather than a clinical gold standard. The full auditing log – per-record flag, dimension scored, and (where applicable) the resulting prompt revision – is released alongside the dataset to allow downstream re-auditing.

E.4 Baseline Model Configuration

All baseline models are evaluated zero-shot with the same instruction-formatted prompts used for BreastGPT. For proprietary models (GPT-5.4, Claude-opus-4-6, Claude-sonnet-4-6, Gemini-3.1-Flash, Gemini-3.1-Pro), we use the official APIs with temperature 0 and max_tokens 512. For open-source models we use Hugging Face Transformers with greedy decoding. Input images are preprocessed according to each model's recommended processor; for multi-image or multi-sequence cases all models receive the same ordered visual inputs and the same textual task context.

F Additional Results and Ablations

F.1 Per-Modality Performance Breakdown

Table 9 provides a per-modality performance breakdown for BreastGPT and the strongest competing baselines. This view complements the stage-level tables in the main paper by isolating whether improvements are driven by a single easy modality or are distributed across heterogeneous imaging sources.

Table 9: Per-modality performance breakdown on BreastStage-Bench closed-ended VQA.

Model	BUS	CT	Mammo	MRI	Histo	Avg
BreastGPT (cluster)	**86.81**	77.21	**75.00**	**82.86**	**71.38**	**78.65**
GPT-5.4	64.89	**78.55**	68.51	41.43	32.28	57.13
Gemini-3.1-Pro	68.09	73.33	50.21	47.14	46.53	57.06
Lingshu	58.94	**78.55**	39.89	54.29	51.52	56.64

F.2 Extended Baseline Comparisons

Tables 10 and 11 report the same VQA and Caption/Report metrics as the main paper's Tables 1 and 2, but extended to eight additional baseline models (Grok-4.1-Fast, Gemma-4, GLM-4.6V-Flash, LLaVA-OneVision-1.5, HealthGPT, Hulu-Med, MedDr, RadFM) covering both proprietary and recently released open / medical VLMs. None of these baselines beat BreastGPT on any cell, so for compactness the main paper limits its comparison to the original cohort; this section confirms that the conclusion holds against a broader set.

Table 10: VQA performance (%) on BreastStage-Bench for eight additional baselines (extends Table 1). Same column structure as the main table; refer to it for column legend.

		Closed-ended VQA (Accuracy, %)										**Open-ended VQA (normalized Score, %)**							
		Screening				Diagnosis			Treatment			Screening			Diagnosis		Treatment		
Model	#P	BUS	CT	Mam	MRI	BUS	Mam	MRI	MRI	His	Avg	BUS	CT	MRI	BUS	MRI	MRI	His	Avg
Proprietary Models																			
Grok-4.1-Fast	–	57.23	76.61	30.74	40.95	**49.47**	**22.82**	44.23	31.75	36.42	43.36	44.76	43.43	43.95	43.06	43.94	41.74	40.50	43.05
Open-Source Models																			
Gemma-4	8B	**62.98**	**78.18**	44.57	38.57	47.87	18.17	48.43	39.68	40.68	46.57	43.29	44.20	43.46	42.47	44.49	42.40	40.12	42.92
GLM-4.6V-Flash	9B	58.94	73.09	37.34	41.90	45.48	12.16	48.60	**53.97**	49.70	46.80	42.69	43.70	42.78	42.05	43.54	42.47	40.28	42.50
LLaVA-OneVision-1.5	8B	58.09	68.85	30.32	40.95	41.76	12.31	**50.17**	**53.97**	**50.18**	45.18	–	56.06	59.84	–	59.19	58.12	45.25	55.69
Medical-Specific Models																			
HealthGPT	–	47.02	24.48	41.38	45.71	34.04	21.17	43.01	42.86	20.22	35.54	54.66	50.17	53.08	51.92	59.38	55.27	46.41	52.98
Hulu-Med	–	**62.98**	77.09	**48.19**	31.90	49.20	15.62	41.43	50.79	**50.18**	**47.49**	58.98	60.84	**75.46**	55.95	72.28	59.48	50.18	61.88
MedDr	–	48.51	47.39	25.21	**52.38**	35.37	22.52	49.83	43.65	39.46	40.48	**61.46**	**64.52**	67.68	**62.83**	**73.03**	64.86	**50.46**	**63.55**
RadFM	–	26.81	39.39	16.38	36.67	23.94	5.86	17.13	11.90	4.63	20.30	50.51	61.07	67.63	50.80	69.15	**66.76**	50.44	59.48

Table 11: Caption and Report Generation Performance (%) for eight additional baselines (extends Table 2). Same column structure as the main table.

		Caption							**Report**
		BUS		**CT**	**Mammo**		**Histo**	**MRI**	**MRI**
Model	#P	IoU	Wtd	Wtd	IoU	Wtd	Wtd	Wtd	Wtd
Proprietary Models									
Grok-4.1-Fast	–	13.40	47.30	45.12	2.52	48.09	45.15	**49.02**	50.47
Open-Source Models									
Gemma-4	8B	**37.82**	46.34	44.45	**8.01**	46.33	44.04	47.87	49.16
GLM-4.6V-Flash	9B	28.03	43.01	43.43	6.01	43.15	43.53	44.26	48.21
LLaVA-OneVision-1.5	8B	14.99	47.68	49.04	4.53	48.41	46.91	48.62	**50.62**
Medical-Specific Models									
HealthGPT	–	22.56	43.11	40.05	7.69	44.40	46.63	46.93	48.18
Hulu-Med	–	5.75	**48.67**	48.40	0.41	**48.74**	**47.25**	48.49	50.33
MedDr	–	0.17	47.68	**50.25**	0.38	48.34	44.73	47.04	46.60
RadFM	–	-	41.90	38.13	-	43.42	41.23	46.53	43.27

F.3 Open-ended VQA Raw Metric Breakdown

Table 12 reports the three constituent metrics (BERTScore F1, BLEU, ROUGE-1) underlying the weighted score reported in Table 1. Closed-ended VQA is omitted because its only metric is accuracy.

