Supplementary Information

Contents

1 Supplementary Methods

1.1 Dataset construction pipelines and prompts

1.1.1 Construction of SonoCorpus

To provide scalable supervision for ultrasound reasoning initialization, we constructed SonoCorpus using an automated multi-stage pipeline. The pipeline uses Qwen3-VL-32B to generate hierarchical chain-of-thought (CoT) rationales and applies ground-truth (GT)-based reverse reasoning, self-reflection, and consensus voting to reduce visually unsupported or inconsistent rationales.

Initial CoT generation and GT-based reverse reasoning We used a multi-temperature decoding strategy, with τ ranging from 0.1 to 1.0 at a step size of 0.1, to generate 10 candidate reasoning paths per image. To ground the early reasoning stages, we adopted a **GT-based reverse reasoning** prompt. The model was provided with verified clinical metadata, including protocol, system, and organ labels, and was instructed to justify these labels using image-based evidence before producing a diagnosis-related assessment (normal, abnormal, or inconclusive). Here, the diagnosis-related assessment is defined as a coarse image-level judgment rather than a fine-grained disease label: `normal` indicates no obvious abnormal sonographic finding, `abnormal` indicates visible abnormal findings, and `inconclusive` indicates insufficient visual evidence or degraded image quality. This design prevents the model from freely guessing basic anatomical context and instead focuses generation on visually grounded explanation and abnormality assessment. Accordingly, the generated diagnosis-related rationales were used for reasoning initialization rather than as expert-confirmed fine-grained disease annotations.

The structured prompt template used for this stage is as follows:

Prompt template for initial CoT generation

You are a medical AI system with expertise in interpreting diagnostic ultrasound images. Given the ground-truth labels, including the scanning protocol, anatomical system, and target organ, generate a concise visual rationale that explains why the image supports these labels. Then provide a coarse image-level diagnosis-related assessment: normal, abnormal, or inconclusive.

Only describe the key visual features supporting each decision.
Ground-truth (GT) labels: `[Protocol GT]`, `[System GT]`, `[Organ GT]`.

Format your response exactly as follows:
`<think>` [Concise visual reasoning validating the protocol, system, organ, and coarse diagnosis-related assessment] `</think>`
`<protocol>` [Protocol GT] `</protocol>`
`<system>` [System GT] `</system>`
`<organ>` [Organ GT] `</organ>`
`<answer>` [normal | abnormal | inconclusive] `</answer>`

Self-reflection and hallucination filtering Because generated rationales may contain visually unsupported sonographic findings, each generated response was passed back to the model together with the original image for a self-reflection step. The model was asked to verify whether the rationale was visually grounded, logically consistent, and compatible with the provided ground-truth labels.

Responses marked as `Valid` were retained as generated, responses marked as `Needs Correction` were retained after replacing the original rationale with the refined rationale, and responses marked as `Hallucinated` or unparsable were discarded.

Prompt template for self-reflection

You are a medical AI system reviewing a generated ultrasound reasoning response. Given the original ultrasound image, the generated rationale, and the predicted assessment, check whether the reasoning is visually grounded and logically consistent.

Previous response:
`<think>` [Generated rationale] `</think>`
`<answer>` [Generated assessment] `</answer>`
Review the response according to the following criteria:

- Are the key sonographic features mentioned in the rationale visible in the image?
- Is the rationale consistent with the verified protocol, system, and organ labels?
- Does the response contain unsupported findings, hallucinated structures, or contradictory reasoning?

Format your refined response exactly as follows:
`<refined_think>` [Refined or confirmed concise reasoning] `</refined_think>`
`<verification_status>` [Valid | Needs Correction | Hallucinated] `</verification_status>`

Consensus voting To reduce stochastic variability from multi-temperature sampling, we applied a **majority consensus voting** mechanism to the coarse diagnosis-related assessments. For a given image i, let V_i denote the set of responses retained after self-reflection, including `Valid` responses and corrected `Needs Correction` responses. We counted the frequency of each assessment label within V_i and identified the mode of the distribution. CoT rationales whose final coarse assessment matched this majority consensus were retained, while minority responses were discarded. This filtering step yielded approximately 7.0 million retained reasoning trajectories with fewer stochastic inconsistencies and visually unsupported rationales.

1.1.2 Construction and quality control of SonoVQA

SonoVQA was constructed using a clinician-in-the-loop pipeline that combines metadata-driven structured question generation with expert-reviewed sonographic feature questions. The benchmark is designed to evaluate hierarchical ultrasound reasoning across protocol, system, organ, sonographic feature, lesion category, and disease diagnosis levels. Structured multiple-choice questions (MCQs) provide objective evaluation for taxonomy-grounded reasoning levels, while yes/no and open-ended questions target visually grounded sonographic feature interpretation.

Metadata-driven MCQ construction at the protocol level The protocol level evaluates whether the model can recognize basic acquisition context before anatomical reasoning. In SonoVQA, this level includes two sub-tasks: ultrasound imaging mode recognition and scanning protocol recognition.

1. Ultrasound imaging mode recognition. We classify images into three imaging modes: B-mode, color Doppler flow imaging (CDFI), and power Doppler imaging (PDI). For each image,

a 3-choice MCQ is generated, with the correct mode as the answer and the other two modes as distractors. The standardized question format is: *"Which ultrasound imaging mode does this image represent? Please choose the correct option."*

2. Scanning protocol recognition. The model is then asked to identify the clinical scanning protocol, such as abdominal, OB-GYN, superficial, musculoskeletal, or peripheral vascular protocol. Ground-truth protocols are derived from the case metadata and anatomical taxonomy. To reduce ambiguity in distractor construction, we used an **anti-ambiguity distractor sampling strategy**. Protocols with highly overlapping acquisition appearances were not used as distractors for each other. For example, abdominal and OB-GYN examinations may share similar convex-probe appearances and acoustic windows; therefore, when the ground-truth protocol is abdominal, OB-GYN is excluded from the distractor pool, and vice versa. Similar exclusions were applied between superficial and musculoskeletal protocols. The remaining valid candidates are randomly sampled to form a 4-choice MCQ.

An example of the standardized format is shown below:

Example format for protocol MCQs

Question: Which scanning protocol is most appropriate for this ultrasound image? Please choose the correct option.
Options:
A) Peripheral vascular protocol
B) Abdominal protocol
C) Superficial protocol
D) Musculoskeletal protocol
Answer: B

Metadata-driven MCQ construction at system, organ, lesion, and diagnosis levels Building on the protocol level, we generated MCQs for progressively higher reasoning levels, including anatomical system, target organ, lesion category, and disease diagnosis. For each level, distractors were sampled from the corresponding taxonomy to create clinically plausible but clearly incorrect alternatives, thereby reducing shortcuts based on obvious anatomical or semantic mismatch.

1. Anatomical system and organ identification. The system-level task asks the model to identify the anatomical system according to the SonoVQA taxonomy, while the organ-level task asks for the target organ depicted in the image. Both levels use 4-choice MCQs. For organ-level questions, distractors were sampled with anatomical constraints to avoid overly ambiguous alternatives from structures that are frequently co-imaged or spatially adjacent. For example, when the ground-truth organ is the liver, adjacent structures such as the gallbladder may be excluded from the distractor pool to avoid penalizing cases where multiple structures are visible in the same field of view. This design encourages the question to focus on the primary target organ rather than incidental co-visible anatomy.

