ULTRAG-BENCH: A MULTI-TASK BENCHMARK FOR ASSESSING LARGE VISION-LANGUAGE MODELS ON PIXEL-LEVEL EVIDENCE GROUNDING IN ULTRASOUND

Anonymous authors
Paper under double-blind review

ABSTRACT

Ultrasound is one of the most widely used medical imaging modalities, and recent large vision-language models(VLMs) have shown increasing capabilities in ultrasound image understanding. However, these models fail to provide pixel-level visual evidence aligned with their semantic predictions, and their fine-grained grounding capability in ultrasound remains largely unclear. We introduce UltraG-Bench, a large-scale multi-task benchmark for evaluating pixel-level evidence grounding in ultrasound. UltraG-Bench is built by annotating 40 public ultrasound segmentation datasets spanning 13 anatomical categories, and comprises three progressive tasks: instruction-guided segmentation, evidence-grounded VQA, and evidence-grounded report generation, with 331125, 666779, and 138832 annotations, respectively. Comprehensive evaluation of 14 state-of-the-art models reveals a substantial gap between semantic understanding and fine-grained pixel-level localization. We further propose UltraG-Agent, which combines the semantic reasoning capabilities of a VLM with the ultrasound-specific segmentation capability of UltraSAM3. Experiments show that UltraG-Agent substantially improves both semantic prediction and pixel-level visual grounding.

1 INTRODUCTION

Ultrasound has become one of the most widely used medical imaging modalities in clinical practice owing to its lack of ionizing radiation, real-time imaging capability, and relatively low cost. Compared with CT and MRI, ultrasound devices are more portable and less expensive, making them highly accessible across diverse clinical settings Fleming et al. (2021). Accurate recognition, localization, and interpretation of anatomical categories and lesions in ultrasound images are fundamental to clinical diagnosis and decision-making Guo et al. (2026).

Nowadays, supervised segmentation methods and general foundation models have continuously improved pixel-level localization of organs and lesions Ronneberger et al. (2015); Ma et al. (2024). On the other hand, vision-language models (VLMs) and multimodal large language models (MLLMs) are increasingly being applied to ultrasound image understanding, including tasks such as visual question answering (VQA) and report generation Guo et al. (2026). However, a substantial gap remains between pixel-level visual localization and high-level semantic understanding in current models. Conventional segmentation models can provide precise spatial boundaries but generally lack open-ended language understanding and semantic generation capabilities, whereas VLMs and MLLMs can answer questions or generate image descriptions but often fail to provide pixel-level visual evidence that is aligned with their semantic predictions Myhre et al. (2026); Jin et al. (2026). This gap is particularly critical in medical applications, where clinical reasoning depends not only on "what the model predicts", but also on "which regions in the image support that prediction." However, this gap has not yet been systematically evaluated by existing benchmarks, leaving the extent of the disconnect between semantic understanding and pixel-level visual grounding in current models poorly understood.

Existing benchmarks often evaluate semantic understanding and pixel-level visual localization in isolation Lau et al. (2018); Liu et al. (2021); He et al. (2020); Hu et al. (2024): segmentation bench-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