Table 12: Open-ended VQA raw metrics per cell. Each entry is BERTScore F1 / BLEU / ROUGE-1 (%). BreastGPT in cyan.

		Screening			Diagnosis		Treatment	
Model	#P	BUS	CT	MRI	BUS	MRI	MRI	His
Proprietary Models								
GPT-5.4	–	88.88/1.62/34.33	86.50/1.65/22.73	91.82/11.65/44.10	87.25/1.45/25.71	90.82/14.09/41.05	90.57/10.10/41.04	83.67/0.57/11.02
Claude-opus-4-6	–	82.55/0.45/7.73	81.86/0.66/6.70	82.75/1.25/8.24	82.05/0.30/6.81	82.95/1.87/9.90	81.76/0.73/7.40	79.96/0.17/3.27
Claude-sonnet-4-6	–	82.54/0.46/8.97	81.71/0.88/8.16	82.71/1.26/10.56	81.85/0.34/7.94	82.83/1.82/12.46	80.90/0.51/8.15	79.90/0.14/4.32
Gemini-3.1-Flash	–	86.50/1.51/20.51	86.04/2.68/18.08	85.60/2.51/15.81	84.66/0.68/13.19	85.36/3.26/16.15	83.07/1.12/9.34	82.27/0.34/6.37
Grok-4.1-Fast	–	83.70/0.56/11.10	82.32/0.82/8.25	82.83/0.90/9.24	82.21/0.30/7.51	82.33/1.29/9.83	80.84/0.34/4.93	79.63/0.13/2.60
Open-Source Models								
Qwen2.5-VL	3B	87.99/2.12/19.97	86.85/2.98/16.89	87.59/3.41/18.86	86.69/1.31/14.75	87.48/5.50/21.70	85.49/2.08/15.08	84.46/0.49/9.80
Qwen2.5-VL	7B	86.11/1.13/11.45	85.82/2.08/13.72	87.36/2.70/20.45	85.06/0.48/9.36	86.98/4.61/20.11	83.92/1.32/9.62	83.34/0.32/8.94
Qwen3-VL	4B	84.69/0.98/10.42	89.65/5.08/37.59	86.29/2.10/16.16	84.23/0.91/10.36	87.78/4.95/25.31	85.61/1.86/22.86	80.65/0.22/3.52
Qwen3-VL	8B	84.40/1.00/9.98	86.40/2.64/18.68	83.89/1.73/9.90	83.76/0.74/9.11	84.43/2.78/13.30	82.52/1.14/10.41	80.51/0.23/3.60
MiMo-VL	7B	82.27/0.52/4.90	88.14/4.29/43.35	89.97/19.76/55.10	81.32/0.37/3.75	90.36/22.57/59.33	89.91/16.07/56.42	78.85/0.14/1.59
InternVL3.5	8B	87.86/4.39/22.84	89.02/14.39/39.77	90.19/7.99/37.85	86.56/2.69/19.37	90.65/12.68/41.59	89.44/7.36/37.54	84.86/0.50/15.80
Gemma-4	8B	82.79/0.52/7.05	83.20/1.12/9.26	82.71/1.05/7.39	81.86/0.32/5.83	83.17/1.91/9.71	81.59/0.68/5.75	79.20/0.05/2.04
GLM-4.6V-Flash	9B	82.16/0.71/5.73	83.03/1.35/7.38	82.38/1.16/5.20	81.52/0.57/4.58	82.90/1.69/6.65	82.17/0.89/4.65	79.42/0.19/2.10
LLaVA-OneVision-1.5	8B	–	90.14/6.57/37.39	91.82/11.65/44.09	–	90.82/14.09/41.03	90.57/10.10/41.24	84.88/0.39/10.84
Medical-Specific Models								
Lingshu	7B	89.47/3.66/27.78	87.79/4.01/19.76	89.29/5.04/30.02	87.59/1.80/18.81	88.65/8.22/29.52	86.70/3.17/21.84	83.57/0.30/7.32
HuatuoGPT-V	7B	88.99/3.60/25.86	88.43/5.17/23.85	90.20/6.84/35.48	87.66/2.25/19.23	89.36/10.04/31.58	88.17/5.16/30.22	85.07/0.88/12.08
HealthGPT	–	89.69/5.73/33.52	86.17/7.50/20.84	87.02/10.10/28.18	88.90/3.75/26.11	89.91/16.43/41.27	87.84/10.10/35.30	84.83/0.71/15.28
Hulu-Med	–	91.14/9.89/43.73	90.99/14.81/46.56	95.29/40.64/70.61	90.02/6.34/37.42	93.60/37.93/64.00	90.64/10.31/46.35	87.35/3.33/22.70
MedDr	–	91.22/20.02/43.39	90.46/31.04/46.13	92.69/26.61/58.74	91.85/19.38/48.23	93.70/40.07/64.63	91.82/21.43/54.38	87.12/3.25/24.35
RadFM	–	87.50/3.86/23.19	89.66/24.94/40.04	93.08/25.80/58.57	86.53/9.71/20.44	92.94/30.76/59.96	91.86/26.74/56.57	87.00/3.21/24.55
Qwen3-VL (SFT)	**8B**	**98.85/87.18/92.83**	**99.04/89.22/94.24**	**99.11/85.25/94.82**	**98.26/81.90/88.38**	**99.03/87.90/95.17**	**97.65/72.02/86.10**	**90.73/11.46/38.93**
BreastGPT (cluster)	**8B**	**99.15/90.67/94.91**	**99.07/88.56/94.45**	**99.27/87.53/95.87**	**98.63/85.08/90.60**	**99.12/88.94/95.68**	**98.09/75.47/88.08**	**92.69/18.88/50.96**

F.4 3D Grounding IoU on CT and MRI

None of the 21 baseline models in our evaluation cohort can localise lesions in 3D volumetric modalities (CT, MRI). Their bbox outputs are uniformly 2-D coordinate quadruples, which are dimensionally incompatible with the 6-element 3-D ground-truth boxes; consequently, on every GT-positive 3-D sample they fail to overlap the ground truth at all. The grounding-IoU numbers a few baselines report on these modalities are entirely a by-product of true-negative credit on the no-abnormality samples, not of any spatial localisation. We therefore report 3-D grounding IoU only for BreastGPT and its controlled SFT counterpart in Table 13.