2. Lesion categorization. To evaluate intermediate diagnostic reasoning, we generated lesion-category questions over broad pathological groups, including tumor, obstructive disease, inflammatory disease, trauma, and diffuse abnormality. In addition, binary tumor versus non-tumor questions were included when appropriate. Distractors were sampled within the lesion-category taxonomy while avoiding combinations that could be clinically ambiguous for the selected case. This strategy helps maintain a single best answer while preserving clinically meaningful alternatives.

3. Disease diagnosis. For disease diagnosis questions, we used a cascading hard-negative sampling strategy to construct 4-choice MCQs. Distractors were preferentially sampled from diagnoses within the same disease sub-category. If insufficient candidates were available, the sampling pool was progressively expanded to diagnoses within the same organ, then the same anatomical system, and finally the global disease taxonomy. This strategy increases the clinical plausibility of distractors while maintaining a well-defined ground-truth answer.

Sonographic-feature question-answer (QA) generation via GPT-4o To evaluate fine-grained sonographic feature interpretation, we used GPT-4o to draft yes/no and open-ended QA pairs from expert-written clinical case descriptions and the corresponding ultrasound images. These questions focused on visually observable ultrasound descriptors, such as echogenicity, margins, internal echoes, vascularity, and posterior acoustic behavior. To reduce diagnostic leakage and keep the questions focused on visual evidence, we applied the following constrained prompt.

Prompt template for sonographic feature QA generation

You are a medical ultrasound imaging expert. Given a short ultrasound case description and the corresponding image, generate concise and factual QA pairs about image-observable sonographic features.

Follow these rules:
1. Each question must correspond to a visual feature directly inferable from the image and supported by the case description, such as echogenicity, margins, internal echoes, vascularity, or posterior acoustic behavior. This includes both yes/no and open-ended questions.
2. Do not include or mention anatomical system names, organ names, specific diseases, diagnoses, patient information, or imaging modality terms.
3. Avoid speculative or interpretive questions. Generate only questions with clear, objective answers that can be visually confirmed.
4. Do not refer to the case description itself in the question or answer.
5. Return only a JSON list of QA pairs.

Clinician-in-the-loop quality control Although vision-language models (VLMs) can efficiently draft sonographic feature QAs, the generated content may contain visually unsupported findings, inaccurate terminology, or unintended diagnostic hints. Therefore, all drafted yes/no and open-ended QA pairs were manually reviewed by experienced sonographers against the source ultrasound image. The review process focused on the following criteria:

1. **Visual verification:** Confirm that the queried feature, such as posterior acoustic shadowing or vascularity, is visible in the provided keyframe.

2. **Terminology standardization:** Refine generated descriptions to use standard sonographic terminology, such as echogenicity, margin, shape, vascularity, calcification, and posterior acoustic features, while avoiding disease labels or risk-stratification categories.

3. **Leakage control:** Remove residual anatomical, diagnostic, or patient-related hints that could allow shortcut reasoning.

4. **Ambiguity removal:** Discard questions when the visual evidence is borderline, poorly visible, or likely to cause substantial inter-observer disagreement.

This review step reduced visually unsupported statements and improved the clinical consistency of the final SonoVQA annotations.

1.1.3 SonoVQA question distribution and representative examples

Table S1 summarizes the composition of SonoVQA across dataset scale, question format, and reasoning taxonomy. Metadata-driven MCQs were generated for protocol, system, organ, lesion-category, and disease-diagnosis levels. The protocol level includes imaging-mode and scanning-protocol recognition, whereas the lesion-category level includes both tumor/non-tumor and fine-grained lesion-category questions. Sonographic-feature questions were constructed separately as yes/no and open-ended QA pairs to evaluate image-observable ultrasound descriptors.

Table S1: Dataset overview and question taxonomy of the SonoVQA benchmark.

Dataset overview			
Clinically curated cases			1,503
Question-answer pairs			14,905
Anatomical systems / target organs			7 / 18
Lesion categories / disease diagnoses			32 / 64
Question format			
Multiple-choice questions			10,521
Yes/no questions			2,768
Open-ended questions			1,616
Question taxonomy			
Reasoning level	**Subtask / target**	**Question format**	**VQA pairs**
Protocol	Imaging mode recognition	MCQ	1,503
Protocol	Scanning protocol recognition	MCQ	1,503
System	Anatomical system identification	MCQ	1,503
Organ	Target organ localization	MCQ	1,503
Sonographic feature	Image-observable feature interpretation	Yes/no + open-ended	4,384
Lesion category	Tumor / non-tumor classification	MCQ	1,503
Lesion category	Fine-grained lesion-category classification	MCQ	1,503
Disease diagnosis	Fine-grained disease diagnosis	MCQ	1,503
Total			**14,905**

We present a representative SonoVQA case to illustrate how questions are organized across the six reasoning levels. Figure S1 shows the source ultrasound image, and Table S2 lists the corresponding QA examples.

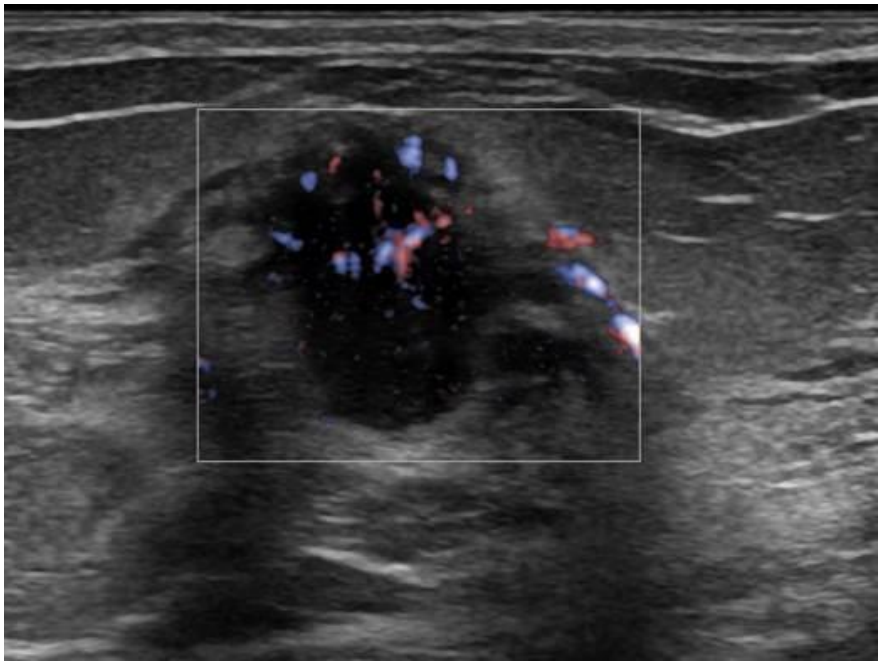


Figure S1: Representative breast ultrasound case from SonoVQA, showing a hypoechoic mass with internal vascularity under power Doppler imaging (PDI).

Table S2: Hierarchical VQA examples demonstrating the progression from low-level perception to high-level clinical reasoning.

Reasoning Level	**Visual Question and Ground Truth Answer**
Protocol	**Question:** Which ultrasound imaging mode does this image represent? **Options:** A: Color Doppler (CDFI) | B: B-mode | C: Power Doppler Imaging (PDI) **Answer:** C: Power Doppler Imaging (PDI)
	Question: Which scanning protocol is most appropriate for this image? **Options:** A: Peripheral vascular protocol | B: Superficial protocol | C: OB-GYN | D: Abdominal **Answer:** B: Superficial protocol
System	**Question:** Which anatomical system does this ultrasound image primarily belong to? **Options:** A: Peripheral vascular | B: Superficial tissues system | C: Musculoskeletal | D: Abdominopelvic **Answer:** B: Superficial tissues system
Organ	**Question:** Which anatomical organ is primarily depicted in this ultrasound image? **Options:** A: Thyroid | B: Uterus | C: Breast | D: Hand **Answer:** C: Breast
Sonographic Feature	**Question:** Is vascularity visible within the lesion? **Answer:** Yes, vascularity is visible within the lesion.
	Question: What is the echogenicity of the lesion compared to surrounding tissue? **Answer:** The lesion appears hypoechoic compared to surrounding tissue.