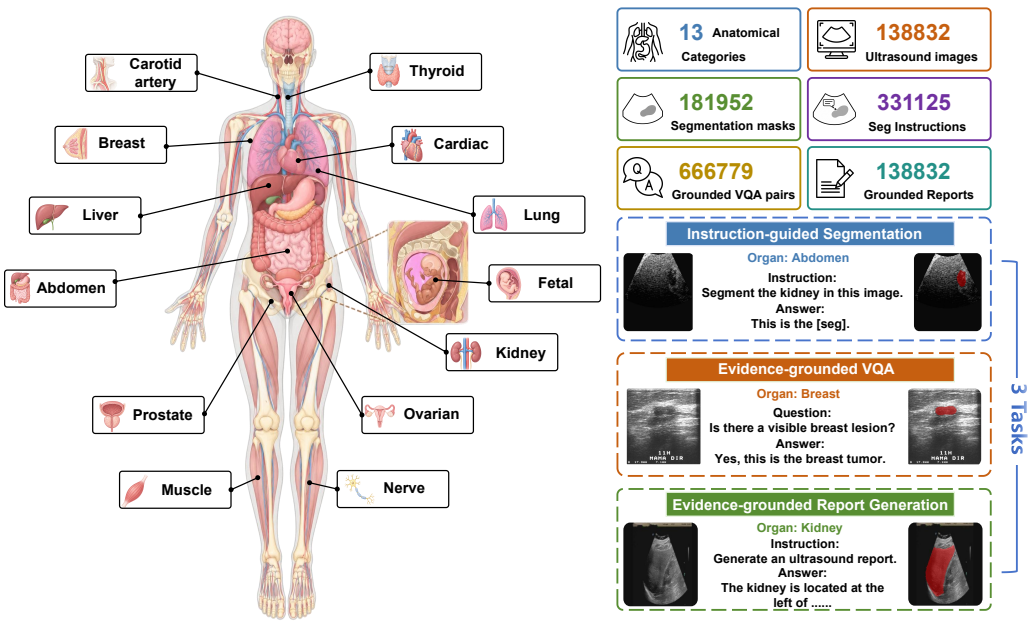


Figure 1: Overview of UltraG-Bench, including 13 anatomical categories, benchmark statistics, and representative examples of the three tasks: Instruction-guided Segmentation, Evidence-grounded VQA, and Evidence-grounded Report Generation.

marks primarily measure the overlap between predicted and ground-truth masks using metrics such as IoU and Dice, whereas VQA and report-generation benchmarks mainly assess the correctness of the generated answers or text. As a result, a model may produce the correct answer while relying on the wrong visual region, or generate fluent and semantically plausible descriptions without providing consistent pixel-level evidence for its claims. Recently, visual evidence-grounded datasets such as GEMeX Liu et al. (2025b) and MIMIC-ILS Choi et al. (2026) have emerged, but such evaluation remains lacking in ultrasound, especially at the pixel level.

To address these challenges, we introduce **UltraG-Bench** as shown in Figure 1, a large-scale multi-task benchmark for Pixel-level Evidence Grounding in ultrasound. UltraG-Bench annotates **40** public ultrasound segmentation datasets spanning **13** anatomical categories, including the abdomen, breast, heart, carotid artery, fetus, kidney, liver, lung, muscle, nerve, ovary, prostate, and thyroid, comprising **138,832** ultrasound images and **181,952** segmentation masks. Based on these pixel-level annotations, we construct three tasks from visual perception to semantic understanding: instruction-guided segmentation, which evaluates whether a model can localize the correct target from a natural-language instruction; evidence-grounded VQA, which assesses answer correctness and the accuracy of the corresponding visual evidence; evidence-grounded report generation, which examines whether a model can consistently associate different semantic statements with the appropriate pixel-level regions during report generation. During dataset construction, we explicitly preserve the correspondence between textual statements and the original segmentation masks, while restricting diagnostic, pathological, and treatment-related claims that are unsupported by the available pixel-level annotations, thereby ensuring that each grounded statement is backed by traceable visual evidence. Based on this benchmark, we systematically evaluate 15 state-of-the-art models, including closed-source general-purpose MLLMs, open-source general-purpose MLLMs, medical MLLMs, and pixel-level grounded medical MLLMs.

Furthermore, we observe a clear complementarity between high-level semantic understanding and fine-grained pixel localization in existing models: MLLMs excel at interpreting complex natural-language instructions and generating semantic responses, whereas specialized ultrasound segmentation models provide stronger pixel-level structural localization. This motivates a further question: Can more reliable Pixel-level Evidence Grounding be achieved through collaboration among pre-trained models? To explore this, we construct **UltraG-Agent**, an ultrasound grounding agent that uses an open-source MLLM for instruction understanding, task planning, and response generation,

while employing UltraSAM3 Xu et al. (2026a), an ultrasound concept-driven segmentation model, as the visual module for target segmentation. Experiments show that UltraG-Agent achieves state-of-the-art performance on UltraG-Bench.