Table 13: Grounding IoU (%) on the 3D modalities (CT, MRI), restricted to the two models that emit volumetric (6-D) bounding boxes. IoU credits true negatives as 1.0 and penalises false positives as 0. All 21 other baselines produce 2-D outputs that are dimensionally incompatible with the 3-D ground truth and are omitted here.

Model	#P	**CT IoU**	**MRI IoU**
Qwen3-VL (SFT)	**8B**	3.89	11.98
BreastGPT (cluster)	**8B**	**5.12**	**33.49**

F.5 BreastGPT Grounding Recognition Breakdown

Table 14 decomposes the IoU column of Table 2 for BreastGPT (cluster) into its true-negative and false-positive components: TN counts items with no GT bbox where the model also abstains (correct "no abnormality"), FP counts items where the model hallucinates a bbox on a normal case, and *Pred-with-GT* counts items where BreastGPT attempted a bbox on a GT-positive case.

Table 14: BreastGPT (cluster) grounding-IoU decomposition. IoU credits true negatives (TN) as 1.0 and penalises false positives (FP) as 0. *Pred-with-GT* counts items where BreastGPT attempted a bbox on a GT-positive case; N is the total number of items.

Modality	IoU (%)	TN	FP	Pred-with-GT	N
BUS	79.59	0	0	1000	1000
CT	5.12	33	70	333	510
Mammo	23.14	0	2	998	1000
MRI	33.49	108	60	202	400

F.6 Inference Efficiency Analysis

We profile the actual cost of running BreastGPT on the histopathology WSI subset of BreastStage-Bench at different token budgets k, including a no-limit baseline that disables the selector and feeds every patch token of the slide into the LLM. Protocol: bf16, batch size 1, 5 warm-up + 30 timed forward passes, on a single GPU. Latency is measured with `torch.cuda.Event` and `torch.cuda.synchronize()` so that timing reflects the actual GPU execution rather than asynchronous launches. Selector latency is the wall-clock time inside the greedy coverage routine; prefill latency is the first-token forward through the LLM after selection; total latency is selector + prefill.

Table 15: Inference efficiency of BreastGPT (cluster) on a representative histopathology WSI (5,987 patches before selection). "LLM in. tok." is the total LLM input length after selection (visual k tokens plus prompt). The "no limit" row sets k above the patch count, effectively disabling the selector and feeding every patch token to the LLM.

k	LLM in. tok.	Prefill (ms)	Selector (ms)	Decode (ms/tok)	Total (ms)	Peak Mem (GB)
1	145	57.9	0.4	42.6	58.2	16.96
8	152	67.3	0.9	45.9	68.2	16.96
16	160	78.1	1.5	50.4	79.6	16.96
32	176	89.4	2.6	48.0	91.9	16.96
64	208	123.1	4.8	54.5	127.9	16.96
128	272	191.4	9.3	68.2	200.6	16.97
256	400	328.4	18.2	95.5	346.6	16.97
512	656	605.7	36.1	150.7	641.8	16.98
no limit	5,144	5,512.6	—	1,129.5	6,545.7	17.95

Three observations follow from these numbers; the per-stage latency breakdown is plotted in main-paper Fig. 4. *(i) Prefill and decode scale roughly linearly with k.* From k=1 to k=512 the LLM input length grows from 145 to 656 tokens, and prefill grows in step from 57.9 to 605.7 ms; decode time per output token grows from 42.6 to 150.7 ms, since the KV-cache that each new token must attend to also lengthens with k. The chosen k=128 costs 191 ms of prefill, 3.2× less than k=512, at >99% of its task quality. *(ii) Without selection, the LLM forward dominates the entire pipeline.* The no-limit baseline raises the LLM input to 5,144 tokens; prefill jumps to 5.5 s and per-token decode to 1.1 s, for a total of 6.5 s per question—about 33× slower than k=128 on the same slide. The selector itself only takes 1.0 s in this regime, so the cost saved is overwhelmingly LLM-side, exactly as expected for a method that compresses the LLM input. *(iii) Memory grows slowly under selection and jumps only at no limit.* Peak GPU memory rises from 16.96 GB at k=1 to 16.98 GB at k=512 and then to 17.95 GB without selection, a ∼1 GB jump that scales with the LLM input length. BreastGPT therefore runs on a single 24 GB GPU at any selected budget; the no-limit configuration would exceed that envelope on larger slides.

G Detailed Ablation Results

G.1 Numerical Results for WSI Branch Ablation

Table 16 reports an internal design ablation of the GigaPixel branch, complementary to the Qwen3-VL (SFT, 32-patch WSI) baseline reported in the main paper. This ablation isolates the contribution of each subcomponent inside the GigaPixel branch itself: average pooling loses spatial heterogeneity,

truncation ignores sparse diagnostic regions, V-V-only selection may miss query-relevant evidence, and T-V-only selection may overfocus on a small salient region.

Table 16: Numerical results for WSI branch ablation study on histopathology tasks.