Continued on next page

Table S2 – Continued from previous page

Reasoning Level	Visual Question and Ground Truth Answer
	Question: Are the lesion margins well-defined? **Answer:** No, the lesion margins appear irregular and poorly defined.
Lesion Category	**Question:** Is the primary lesion a tumor or a non-tumor condition? **Options:** A: Tumor | B: Non-tumor **Answer:** A: Tumor
	Question: Which disease category is primarily represented in this image? **Options:** A: Obstructive disease | B: Tumor | C: Trauma | D: Inflammatory disease **Answer:** B: Tumor
Disease Diagnosis	**Question:** Which specific diagnosis is primarily represented in this ultrasound image? **Options:** A: Fibroadenoma | B: Breast carcinoma | C: Breast cyst | D: Mastitis **Answer:** B: Breast carcinoma

1.2 Training configurations and hyperparameters

1.2.1 Stage 1: supervised fine-tuning on SonoCorpus

The first stage performs supervised fine-tuning (SFT) on SonoCorpus, which contains approximately 7.0 million retained reasoning trajectories. This stage initializes the model with the expected ultrasound reasoning format before reinforcement alignment. We use the open-source LLaMA-Factory framework [1] for implementation. During SFT, the vision tower and multimodal projector are frozen, and full-parameter fine-tuning is applied to the language model backbone. The detailed hyperparameters for the SFT stage are summarized in Table S3.

Table S3: Hyperparameters for stage 1: supervised fine-tuning on SonoCorpus.

Hyperparameter	Value
Base Model	Qwen3-VL-Instruct (8B/32B)
Framework	LLaMA-Factory [1]
Fine-tuning Type	Full fine-tuning of language model
Frozen Components	Vision Tower, Multimodal Projector
Optimization Strategy	DeepSpeed ZeRO-3
Precision	BF16
Learning Rate	1.0×10^{-5}
Learning Rate Scheduler	Cosine
Warmup Ratio	0.1
Training Epochs	1
Per-Device Train Batch Size	12
Gradient Accumulation Steps	8
Evaluation Interval	Every 1,000 steps
Cutoff Length (Tokens)	4,096
Image Max Pixels	262,144

1.2.2 Stage 2: GRPO alignment on the multi-task mixture

Starting from SonoReasoner-Init, we apply Group Relative Policy Optimization (GRPO) on the multi-task dataset mixture described in the Methods. This stage aligns the model with diagnosis classification, lesion lo-

calization, report generation, and SonoVQA while preserving the structured `<think>...</think><answer>...</answer>` response format.

We use the open-source EasyR1 framework [2] for GRPO training. During GRPO, the vision tower remains frozen, and the language model policy is updated through reinforcement alignment. Rollout generation is accelerated with vLLM, and the actor model generates a group of $n = 5$ responses for each prompt to compute group-relative advantages.

To encourage workflow-consistent reasoning and reliable structured outputs across heterogeneous tasks, we use the following prompt template during rollout generation.

Prompt template for GRPO alignment

You are a medical ultrasound expert.
Reason like a clinician performing ultrasound interpretation when reasoning is required. Follow a natural clinical order when applicable: scanning protocol → anatomical system → target organ → diagnosis.

Output format (always):
`<think>...</think><answer>...</answer>`
Rules:
- `<think>` should reflect clinically meaningful reasoning when needed.
- Do not force unnecessary steps if the question does not require them.
- No extra text outside the tags.

You MUST output using EXACTLY this format with NO line breaks between tags:
`<think>...</think><answer>...</answer>`

For each prompt, the policy samples a group of five responses, which are scored by the unified reward function described in the Methods. Group-relative advantages are computed within each response group, and the policy is updated with a Kullback-Leibler (KL) penalty against the frozen reference policy. This implementation follows the critic-free GRPO setting, in which no separate value model is trained. During alignment, all task types share the same response schema, while task-specific reward functions are applied to the content inside the `<answer>` block.

The detailed hyperparameters for the GRPO stage, including rollout, KL regularization, sequence length, and distributed training settings, are summarized in Table S4.

1.3 Evaluation protocols, reward definitions, and core prompt templates

To support reproducible comparisons across baseline models and scalable evaluation of unstructured generative tasks, we standardized the prompting, answer extraction, and VLM-judge protocols used in all evaluations. This section provides the prompt templates, answer-extraction rules, task-specific reward definitions, and VLM-judge rubrics used for baseline inference, GRPO alignment, and generative-task scoring.

1.3.1 Prompt templates for baseline evaluation

For baseline evaluation, all models were queried with a unified instruction template that requested a concise image-based rationale followed by a final answer. This standardized output format reduces variation caused by model-specific response styles and facilitates consistent answer extraction across closed-source, open-weight, and medical-domain VLMs.

The prompt template provided to baseline models was as follows:

Table S4: Hyperparameters for Stage 2: GRPO alignment.

Hyperparameter	**Value**
Reinforcement learning (RL) Algorithm	GRPO
Framework	EasyR1 [2]
Base Policy	SonoReasoner-Init (8B/32B)
Frozen Components	Vision Tower
Trainable Components	Language Model Policy
Distributed Strategy	FSDP Full Sharding
Actor Offloading	Parameter and Optimizer Offloading
Reference Policy Offloading	CPU Offloading
Rollout Engine	vLLM [3]
Reward Type	Sequential Reward Function
Optimizer	AdamW (bfloat16 (BF16))
Learning Rate	1.0×10^{-6}
Weight Decay	0.01
Warmup Ratio	0.0
Max Gradient Norm	1.0
Rollout Group Size (n)	5
Rollout Batch Size	7
Actor Global Batch Size	7
Micro Batch Size per Device	1
Rollout Temperature	1.0
Rollout Top-p	1.0
Validation Temperature / Top-p	0.6 / 0.95
KL Penalty Estimator	Low-variance KL
KL Coefficient (β)	0.01
Max Prompt Length (Tokens)	4,096
Max Response Length (Tokens)	2,048
Image Max Pixels	262,144
Precision	BF16
Gradient Checkpointing	Enabled
Training Epochs	5
Number of GPUs	8
Random Seed	1

Prompt template for baseline models

You are a medical ultrasound expert.
You MUST output using exactly this format with no line breaks between tags:
<think>Your short and concise image-based reasoning.</think><answer>Your final answer.</answer>.
Do not repeat or rewrite any part of the question, the options, the prompt, or the input text.

Lenient answer extraction strategy. Generative models may occasionally deviate from the requested XML-style format. To avoid unfairly penalizing minor formatting deviations, we used a regular-expression parser with hierarchical fallbacks. When the <answer> tag was missing or malformed, the parser attempted to extract the final semantic answer from the response. All baseline evaluations were conducted with a low decoding temperature ($\tau = 0.1$) and task-dependent maximum token limits, such as 512 or 1024 tokens, to reduce sampling variability and avoid truncation.