The main contributions are as follows:

- **UltraG-Bench:** We annotate 40 public ultrasound segmentation datasets spanning 13 anatomical categories, comprising 138,832 ultrasound images and 181,952 segmentation masks. We construct three tasks, including instruction-guided segmentation, mask-grounded VQA, and mask-grounded report generation, to evaluate capabilities of existing models.
- **Comprehensive evaluation:** Under a unified evaluation framework, we benchmark 16 state-of-the-art models, including closed-source and open-source general MLLMs, medical MLLMs, pixel-level grounded medical MLLMs, and prompt-driven segmentation models.
- **UltraG-Agent:** We propose an ultrasound agent framework that combines the instruction understanding and task-planning capabilities of an open-source MLLM with the specialized ultrasound segmentation capability of UltraSAM3. UltraG-Agent achieves state-of-the-art performance among all evaluated models.

2 Related Work

Medical Evidence-Grounded Benchmarks. Early medical VQA datasets, such as VQA-RAD, SLAKE, PathVQA, and OmniMedVQA, pair medical images with questions and answers Lau et al. (2018); Liu et al. (2021); He et al. (2020); Hu et al. (2024). They provide important benchmarks for medical visual understanding and question answering. As interpretability and visual evidence become more important, later datasets begin to link text with local image regions. MS-CXR, Chest ImaGenome, and PadChest-GR provide annotations that connect radiology findings, anatomical structures, and imaging attributes with corresponding bounding boxes in chest X-rays Boecking et al. (2022); Wu et al. (2021); De Castro et al. (2025). These annotations enable image–text alignment to be evaluated at the region level. For VQA, GEMeX Liu et al. (2025b) introduces textual explanations and visual regions into Med-VQA. GEMeX-RMCoT Liu et al. (2025a) further links intermediate reasoning steps to relevant anatomical regions. MIMIC-ILS Choi et al. (2026) extends referring image segmentation to chest X-rays, providing finer pixel-level evidence. However, existing benchmarks focus on radiological modalities, while ultrasound remains lacking. Moreover, most datasets focus on a single task. In contrast, UltraG-Bench builds a multi-task Pixel-level Evidence Grounding benchmark from 40 public ultrasound segmentation datasets. We have included a comparison with existing medical evidence-grounded benchmarks in Table 1.

Table 1: Comparison of UltraG-Bench with representative medical evidence-grounded datasets.

Benchmark	Modality	Scale	Spatial Evidence	Region Grounding	Instr.-Seg.	Grounded VQA	Grounded Report	Coverage
MS-CXR	CXR	1,162 pairs	Bounding Box	✓	✗	✗	✗	Chest
Chest ImaGenome	CXR	242,072 images	Bounding Box	✓	✗	✗	✗	Chest
PadChest-GR	CXR	4,555 studies	Bounding Box	✓	✗	✗	✓	Chest
GEMeX	CXR	1.61M VQA pairs	Bounding Box	✓	✗	✓	✗	Chest
GEMeX-RMCoT	CXR	202K instances	Bounding Box	✓	✗	✓	✗	Chest
MIMIC-ILS	CXR	1.1M pairs	Pixel Mask	✓	✓	✗	✗	Chest
UltraG-Bench	**Ultrasound**	**1.14M annotations**	**Pixel Mask**	✓	✓	✓	✓	**13 anatomies**

Large Vision-Language Models for Medical Evidence Grounding. Recently, both general-purpose and medical multimodal large language models have made rapid progress in medical image understanding. General MLLMs such as GPT Singh et al. (2025), Claude Anthropic (2025), Qwen-VL Bai et al. (2025), InternVL Wang et al. (2025), and MiniCPM-V Yao et al. (2024) show strong visual understanding, natural-language interaction, and open-ended reasoning. Medical MLLMs such as LLaVA-Med Li et al. (2023), MedGemma Sellergren et al. (2025), and Lingshu Xu et al. (2026b) further improve medical image and clinical understanding through medical pretraining or instruction tuning. However, these models are often unclear whether their predictions are supported by the correct visual evidence. To provide fine-grained spatial outputs, another line of work endows MLLMs with segmentation capability. LISA Lai et al. (2024) connects special tokens such as '[SEG]' to a segmentation decoder and enables reasoning segmentation from complex language