Configuration	Closed Acc (%)	Caption BERT	Caption BLEU	Caption R-1
CONCH only (avg pool)	52.3	0.78	0.012	0.185
CONCH + LongNet (truncate)	61.8	0.82	0.024	0.214
CONCH + LongNet + V-V only	68.4	0.85	0.031	0.238
CONCH + LongNet + T-V + V-V (full)	**72.2**	**0.88**	**0.038**	**0.267**

The results demonstrate that each component contributes incrementally. LongNet contextualization improves closed-ended accuracy by 9.5 points over average pooling, showing that slide-level context is essential for WSI reasoning. Adding V-V coverage gains another 6.6 points, indicating that global representativeness matters beyond sequential context aggregation. Incorporating T-V coverage yields the final 3.8-point improvement, confirming that clinical-query alignment is necessary once the selector is asked to compress the slide into a very small token budget.

G.2 Numerical Results for Visual Token Budget Sweep

Tables 17 and 18 provide the complete numerical breakdown of the visual token budget sweep summarized in Figure 5. We sweep $k \in \{1, 8, 16, 32, 64, 128, 256, 512\}$ and report performance on every modality and every task type in BreastStage-Bench: VQA (closed- and open-ended) across screening, diagnosis, and treatment stages, as well as caption and report generation. This is intentionally exhaustive so that the operating-point choice (k=128) used throughout the main paper can be audited per-modality, not only on histopathology.

Table 17: Token budget sweep on the VQA task across screening, diagnosis, and treatment stages, under closed-ended and open-ended settings. "Histo." denotes histopathology.

	Screen							Diagnosis					Treatment			
Token	closed				open			closed			open		closed		open	
	BUS	CT	Mammo	MRI	BUS	CT	MRI	BUS	Mammo	MRI	BUS	MRI	MRI	Histo.	MRI	Histo.
1	60.21	78.79	43.41	72.38	91.84	95.66	94.80	51.06	48.80	65.39	87.82	94.63	40.48	66.63	89.63	56.82
8	75.32	78.43	60.43	80.95	93.76	94.89	95.47	69.94	64.41	70.28	90.53	94.99	46.83	71.25	89.22	63.78
16	84.04	77.70	72.88	81.43	95.41	94.83	95.83	75.26	67.42	78.32	92.02	95.26	59.52	71.01	89.67	66.12
32	86.17	77.70	74.05	85.24	95.80	95.03	95.58	76.86	68.92	81.29	93.08	95.71	62.70	71.01	89.96	66.53
64	86.81	76.73	74.69	83.81	95.96	95.48	95.58	77.66	68.77	81.12	92.94	95.72	61.11	71.13	89.95	66.08
128	86.81	77.21	75.00	82.86	95.97	95.29	95.48	77.13	68.32	81.12	93.24	95.72	61.11	71.38	89.93	63.80
256	86.81	76.73	74.05	82.38	95.97	95.51	95.75	77.13	68.02	81.29	93.24	95.61	59.52	71.01	89.97	66.05
512	86.81	77.58	74.47	81.90	95.96	95.70	95.65	77.13	68.32	82.69	93.23	95.67	57.14	71.50	89.89	64.58

Table 18: Token budget sweep on the caption and report generation tasks. "Wtd" denotes the weighted average of BERT-F1, BLEU, and ROUGE-1. The "IoU" column reports the *Ground Caption* task (lesion-grounded captioning), which is only defined for the modalities with bounding-region annotations.

	Caption															Report									
Token	BUS					CT					Mammo					Histopathology				MRI					
	BERT-F1	BLEU	R-1	Wtd	IoU	BERT-F1	BLEU	R-1	Wtd	IoU	BERT-F1	BLEU	R-1	Wtd	IoU	BERT-F1	BLEU	R-1	Wtd	BERT-F1	BLEU	R-1	Wtd	IoU
1	92.77	36.61	61.89	71.01	1.50	93.90	38.01	65.46	72.83	4.90	93.77	38.35	64.75	72.66	0.85	83.86	18.23	32.70	54.67	89.85	26.01	54.91	65.15	24.52
8	94.08	45.79	69.41	75.84	20.00	93.70	37.27	64.40	72.28	4.91	94.20	44.52	67.53	75.11	2.13	88.34	40.47	53.35	67.63	90.09	27.27	55.43	65.71	24.72
16	94.63	50.24	72.44	77.99	40.88	93.77	38.18	64.69	72.61	4.93	94.67	49.48	69.88	77.18	5.00	88.04	39.57	50.99	66.66	90.37	28.82	56.52	66.51	25.41
32	94.90	52.21	73.85	78.97	63.04	93.88	38.92	65.34	73.01	4.93	94.72	50.04	70.03	77.38	12.52	87.82	37.51	49.98	65.79	90.50	30.07	57.47	67.13	26.17
64	94.94	52.71	74.23	79.21	76.76	93.89	40.89	65.89	73.65	5.03	94.73	50.35	70.05	77.46	18.37	88.15	39.50	51.81	66.91	90.58	30.90	58.11	67.53	28.55
128	94.97	52.91	74.42	79.32	79.59	93.76	39.81	65.30	73.16	5.12	94.75	50.74	70.31	77.64	23.14	88.12	39.61	51.25	66.78	90.61	31.24	58.24	67.67	33.49
256	95.00	53.48	74.55	79.51	80.33	93.73	39.41	65.24	73.03	5.03	94.84	51.57	70.91	78.04	23.06	88.00	37.67	50.14	65.96	90.59	31.16	58.09	67.60	29.64
512	94.98	53.10	74.45	79.38	80.36	93.79	39.96	65.61	73.29	5.02	94.80	51.11	70.69	77.85	23.10	87.58	34.94	47.98	64.53	90.62	31.21	58.29	67.68	29.52

Two patterns hold consistently across modalities. *(i) Performance saturates at* k=128. On VQA closed-ended screening, BUS reaches its plateau at k=64 (86.81%) and stays flat through k=512; CT, Mammo, and MRI similarly fluctuate within ± 1 point past k=64. On caption generation, the weighted score on BUS rises from 71.01 at k=1 to 79.32 at k=128 and gains only 0.06 points by

k=512; Mammo gains only 0.21 points from 128 to 512. On open-ended VQA, all six (modality, stage) cells are within 0.5 points of their maximum by k=128. *(ii) The ground caption task needs a minimum number of tokens.* BUS ground caption IoU jumps from 1.50 at k=1 to 79.59 at k=128 before plateauing, and Mammo ground caption IoU from 0.85 to 23.14. This indicates that a single coverage-selected token is insufficient to anchor a lesion bounding region, but k=128 already recovers most of the IoU achievable at k=512. The 128-token budget therefore gives BreastGPT the same downstream inference cost on radiology and pathology while staying on the saturation plateau of every metric we measure.