1.3.2 Evaluation details for generative and reasoning tasks

As detailed in the Methods, the unified reward function combines task correctness, reasoning structure, and format adherence. Unless otherwise specified, the reward weights follow the main Methods: 0.7 for R_{task}, 0.2 for $R_{\text{reasoning}}$, and 0.1 for R_{format}. All reward components are normalized to $[0, 1]$ before weighted aggregation.

For discriminative tasks, R_{task} is computed using deterministic task-specific criteria, such as canonical label matching for classification and average precision at an IoU threshold of 0.50 (AP50)-based localization scoring for lesion detection. For generative tasks, including report generation and open-ended VQA, semantic evaluation is required. We therefore use VLM judges that take both the ultrasound image and the generated text as input. Specifically, we use Qwen3-VL-Instruct-32B [4] to compute $R_{\text{reasoning}}$, which evaluates whether the <think> block follows the expected clinical reasoning order. We use GPT-4o [5] to compute VLM Scores for report generation and open-ended VQA. During GRPO, these VLM-judge scores are used as semantic components of R_{task} for generative tasks. During evaluation, the same judge protocols are used to report VLM Score. In all judge prompts, the original ultrasound image is included to support visually grounded assessment.

Table S5 summarizes the task-specific reward definitions used during GRPO alignment.

Table S5: Task-specific reward definitions used during GRPO alignment.

Task	**Task reward definition**
Diagnosis classification	Canonical matching between the predicted answer and the ground-truth diagnostic category after label normalization.
Lesion localization	Predicted bounding boxes were parsed from the <answer> block and matched to ground-truth boxes. The reward combined AP_{50}-style correctness with the mean IoU of correctly matched boxes.
Report generation	Hybrid semantic reward combining BERTScore and a multimodal VLM-judge score based on the ultrasound image and reference report.
Clinical VQA	Exact option matching for MCQs; hybrid BERTScore and VLM-judge scoring for yes/no and open-ended questions.

Clinical reasoning order judge ($R_{\text{reasoning}}$) To compute $R_{\text{reasoning}}$, we use Qwen3-VL-Instruct-32B as a VLM judge. The judge evaluates whether the <think> block follows the protocol-system-organ-diagnosis workflow and whether the mentioned reasoning steps are visually plausible given the ultrasound image.

VLM judge: reasoning order score via Qwen3-VL-Instruct-32B

Input data:
- Ultrasound image(s): [Interleaved image tokens]
- Question: {question}
- Predicted thinking: {pred_think}
- Predicted answer: {pred_answer}

Role: You are an expert clinical ultrasound reasoning judge.
Task: Evaluate the structural integrity and visual factuality of the reasoning trajectory. Do not assume or introduce information not explicitly stated.

Evaluation criteria (score 0.00–1.00):
Evaluate whether the predicted thinking follows the clinical reasoning order: protocol → system → organ → diagnosis.

- **Explicit and visual plausibility:** Protocol, system, or organ should be counted only if explicitly stated as a reasoning step and visually plausible based on the image. Do not give credit for implied information or generic workflow summaries.
- **Order requirement:** Earlier diagnostic stages should logically precede later stages.
- **Hallucination penalty:** Do not judge the medical correctness of the final diagnosis, but penalize severe visual hallucinations within the reasoning steps.

Output format:
Score: <0.00–1.00>
Reason: <≤4 concise sentences>

Clinical efficacy entity sets for report generation For report generation, Clinical Efficacy (CE) metrics assess whether the generated report preserves key clinical information rather than only measuring textual similarity. Following prior ultrasound report generation evaluation protocols [6], we define organ-specific entity sets based on sonographer-recommended clinical findings. Each generated and reference report is converted into a multi-label vector over the corresponding entity set, where each entity is labelled as 1 if it is mentioned and 0 otherwise. CE-F1 and CE-Recall are computed by comparing the entity vectors extracted from the generated and reference reports. The entity sets used in this study are listed in Table S6.

Table S6: Organ-specific clinical entities used for Clinical Efficacy (CE) evaluation in report generation.

Organ / dataset	Clinical entities
Liver	liver, hepatic capsule, echogenicity, hepatic vein, kidney, intrahepatic duct, bile duct, gallbladder, liver margin, pancreas, pancreatic duct, lesion, spleen, CDFI, nodule.
Thyroid	thyroid gland, glandular tissue, echogenicity, lesion, CDFI, lymph node, border, shape, nodule, left lobe, right lobe, thyroid margin.
Breast	breast, gland, CDFI, axilla, echogenicity, nodule, lymph node, mammary duct, lesion, subcutaneous fat layer, tumor.

Report generation judge for VLM Score For report generation, the VLM judge provides a semantic score in $[0, 1]$ by comparing the generated report with the reference report using the ultrasound image as visual context. This score is used as the VLM Score during evaluation and as a semantic component of R_{task} during GRPO alignment.

VLM judge: report generation via GPT-4o

Input data:
- Ultrasound image(s): [Interleaved image tokens]
- Reference report: {reference}
- Generated report: {response}

Role: You are an experienced ultrasound physician.
Task: Compare the generated report with the reference report based on the provided ultrasound image.

Evaluation criteria (score 0.00–1.00):

- **Factual consistency:** Does the generated report accurately reflect the visual findings in the ultrasound image?
- **Information coverage:** Are the key pathological findings and anatomical structures from the reference report preserved?
- **Hallucination penalty:** Are there fabricated lesions, incorrect measurements, or contradictions?
- **Clinical logic:** Is the wording professional and is the reporting logic clinically coherent?

Output format:
Score: <0.00–1.00>
Reason: <≤4 concise sentences>

Open-ended VQA judge for VLM Score For yes/no and open-ended VQA, the VLM judge provides a semantic score in $[0, 1]$ by comparing the predicted answer with the reference answer using the ultrasound image and question as context. This score is used as the VLM Score during evaluation and as a semantic component of R_{task} during GRPO alignment.

VLM judge: open-ended VQA via GPT-4o

Input data:
- Ultrasound image(s): [Interleaved image tokens]
- Question: {problem}
- Reference answer: {gt_answer}
- Predicted answer: {pred_answer}

Role: You are an experienced ultrasound physician.
Task: Evaluate whether the predicted answer is semantically equivalent to the reference answer based on the provided ultrasound image and question.

Evaluation criteria (score 0.00–1.00):

- **Semantic equivalence:** Does the prediction convey the same clinical meaning as the reference answer?
- **Visual grounding:** Is the predicted feature, such as echogenicity or vascularity, consistent with the ultrasound image?

Output format:
Score: <0.00–1.00>
Reason: <≤4 concise sentences>

1.4 Baseline models and inference settings

For reproducibility, we report the model identifiers or released checkpoints used for baseline evaluation in Table S7. Proprietary VLMs are listed by their reported model names, whereas open-weight and medical-domain VLMs are listed by the checkpoint identifiers used in our experiments.

All baselines were queried using the unified baseline prompt. Unless otherwise specified, inference was conducted with temperature $\tau = 0.1$. The maximum generation length was set to 512 tokens for diagnosis classification, lesion localization, and multiple-choice VQA, and to 1024 tokens for report generation and open-ended VQA. Images were converted to RGB and supplied as image inputs. Predictions were parsed from the `<answer>` block using the answer-extraction strategy described in Section 1.3.1.

Table S7: Baseline models and released checkpoints evaluated in this study.