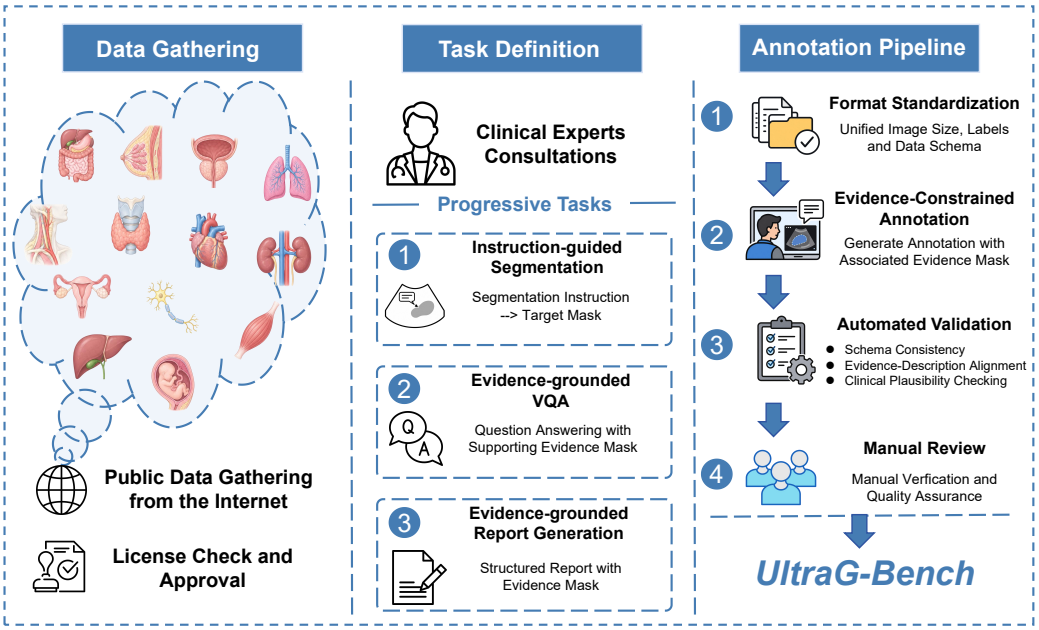


Figure 2: Construction pipeline of the UltraG-Bench, including public data collection, progressive task design, evidence-constrained annotation, automated validation, and expert review.

instructions. Later models such as UniBiomed Wu et al. (2026) extend this idea to medical images and produce pixel-level masks from medical text instructions. However, their true capability for ultrasound Pixel-level Evidence Grounding remains underexplored. More recently, prompt-driven segmentation models such as SAM3 Carion et al. (2026) have emerged. Inspired by this paradigm, we develop UltraG-Agent to evaluate agent-based grounding on our benchmark.

3 ULTRAG-BENCH

Overview. We construct **UltraG-Bench**, a multi-task benchmark for evaluating **Pixel-level Evidence Grounding** in ultrasound images. UltraG-Bench is built upon 40 publicly available ultrasound image segmentation datasets, covering 13 anatomical categories, including the abdomen, breast, heart, carotid artery, fetus, kidney, liver, lung, muscle, nerve, ovary, prostate, and thyroid. Based on the pixel-level segmentation masks provided by the original datasets, we further construct three progressively challenging tasks that bridge visual perception and semantic understanding. The overall construction pipeline is illustrated in Figure 2.