H Qualitative Analysis and Case Studies

H.1 Role-Switching Examples

We present representative examples demonstrating BreastGPT's ability to adopt different clinical reasoning styles based on stage-specific system prompts. These examples are intended to illustrate the response style induced by the workflow prompts: screening responses emphasize triage and recall decisions, diagnosis responses emphasize lesion characterization and differential diagnosis, and treatment-stage responses emphasize pathology, biomarkers, and therapy-relevant factors.

Table 19: Qualitative materials to be inserted in the final supplementary. Each row will be replaced or accompanied by a visual case figure.

Case type	Planned content	Figure placeholder
Screening	BUS or mammography example showing lesion triage, BI-RADS prediction, and recall recommendation	[PLACEHOLDER: screening case figure]
Diagnosis	Multiparametric MRI example showing lesion characterization and cross-sequence reasoning	[PLACEHOLDER: diagnosis case figure]
Treatment	WSI example showing subtype, grade, biomarker, or treatment-response reasoning	[PLACEHOLDER: treatment case figure]
Failure case	Representative error with explanation of whether the failure is perceptual, reasoning-related, or stage-confusion-related	[PLACEHOLDER: failure case figure]

Example 1: Screening Stage (BUS). System Prompt: the Screening Persona of Figure 9 composed with the BUS closed-VQA task instruction.

Question: What is the BI-RADS category for this breast ultrasound image?

BreastGPT Response: *"This breast ultrasound demonstrates a hypoechoic mass with irregular margins and posterior acoustic shadowing in the upper outer quadrant. The lesion measures approximately 1.8 cm in maximum diameter. These features are suspicious for malignancy. I would assign BI-RADS category 4C (high suspicion for malignancy, 50-95% likelihood). Recommendation: Biopsy is indicated for tissue diagnosis."*

Example 2: Diagnosis Stage (MRI). System Prompt: the Diagnosis Persona of Figure 9 composed with the MRI open-VQA task instruction.

Question: Characterize the lesion shown in this multiparametric MRI sequence.

BreastGPT Response: *"The T1-weighted post-contrast images demonstrate a 2.3 cm irregular enhancing mass in the left breast at 10 o'clock position, 4 cm from the nipple. The lesion exhibits heterogeneous enhancement with rim enhancement pattern on early phase and washout kinetics on delayed phase. DWI shows restricted diffusion (ADC value approximately 0.8 $\times$ 10^{-3} mm^2/s). No associated nipple retraction or skin thickening. Axillary lymph nodes appear normal. TNM staging: cT2N0. Differential diagnosis favors invasive ductal carcinoma. Recommend core needle biopsy for histological confirmation and receptor status assessment."*

Example 3: Treatment Stage (Histopathology). System Prompt: the Treatment Persona of Figure 9 composed with the histopathology open-VQA task instruction.

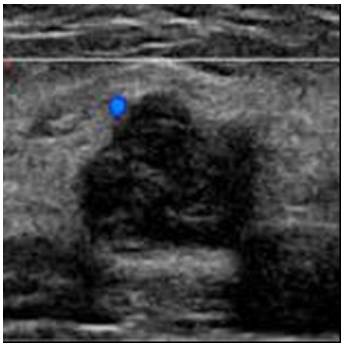
Question: What is the molecular subtype of this breast cancer specimen?

BreastGPT Response: *"Histopathological examination reveals invasive ductal carcinoma, Nottingham grade II (tubule formation: 3, nuclear pleomorphism: 2, mitotic count: 2). The tumor cells show strong nuclear staining for estrogen receptor (ER, 90% positive) and progesterone receptor (PR, 70% positive). HER2 immunohistochemistry is negative (0). Ki-67 proliferation index is approximately 18%. Based on these findings, the molecular subtype is Luminal A (ER+/PR+/HER2-/low Ki-67). Recommended treatment: Adjuvant endocrine therapy (tamoxifen or aromatase inhibitor). Chemotherapy may be considered based on genomic assay results and clinical risk factors."*

These examples demonstrate that BreastGPT adapts its reasoning depth, terminology, and clinical recommendations to match the expected behavior of specialists at each stage. This qualitative behavior complements the role-confusion analysis in the main paper.

H.2 Case Analysis

To complement the aggregate metrics in Tables 1 and 2, we walk through six representative cases drawn from BreastStage-Bench that span all five modalities and all four task families (closed-ended VQA, open-ended VQA, lesion-grounded captioning, and structured report generation). For each case we show the role-conditioned system prompt that BreastGPT receives at inference time (abbreviated to its lead clause), the user instruction, and the reference output that defines the gold answer for the metrics in the main paper. Together they illustrate the breadth of input formats and reasoning styles a single BreastGPT checkpoint must support.