Model	**Model identifier / checkpoint**
Proprietary VLMs	
GPT-5.1	GPT-5.1
GPT-4.1	GPT-4.1
Claude-Opus-4.5	Claude-Opus-4.5
Gemini-3-Pro	Gemini-3-Pro
Gemini-2.5-Pro	Gemini-2.5-Pro
Grok-4	Grok-4
General-purpose open-weight VLMs	
Qwen3-VL-8B	`Qwen/Qwen3-VL-8B-Instruct`
Qwen3-VL-32B	`Qwen/Qwen3-VL-32B-Instruct`
InternVL3.5-8B	`OpenGVLab/InternVL3_5-8B`
InternVL3.5-14B	`OpenGVLab/InternVL3_5-14B`
InternVL3.5-38B	`OpenGVLab/InternVL3_5-38B`
Medical-domain VLMs	
MedGemma-4B	`google/medgemma-4b-it`
MedGemma-27B	`google/medgemma-27b-it`
HuatuoGPT-V-7B	`FreedomIntelligence/HuatuoGPT-Vision-7B`
HuatuoGPT-V-34B	`FreedomIntelligence/HuatuoGPT-Vision-34B`
Lingshu-7B	`lingshu-medical-mllm/Lingshu-7B`
Lingshu-32B	`lingshu-medical-mllm/Lingshu-32B`
Hulu-Med-7B	`ZJU-AI4H/Hulu-Med-7B`
Hulu-Med-32B	`ZJU-AI4H/Hulu-Med-32B`
Baichuan-M2-32B	`baichuan-inc/Baichuan-M2-32B`

1.5 Human–AI assistance protocol

Human–AI assistance analysis was conducted on the disease-diagnosis questions from the held-out SonoVQA test split. A sonographer with three years of clinical ultrasound experience first answered these questions independently. For each question, the reader was shown the ultrasound image, the question, and the candidate answer options, without access to model predictions or ground-truth labels. In a second assisted round, the same reader answered the questions with access to the corresponding SonoReasoner-32B output as a reference. No time limit was imposed in either round. Accuracy was computed at the question level by comparing the selected option with the ground-truth answer. Given that this analysis involved a single sonographer, the results were interpreted as exploratory evidence of assisted reading rather than as a multi-reader clinical validation.

Table S8 summarizes the subgroup statistics for the unaided sonographer, SonoReasoner-32B alone, and the AI-assisted sonographer.

Table S8: Subgroup statistics for the human–AI assistance analysis on SonoVQA disease-diagnosis questions. Values are accuracies in percent. Δ denotes AI-assisted minus unaided sonographer accuracy.

Disease type	#QAs	Unaided	SonoReasoner	AI-assisted	Δ
Overall	**336**	**55.65**	**49.70**	**59.82**	**+4.17**
Tumor	227	62.11	46.70	63.00	+0.89
Obstructive disease	54	50.00	61.11	66.67	+16.67
Inflammatory disease	29	31.03	44.83	31.03	+0.00
Trauma	14	14.29	28.57	28.57	+14.28
Diffuse abnormality	12	66.67	91.67	75.00	+8.33

2 Supplementary Results

2.1 Impact of SFT data scale on reasoning initialization

To examine how the amount of reasoning-initialization data affects downstream performance, we conducted an ablation study by varying the volume of SonoCorpus used for SFT. The language model parameters were updated using 10%, 50%, and 100% subsets of the approximately 7.0 million retained reasoning trajectories, while the vision encoder and multimodal projector remained frozen. The resulting models were evaluated across downstream clinical tasks and hierarchical SonoVQA levels.

As shown in Fig. S2a, increasing the SFT data scale consistently improves average performance across diagnosis classification, lesion localization, report generation, and clinical VQA. The gains are observed not only in generative tasks, such as report generation and VQA, but also in lesion localization, suggesting that structured reasoning initialization can benefit both semantic and spatial clinical tasks.

A level-wise analysis on SonoVQA further shows different scaling patterns across the reasoning hierarchy (Fig. S2b). Lower-level tasks, such as protocol and system recognition, show relatively early saturation, whereas higher-level tasks, particularly lesion categorization and disease diagnosis, continue to improve as more SFT data are used. These results suggest that basic acquisition and anatomical recognition can be learned with fewer examples, while disease-level reasoning benefits more from large-scale structured reasoning supervision.

2.2 Training dynamics of GRPO alignment

To examine how the reasoning reward affects GRPO alignment, we monitored training dynamics on a held-out validation set. We tracked both the reasoning reward ($R_{\text{reasoning}}$), which evaluates workflow consistency in the `<think>` block, and the task reward (R_{task}), which measures task-specific correctness or semantic quality. We compared the full SonoReasoner policy with an ablated variant trained without the reasoning reward (w/o $R_{\text{reasoning}}$).

As shown in Fig. S3a, the full model rapidly improves $R_{\text{reasoning}}$ during the early training phase and maintains a high reasoning score thereafter. This indicates that the model quickly adapts to the expected protocol-system-organ-diagnosis reasoning order. In parallel, R_{task} increases steadily across training steps. In contrast, the ablated variant shows a much lower $R_{\text{reasoning}}$ throughout training and reaches a lower R_{task} by the end of alignment. These results suggest that the reasoning reward helps stabilize the desired clinical reasoning structure and supports downstream task optimization.

We further analyzed the task-wise evolution of R_{task} across classification, lesion localization, report generation, and SonoVQA open-ended evaluation (Fig. S3b). The full model consistently outperforms the ablated variant across all four task formats, indicating that the benefit of $R_{\text{reasoning}}$ is not limited to a single task type. These findings support the use of reasoning-structure guidance as a complementary training signal during multi-task GRPO alignment.

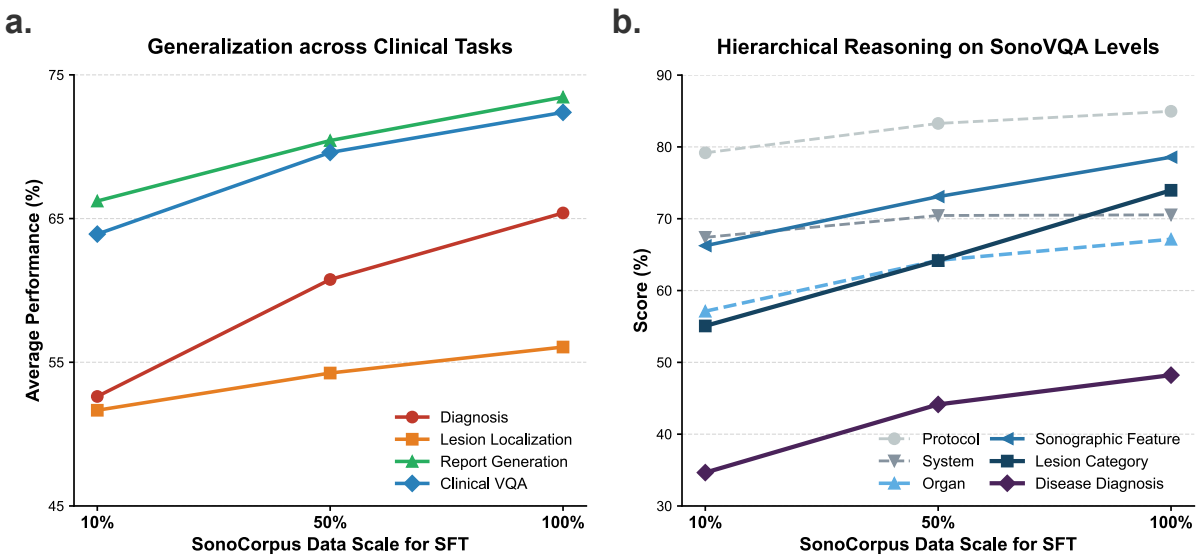


Figure S2: **Impact of SFT data scale on reasoning initialization.** **a**, Average performance across downstream clinical tasks under different SonoCorpus data scales for SFT. Increasing the amount of SFT data improves performance across diagnosis classification, lesion localization, report generation, and clinical VQA. **b**, Level-wise performance on SonoVQA. Protocol and system recognition show relatively early saturation, whereas higher-level reasoning tasks, including lesion categorization and disease diagnosis, continue to benefit from larger-scale reasoning initialization.