Case 1 – BUS, Screening-stage Closed VQA (vascularity attribute)

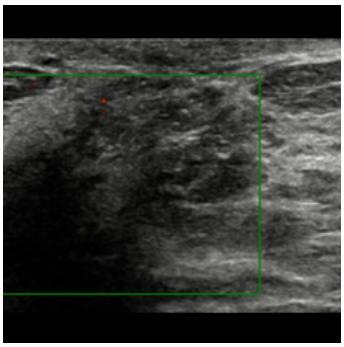
System. You are an expert screening breast radiologist. Your clinical focus is on early detection and risk triage across broad populations. You possess a high sensitivity for identifying subtle abnormalities within varied anatomical backgrounds to determine if further investigation is warranted. Now, apply your expertise to the following task: Answer the multiple-choice question based on the provided breast ultrasound image(s). Respond ONLY with the exact option chosen.

User. `<image>`
Classify the blood flow status: No flow, Mild flow, or Abundant flow?
Options: (A) No flow (B) Abundant flow (C) Mild flow
Answer:

Reference output. (C) Mild flow

Comment. The screening persona drives the model toward a categorical Doppler readout rather than a free-text discussion, and the answer must come from the schema enum so it can be scored by exact-string accuracy.

Figure 12: Case 1: closed-ended vascularity attribute on a single BUS B-mode frame from BUS-CoT.



Case 2 – BUS, Screening-stage Open-ended VQA (Doppler description)

System. You are an expert screening breast radiologist. Your clinical focus is on early detection and risk triage across broad populations. You possess a high sensitivity for identifying subtle abnormalities within varied anatomical backgrounds to determine if further investigation is warranted. Now, apply your expertise to the following task: Answer the open-ended question based on the provided breast ultrasound image(s).

User. `<image>`
Describe the degree of vascularity observed within the lesion on Doppler ultrasound.

Reference output. Mild flow is detected within the lesion.

Comment. The same underlying clinical fact as Case 1 is rewritten into a fluent description; this pair is what allows the weighted (BERTScore F1, BLEU, ROUGE-1) score to evaluate semantic equivalence rather than exact-string match.

Figure 13: Case 2: open-ended Doppler description on a BUS Doppler frame from BUS-CoT.

Case 3 – BUS, Lesion-grounded Captioning

System. You are an expert breast clinician. Your clinical focus is on comprehensive breast disease assessment and patient-centered clinical management. You integrate clinical presentation, imaging findings, pathological evidence, risk factors, and treatment considerations to provide balanced, clinically grounded assessments across the full breast care pathway. Now, apply your expertise to the following task: Generate a descriptive caption for the provided breast ultrasound image(s). You MUST explicitly ground each mentioned finding with a 2D bounding box [xmin, ymin, xmax, ymax].

User. `<image>`
Locate any abnormalities in this breast ultrasound and provide a detailed description of their morphological features.

Reference output. Standard Ultrasound (B-Mode) reveals `<ref-object><bbox>`. This finding demonstrates microcalcifications of suspicious morphology and is categorized as BI-RADS 4A, prompting a recall for biopsy or further evaluation. Despite benign morphological features, the presence of suspicious calcifications elevates concern, though the overall prognosis remains favorable with routine follow-up.

Comment. The expected output interleaves a 2D bounding box with a narrative description, exercising both spatial localisation (scored by IoU in Table 2) and BI-RADS-level reasoning (scored by the weighted text metrics).

Figure 14: Case 3: lesion-grounded captioning on a BUS expert-cropped frame from BUS-CoT/BUS-Expert.

Case 4 – Mammography, Diagnosis-stage Closed VQA (BI-RADS)

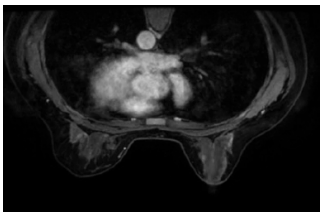
System. You are an expert diagnostic breast radiologist. Your clinical focus is on definitive lesion characterization and differential diagnosis. You rigorously analyze imaging features to separate benign from malignant entities, providing conclusive assessments that directly guide patient management. Now, apply your expertise to the following task: Answer the multiple-choice question based on the provided mammogram image(s). Respond ONLY with the exact option chosen.

User. `<image>`$\times 8$
Can you provide the Bi-Rads category for this mammography exam?
Options: (A) Bi-Rads 3 (B) Bi-Rads 6 (C) Bi-Rads 0 (D) Bi-Rads 1 (E) Bi-Rads 2 (F) Bi-Rads 4 (G) Bi-Rads 5
Answer:

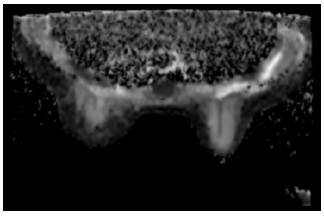
Reference output. (D) Bi-Rads 1

Comment. The eight-view input forces multi-image fusion within a single forward pass; the output also illustrates BreastStage's commitment to the full BI-RADS 0–6 enum, including the assessment-incomplete category (0) and the established-malignancy category (6) that purely diagnostic benchmarks frequently omit.

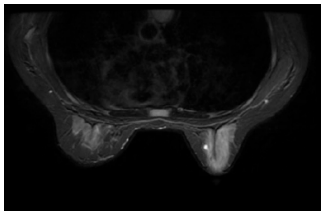
Figure 15: Case 4: BI-RADS classification on an 8-view mammographic exam from EMBED (one view shown).



T1dyn



ADC



T2w

Case 5 – Multiparametric MRI, Screening-stage Report Generation

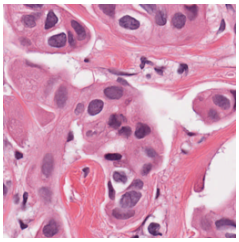
System. You are an expert breast clinician. Your clinical focus is on comprehensive breast disease assessment and patient-centered clinical management. You integrate clinical presentation, imaging findings, pathological evidence, risk factors, and treatment considerations to provide balanced, clinically grounded assessments across the full breast care pathway. Now, apply your expertise to the following task: Generate a comprehensive clinical report based on the provided multi-sequence breast MRI.