2.3 Qualitative examples across hierarchical and downstream tasks

To complement the quantitative analyses, we provide additional qualitative examples illustrating how SonoReasoner applies the protocol–system–organ reasoning structure across hierarchical SonoVQA and downstream ultrasound tasks. These examples show how the model anchors each image to the appropriate acquisition protocol and anatomical context before producing task-specific outputs, including VQA answers, diagnostic classifications, lesion bounding boxes, and structured ultrasound reports.

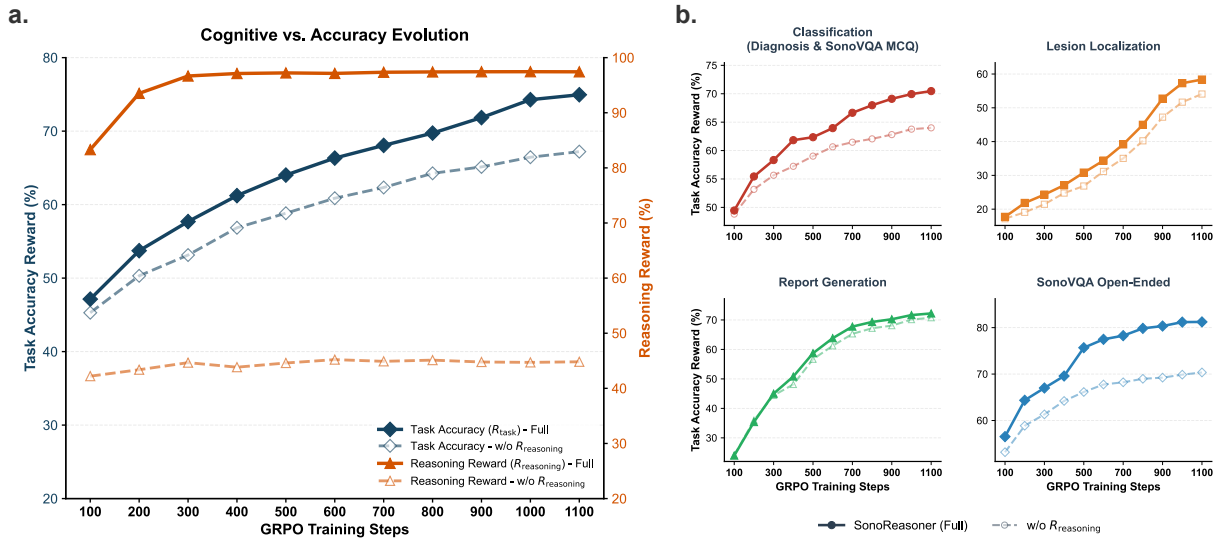
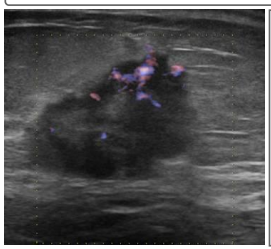


Figure S3: **Training dynamics during GRPO alignment.** Evaluations are performed on a held-out validation set across training steps. **a**, Comparison between the full model and the variant without the reasoning reward. The full model shows a rapid increase in $R_{\text{reasoning}}$ and a steady improvement in R_{task}, whereas removing $R_{\text{reasoning}}$ leads to lower reasoning and task rewards throughout training. **b**, Task-wise evolution of R_{task} across classification, lesion localization, report generation, and SonoVQA open-ended evaluation. The full model consistently achieves higher rewards than the ablated variant across all task formats.

a.

SonoVQA Case 1 | Breast carcinoma with Doppler vascularity
Pathway: Superficial protocol → Superficial tissues system → Breast → Tumor → Breast carcinoma

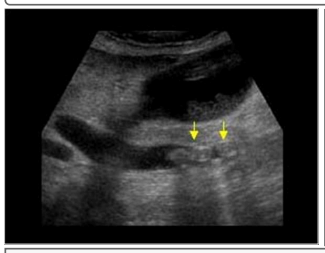


Level	Question focus	SonoReasoner answer
Protocol	Scanning protocol	✓ Superficial protocol
System	Anatomical system	✓ Superficial tissues system
Organ	Target organ	✓ Breast
Feature	Intralesional Vascularity	✓ Vessel poles visible
Lesion	Disease category	✓ Tumor
Diagnosis	Specific diagnosis	✓ Breast carcinoma Doppler criteria

Representative reasoning trace:
...obtained using a superficial ultrasound protocol and shows a breast lesion within the superficial tissues system. Internal Doppler vascular signals are present within the hypoechoic lesion. The vascularized mass-like appearance supports a tumor category and is consistent with breast carcinoma-related Doppler findings...

b.

SonoVQA Case 2 | Sludge and gallstones in the extrahepatic bile ducts
Pathway: Abdominal protocol → HPB system → Gallbladder & bile ducts → Obstructive disease → Extrahepatic bile duct sludge/gallstones



Level	Question focus	SonoReasoner answer
Protocol	Scanning protocol	✓ Abdominal protocol
System	Anatomical system	✓ HPB system
Organ	Target organ	✓ Gallbladder and bile ducts
Feature	Acoustic shadowing	✓ Acoustic shadows visible
Lesion	Disease category	✓ Obstructive disease
Diagnosis	Specific diagnosis	✓ Bile duct stones/sludge

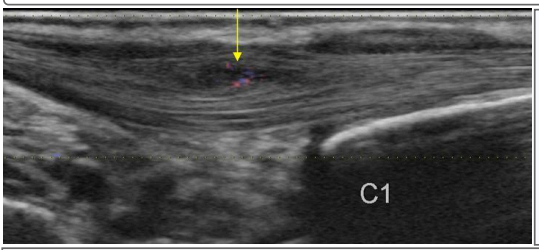
Representative reasoning trace:
...obtained using an abdominal ultrasound protocol and shows the gallbladder and bile ducts within the hepatopancreatobiliary system. Hyperechoic stone-like foci with posterior acoustic shadowing are present in a tubular biliary structure. These findings support obstructive biliary disease and are consistent with sludge and gallstones in the extrahepatic bile ducts...

Figure S4: **Additional hierarchical SonoVQA reasoning examples (I).** Representative SonoVQA cases showing complete reasoning across protocol, anatomical system, target organ, sonographic feature, lesion category, and disease diagnosis levels. **a**, A breast carcinoma case acquired under a superficial protocol, where SonoReasoner identifies the superficial tissues system, breast target, intralesional vascularity, tumor category, and breast carcinoma-related Doppler findings. **b**, An extrahepatic bile duct sludge/gallstone case acquired under an abdominal protocol, where the model follows the hepatopancreatobiliary pathway, identifies gallbladder and bile duct involvement, recognizes acoustic shadowing, and links the findings to obstructive disease.

a.

SonoVQA Case 3 | De Quervain tenosynovitis

Pathway: Musculoskeletal protocol → Musculoskeletal system → Hand/wrist → Inflammatory disease → De Quervain tenosynovitis



Level	Question focus	SonoReasoner answer
Protocol	Scanning protocol	✓ Musculoskeletal protocol
System	Anatomical system	✓ Musculoskeletal system
Organ	Target organ	✓ Hand/wrist
Feature	Vascularity	✓ Vascularity present
Lesion	Disease category	✓ Inflammatory disease
Diagnosis	Specific diagnosis	✓ De Quervain tenosynovitis

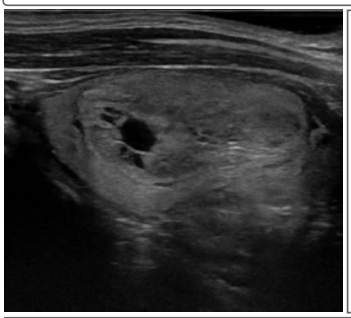
Representative reasoning trace:

...obtained using a musculoskeletal ultrasound protocol and shows the hand/wrist tendon region. A hypoechoic area with Doppler vascular signal is visible around the tendon sheath. The vascular and inflammatory appearance supports an inflammatory tendon-sheath process and is consistent with De Quervain tenosynovitis...

b.

SonoVQA Case 4 | Benign thyroid nodule with cystic degeneration

Pathway: Superficial protocol → Superficial tissues system → Thyroid → Tumor → Thyroid benign lesions



Level	Question focus	SonoReasoner answer
Protocol	Scanning protocol	✓ Superficial protocol
System	Anatomical system	✓ Superficial tissues system
Organ	Target organ	✓ Thyroid
Feature	Cystic component	✓ Cystic changes visible
Lesion	Disease category	✓ Tumor
Diagnosis	Specific diagnosis	✓ Thyroid benign lesions

Representative reasoning trace:

...obtained using a superficial ultrasound protocol and shows a thyroid lesion within the superficial tissues system. A well-defined hypoechoic thyroid nodule with internal anechoic/cystic areas is visible. The circumscribed margin and cystic degeneration support a tumor-like lesion category and are consistent with thyroid benign lesions...

Figure S5: **Additional hierarchical SonoVQA reasoning examples (II). a**, A De Quervain tenosynovitis case acquired under a musculoskeletal protocol, where SonoReasoner localizes the hand/wrist tendon region, identifies Doppler vascularity, and supports an inflammatory tendon-sheath process. **b**, A benign thyroid nodule case acquired under a superficial protocol, where the model identifies the superficial tissues system, thyroid target, cystic components, tumor-like lesion category, and thyroid benign lesion diagnosis. Together with Fig. S4, these examples extend the main-text VQA visualization by showing complete hierarchical reasoning traces across diverse anatomical systems and disease categories.

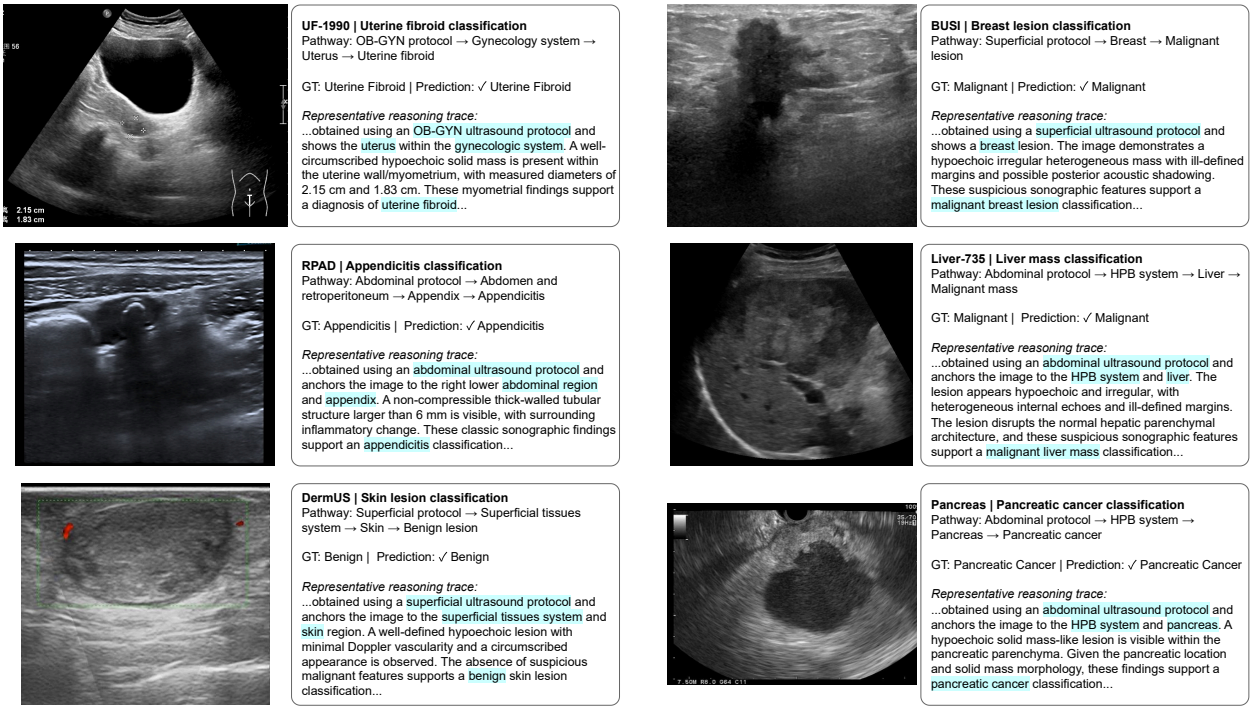
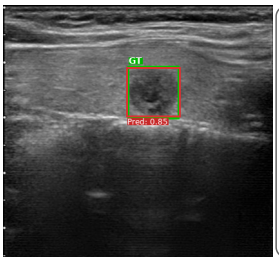
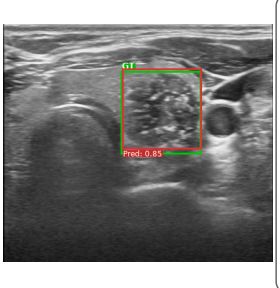


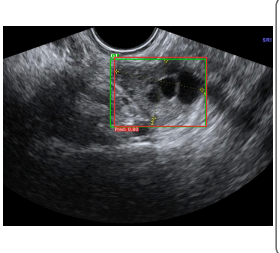
Figure S6: **Additional diagnosis classification examples.** Representative correct classification examples across six downstream ultrasound datasets. Each panel shows the input ultrasound image, task-specific reasoning pathway, ground-truth label, model prediction, and a concise reasoning trace. SonoReasoner first anchors the image to the corresponding protocol and anatomical context, such as obstetrics and gynecology (OB-GYN)–uterus for uterine fibroid classification, abdominal–hepatopancreatobiliary (HPB)–liver for liver mass classification, or superficial–skin for skin lesion classification, before using task-relevant sonographic features to support the final category. These examples illustrate anatomy-guided classification across gynecologic, breast, appendiceal, hepatic, dermatologic, and pancreatic ultrasound tasks.



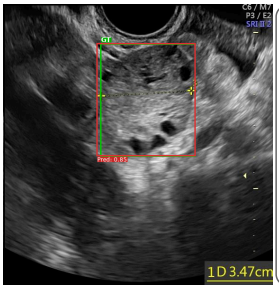
TN3K | Thyroid nodule localization
Pathway: Superficial protocol → Head and neck system → Thyroid → Localization
Representative reasoning trace:
...obtained using a superficial ultrasound protocol and anchors the image to the head and neck system and thyroid gland. A well-defined rounded lesion is visible within the thyroid parenchyma, supporting precise lesion localization...