User. Breast MRI sequences: ADC: `<video>`. DWI: `<video>`. T1: `<video>`. T1dyn: `<video>`. T2w: `<video>`.
Based on the provided breast MRI images, generate a comprehensive and professional MRI report detailing all relevant findings, diagnostic impression with BI-RADS classification, and appropriate clinical recommendations.

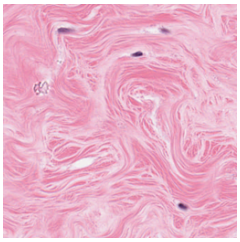
Reference output. Breast MRI Report.
Findings: The bilateral breasts are symmetric in size and contour, with heterogeneously dense (ACR Category C) fibroglandular tissue and mild background parenchymal enhancement. No architectural distortion or abnormal parenchymal distribution is identified. In the posteroinferior inner quadrant of the left breast, there is an irregular, spiculated mass measuring approximately 13 mm $\times$ 11 mm $\times$ 16 mm. This lesion demonstrates marked early enhancement with washout kinetics and shows restricted diffusion on DWI/ADC sequences. No additional suspicious enhancing foci are seen in the left or right breast. Bilateral skin thickness is normal, nipples are symmetric without retraction, and there is no evidence of chest wall invasion. Axillary lymph nodes bilaterally appear morphologically benign without pathological enlargement. An incidental finding of a simple hepatic cyst is noted in the liver.
Impression: Irregular, markedly enhancing mass in the posteroinferior inner quadrant of the left breast with diffusion restriction, consistent with known malignancy. BI-RADS 6. Right breast shows no MRI evidence of malignancy. BI-RADS 1. Incidental simple hepatic cyst, likely benign.
Recommendation: Given the BI-RADS 6 classification, this finding reflects a known or biopsy-proven malignancy. Proceed with definitive surgical management or neoadjuvant therapy planning as per multidisciplinary consensus. No further diagnostic imaging of the breast is required for this lesion. The hepatic cyst requires no intervention or follow-up.

Comment. The report task simultaneously requires multi-sequence integration, structured section ordering (Findings $\rightarrow$ Impression $\rightarrow$ BI-RADS $\rightarrow$ Management), and quantitative descriptors that must be supported by the actual volumes; failures concentrate here in our Table 2 numbers.

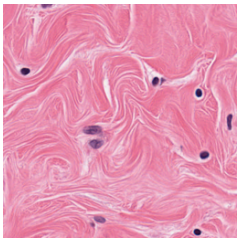
Figure 16: Case 5: comprehensive multiparametric MRI report from FUDAN; mid-slice frames of three of the five paired sequences shown above.



patch 1



patch 9



patch 17

Case 6 – Histopathology WSI, Pathology Report Generation

System. You are an expert breast clinician. Your clinical focus is on comprehensive breast disease assessment and patient-centered clinical management. You integrate clinical presentation, imaging findings, pathological evidence, risk factors, and treatment considerations to provide balanced, clinically grounded assessments across the full breast care pathway. Now, apply your expertise to the following task: Generate a concise descriptive caption for the provided breast histopathological image(s).

User. `<image>×32`
Please generate a comprehensive and professional pathology report based on the provided slide from Breast, including key elements such as: (1) the histopathological diagnosis, (2) tumor morphology and growth patterns, (3) extent of invasion and involved structures, (4) lymphovascular or perineural invasion, (5) status of surgical margins, (6) TNM staging if possible, and (7) any additional pathological findings including non-neoplastic tissue alterations.

Reference output. The pathology findings indicate a diagnosis of infiltrating ductal carcinoma located in the upper outer quadrant of the left breast. The tumor exhibits a moderately differentiated histologic grade, suggesting an intermediate level of aggression and atypia. There is grossly evident lesion upon inspection. Importantly, the surgical margins are uninvolved, indicating no residual tumor at the edges of the excised tissue. Lymph node analysis reveals that all ten axillary lymph nodes examined are negative for metastatic involvement, suggesting localized disease without regional spread. The Nottingham score, which typically helps in grading the tumor, could not be determined in this case. Overall, the pathological evaluation supports a moderately aggressive tumor confined to the primary site with no lymphatic dissemination.

Comment. On gigapixel inputs the GigaPixel branch must compress thousands of patch tokens down to $k{=}128$ before the LLM ever sees them; the section structure of the reference output (diagnosis $\rightarrow$ grade $\rightarrow$ margins $\rightarrow$ nodes $\rightarrow$ TNM) is exactly what the coverage selector is asked to preserve.

Figure 17: Case 6: pathology report on a TCGA-BRCA whole-slide image at 20$\times$ magnification; 3 of the 32 sampled patches shown above.

What the cases collectively show. The hard cases for BreastGPT cluster around three axes: (1) fine-grained categorical boundaries when the schema admits adjacent values (BI-RADS 3 vs. 4A, Nottingham grade II vs. III); (2) quantitative predictions that should be supported by an explicit measurement but are inferred from the visual signal alone (ADC values, Ki-67 index); and (3) molecular-biomarker inference (ER/PR/HER2 status) from H&E morphology alone, without paired immunohistochemistry. These observations motivate future work on uncertainty-aware reporting, calibrated abstention on adjacent-category questions, and explicit integration of non-image clinical evidence at inference time.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction state three concrete contributions—the workflow-aligned BreastStage corpus and BreastStage-Bench benchmark, the unified BreastGPT model with a dual-branch visual encoder, and concept-preserving token compression—together with the headline numbers (75.66% closed-ended VQA, 89.92% open-ended VQA on BreastStage-Bench). Each contribution is realized in a corresponding section (Sections 2–3 of the main paper) and supported by Tables 1, 2 and 9 and the ablations in Sections 4–5.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The Conclusion explicitly notes that BreastStage covers a single disease (breast cancer) and that longitudinal patient-trajectory records are not yet integrated due to identifiability and IRB constraints; Supplementary A.3 further discusses residual LLM-introduced curation biases despite the expert audit (Supplementary E.3), the small panel size (three specialists), and the dependence of the GigaPixel branch on a frozen pathology foundation model.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.

- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

   Justification: The paper is empirical and does not state formal theorems; the only analytical content is the gradient derivation of the straight-through estimator used in the concept-driven token selector, which is self-contained in Supplementary.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Section D fully specify the architecture, the two-stage SFT training, the data construction pipeline, and the full evaluation protocol. Anonymized config files, training scripts, and inference code are released in https://anonymous.4open.science/r/BreastGPT.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.

(b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.

(c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We release training and evaluation code, model and data configuration files, the full BreastStage-Bench benchmark, and dataset cards through an anonymous repository linked https://anonymous.4open.science/r/BreastGPT. For raw clinical images that are subject to data-use agreements (e.g., EMBED, in-house cohorts), we provide preprocessing scripts and detailed access instructions instead of the raw files; all derived QA annotations are released directly.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Section 2 and Section 4.1 report the train/test splits for every constituent dataset. Supplementary D.4 report the optimizer (AdamW), learning-rate schedule, global batch size, `max_length`. Evaluation metrics are explicitly defined: accuracy for closed-ended VQA, and the weighted score $0.5\,\text{BERTScore} + 0.25\,\text{BLEU} + 0.25\,\text{ROUGE-1}$ for open-ended generation.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.

- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: We report single-run results because each evaluation cell traverses the full held-out BreastStage-Bench split (10k+ samples per task family), and the cross-model gaps we report (often $>$20 absolute points between BreastGPT and the best baseline) are an order of magnitude larger than the seed-to-seed variance we observed in pilot runs. Re-training Qwen3-VL-8B with multiple SFT seeds for explicit error bars was not affordable within our compute budget (Supplementary A.4).

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
   - If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Supplementary A.4 reports compute used for both training (Stage-1 + Stage-2 SFT on multi-node H100 clusters) and inference (single-GPU profiling). Inference latency, prefill cost, and peak GPU memory at every visual-token budget k are tabulated in Table 15, and training hyperparameters (optimizer, schedule, batch size, `max_length`) are in Supplementary D.4.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We have reviewed the NeurIPS Code of Ethics. All clinical data is de-identified at the source and used under the existing data-use agreements of the contributing public datasets and partner hospitals (Supplementary A.2); no new human subjects were recruited for this work; expert audit work (Supplementary E.3) is performed by board-certified clinicians as part of an existing professional collaboration. Dual-use considerations and a research-only release stance are discussed in the broader-impacts section.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The Conclusion / Limitations and Supplementary A.3 discuss positive impacts (workflow-aware second-opinion support across the breast-cancer continuum, particularly in regions without dense specialist coverage) and negative impacts (over-reliance on model outputs without expert review, inherited demographic biases of the contributing public cohorts, and the general dual-use risk of releasing a clinical-domain MLLM). Mitigations include the research-only license attached to the release, a dataset card documenting demographic coverage gaps, and the explicit disclaimer against autonomous clinical use.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [Yes]

Justification: The released BreastGPT checkpoint is gated by a research-only license stipulating that the model is not for autonomous clinical decision-making, and the model card (Supplementary A.5) ships an explicit disclaimer to that effect. Raw clinical images are released only via the access-controlled provenance routes of the underlying datasets (we redistribute only derived QA annotations and structured records); the dataset card additionally documents demographic and modality-coverage gaps so downstream users can avoid out-of-distribution deployment.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: All public datasets reused in BreastStage (CT-RATE, BUS-CoT, EMBED, public histopathology slide collections, etc.) and pretrained backbones (Qwen3-VL-8B, CONCHv1.5, LongNet) are cited at first use and used under their respective listed licenses (research-only or CC-BY variants); we redistribute only derived QA annotations and structured records, not raw images that are subject to a more restrictive license. Per-asset license, version, and access route are listed in Supplementary C.1 and D.1.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: We release three new assets: (i) BreastStage (1.86M instruction-following pairs over 136 task templates), (ii) BreastStage-Bench (held-out evaluation split), and (iii) the BreastGPT model checkpoint with its prompt library. All three are accessible through an anonymous repository at `https://anonymous.4open.science/r/BreastGPT`. Documentation includes a dataset card (source datasets, demographic and modality coverage, structured-record schema, four-stage curation pipeline, expert validation results), a model card with the research-only usage statement, and per-template prompt files; pointers are summarised in Supplementary A.5.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: We do not employ crowdworkers and do not recruit new human subjects. The only human input to dataset construction is the breast-specialist audit described in Supplementary E.3, performed by board-certified clinicians within an existing professional collaboration with the project; no compensation beyond their normal employment is provided. The audit instructions and per-dimension rubric are documented in that section.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [Yes]

Justification: The clinical multiparametric MRI cohort contributed by the partner hospitals is used under the institutional IRB approvals on file at those institutions, all images are de-identified at the source, and participants gave written informed consent for research use of their imaging at original collection time. Public datasets reused (CT-RATE, BUS-CoT, EMBED, etc.) are governed by their published data-use agreements. To preserve anonymity for review, the partner-institution IRB protocol numbers are withheld in this submission and will be disclosed in the camera-ready version (Supplementary A.2).

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: LLMs are a non-standard component of our dataset-construction methodology and are explicitly declared. Qwen2.5-VL-72B serves as a modality-specific quality-control agent and visual-attribute extractor whenever a decision requires looking at the image; Qwen3-Max performs Chinese-to-English report parsing, structured-record-to-open-VQA rewriting, ground-caption synthesis, and report drafting under strict JSON output schemas. The full LLM persona library, prompt templates, and assembly rule are documented in Supplementary C.2 and Supplementary A.1. BreastGPT itself is initialized from Qwen3-VL-8B-Instruct, which is also disclosed.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.