TN3K | Thyroid nodule localization
Pathway: Superficial protocol → Head and neck system → Thyroid → Localization
Representative reasoning trace:
...obtained using a superficial ultrasound protocol and anchors the image to the head and neck system and thyroid gland. A hypoechoic, heterogeneous thyroid nodule is visible within the thyroid parenchyma, with relatively clear lesion boundaries. These anatomical and morphological cues support accurate bounding-box localization...

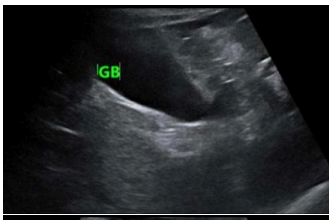


MMOTU2D | Ovarian mass localization
Pathway: OB-GYN protocol → Gynecology system → Ovary → Localization
Representative reasoning trace:
...obtained using an OB-GYN ultrasound protocol and anchors the image to the gynecology system and ovary. A well-defined rounded adnexal lesion with heterogeneous internal echoes is visible near the center of the image. The relatively clear lesion boundary supports accurate bounding-box localization...



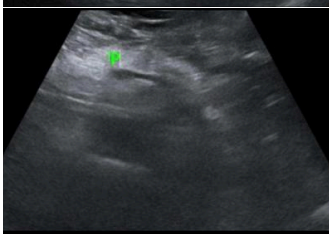
MMOTU2D | Ovarian mass localization
Pathway: OB-GYN protocol → Gynecology system → Ovary → Localization
Representative reasoning trace:
...obtained using an OB-GYN ultrasound protocol and anchors the image to the gynecology system and ovary. A well-defined adnexal lesion is visible near the center of the image, with a relatively clear lesion boundary. These anatomical and morphological cues support accurate bounding-box localization...

Figure S7: **Additional lesion localization examples.** Representative lesion localization examples from TN3K and MMOTU2D. Ground-truth and predicted bounding boxes are overlaid on the ultrasound images, and each example includes the corresponding reasoning pathway and a concise localization rationale. SonoReasoner first identifies the acquisition protocol and target organ, such as superficial head-and-neck thyroid imaging or OB-GYN ovarian imaging, before localizing the lesion boundary. The close agreement between ground-truth and predicted boxes illustrates how anatomical grounding supports spatial localization rather than relying only on visually salient image regions.



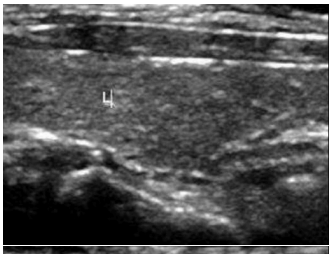
US-Report | Liver report generation
Pathway: Abdominal protocol → Abdomen and retroperitoneum → Liver → Report Generation

Representative reasoning trace:
...obtained using an abdominal ultrasound protocol and anchors the images to the abdomen and retroperitoneum with liver-focused assessment. The liver shows mildly increased and heterogeneous parenchymal echogenicity without a definite focal mass, while the bile ducts and portal vein are not dilated. These multi-organ abdominal findings support structured liver ultrasound report generation...



Ground Truth: The liver appears full in shape. The hepatic parenchymal echogenicity is finely increased and mildly heterogeneous. No intrahepatic or extrahepatic bile duct dilatation is observed. The main portal vein is not widened, and no definite focal liver lesion is identified. The gallbladder, pancreas, and spleen show no obvious abnormality.

Prediction: The liver is normal in shape with a smooth surface. The hepatic parenchymal echogenicity is diffusely mildly increased and slightly heterogeneous, without a definite focal lesion. The main portal vein is not dilated. The gallbladder wall is not thickened, and no obvious abnormal echo or bile duct dilatation is observed. The pancreas and spleen show no definite focal lesion.

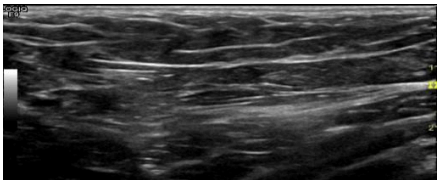


US-Report | Thyroid report generation
Pathway: Superficial protocol → Head and neck system → Thyroid → Report Generation

Representative reasoning trace:
...obtained using a superficial ultrasound protocol and anchors the images to the head and neck system and thyroid gland. A hypoechoic nodule is identified in the right thyroid lobe, with clear margins, regular morphology, and no obvious Doppler flow signal. These thyroid and CDFI findings support structured thyroid ultrasound report generation...

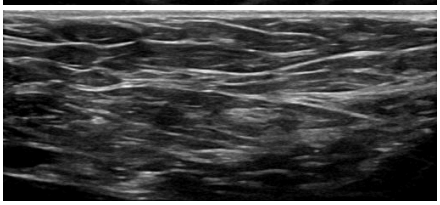
Ground Truth: The thyroid gland is normal in size and shape, with heterogeneous parenchymal echogenicity. A hypoechoic nodule is observed in the middle portion of the right thyroid lobe, measuring approximately [measurement]. The nodule has clear margins and a regular shape, with peripheral echogenic foci and posterior comet-tail artifact. CDFI shows no detectable blood flow signal. No obvious enlarged cervical lymph nodes are observed.

Prediction: The bilateral thyroid lobes show heterogeneous echogenicity. A hypoechoic nodule is observed in the right thyroid lobe, measuring approximately [measurement], with clear margins and a regular shape. CDFI shows no abnormal blood flow signal within the nodule or thyroid gland. No obvious enlarged cervical lymph nodes are observed.



US-Report | Breast report generation
Pathway: Superficial protocol → Breast → Report Generation

Representative reasoning trace:
...obtained using a superficial ultrasound protocol and anchors the images to the breast. The glandular echotexture appears mildly heterogeneous, without ductal dilatation or a discrete focal mass. These bilateral breast findings support structured breast ultrasound report generation...



Ground Truth: The glandular architecture of both breasts is mildly disordered. No ductal dilatation is observed, and no definite space-occupying lesion is identified in either breast.

Prediction: The bilateral breast parenchyma shows heterogeneous echogenicity. No obvious ductal dilatation is observed, and no definite space-occupying lesion is identified in either breast.

Figure S8: **Additional report generation examples.** Representative structured report generation examples for liver, thyroid, and breast ultrasound. For each case, the figure shows the input image pair, reasoning pathway, concise model rationale, reference report, and generated report translated into English for readability. SonoReasoner anchors the images to the appropriate protocol and anatomical target, identifies key reportable findings such as hepatic echogenicity, thyroid nodule morphology and CDFI findings, or breast ductal and parenchymal features, and generates a structured report that preserves the clinically relevant entities from the reference report. These examples complement the quantitative report-generation metrics by illustrating how the model converts multi-image ultrasound evidence into clinically organized report text.

Supplementary References

[1] Zheng, Y., Zhang, R., Zhang, J., Ye, Y. & Luo, Z. Llamafactory: Unified efficient fine-tuning of 100+ language models. In *Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 3: system demonstrations)*, 400–410 (2024).

[2] Zheng, Y. *et al.* EasyR1: An efficient, scalable, multi-modality reinforcement learning training framework. `https://github.com/hiyouga/EasyR1` (2025). Accessed: 2026-05-08.

[3] Kwon, W. *et al.* Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th symposium on operating systems principles*, 611–626 (2023).

[4] Bai, S. *et al.* Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631* (2025).

[5] Hurst, A. *et al.* Gpt-4o system card. *arXiv preprint arXiv:2410.21276* (2024).

[6] Li, J. *et al.* Ultrasound report generation with cross-modality feature alignment via unsupervised guidance. *IEEE Transactions on Medical Imaging* **44**, 19–30 (2024).