3.1 TASK DEFINITION

Instruction-guided Segmentation. Instruction-guided Segmentation evaluates whether a model can recognize and segment a target structure according to a natural-language instruction. Given an ultrasound image and a textual instruction describing the target object, such as *"Segment the kidney in this image."*, the model is required to understand the target concept specified by the instruction and output the corresponding pixel-level segmentation region. Each instruction is explicitly associated with one or more target masks from the original dataset. Therefore, this task jointly evaluates natural-language instruction understanding and fine-grained spatial localization.

Evidence-grounded VQA. Evidence-grounded VQA extends conventional medical visual question answering by additionally requiring pixel-level visual evidence that supports the generated answer. Given an ultrasound image and a related question, the model is required to generate a textual answer while simultaneously segmenting the region that supports the answer. Unlike conventional VQA benchmarks that evaluate only answer correctness, UltraG-Bench explicitly preserves the correspondence among the question, answer, and original segmentation mask. Consequently, even if

a model produces a semantically correct answer, it is not considered to achieve reliable evidence grounding if the predicted visual evidence does not align with the ground-truth target region. This task therefore jointly evaluates semantic understanding and answer-evidence consistency.

Evidence-grounded Report Generation. Evidence-grounded Report Generation further extends pixel-level visual grounding to long-form text generation. Given an ultrasound image and a report-generation instruction, the model is required to generate a structured ultrasound description and associate image-supported semantic statements with their corresponding pixel-level evidence regions. Specifically, statements in the generated report are linked to target masks, establishing fine-grained **statement–evidence** correspondence. This task evaluates not only the semantic quality of the generated text, but also whether the model can maintain consistency between textual descriptions and pixel-level visual evidence throughout a longer generation process. Diagnostic, pathological, treatment-related, or other clinical inferences that cannot be supported by the available pixel-level annotations are excluded from grounded statements, thereby avoiding supervision based on untraceable visual evidence.

3.2 Dataset Construction

3.2.1 Dataset Collection and Preprocessing

We collect publicly available ultrasound image segmentation datasets. Before incorporating them into UltraG-Bench, we verify their public-use licenses and organize their images, segmentation annotations, and target-category information. The resulting collection covers 13 anatomical categories. Detailed descriptions of our benchmark, including data sources, target categories, and statistics, are provided in the Appendix A.1.

Since the source datasets differ substantially in image organization, mask representation, category naming, and annotation format, we first perform unified preprocessing. Specifically, segmentation annotations from different sources are converted into a COCO-style annotation format. Each image is assigned a unique identifier, and every segmentation mask is explicitly associated with its corresponding target-category name. The standardized representation contains unified image information, target categories, segmentation regions, and annotation identifiers, providing a consistent pixel-level evidence representation for the three downstream tasks. After standardization, UltraG-Bench contains 138,832 ultrasound images and 181,952 pixel-level segmentation masks. These standardized target-mask correspondences serve as the foundation for subsequent annotation.

3.2.2 Evidence-constrained Annotation

Based on the standardized segmentation data, we further construct annotations for the three tasks. Instead of allowing an MLLM to freely generate questions or reports, we adopt an Evidence-constrained Annotation strategy. For each image, the existing target-category names and their corresponding segmentation masks are first extracted and used as the factual basis available during annotation generation. The detailed annotation prompts are provided in the Appendix A.3.

We use GPT-5.5 Singh et al. (2025) to generate annotations. For each sample, the model is only allowed to construct segmentation instructions, VQA pairs, and grounded reports based on the provided target names and pixel-level evidence. For Instruction-guided Segmentation, we generate target segmentation instructions with diverse expressions while preserving their correspondence with the target masks. For example, *"Segment the liver."* is directly grounded by the liver mask.

For **Evidence-grounded VQA**, we construct questions and answers related to visible targets, spatial locations, object counts, and other information supported by the available annotations. For example, given the question *"Which abdominal structure occupies the upper-central region of the image?"*, the answer is *"Liver."*, and the corresponding liver mask is used as supporting visual evidence.

For **Evidence-grounded Report Generation**, we generate structured ultrasound descriptions based on the target categories and their pixel-level masks, focusing primarily on four types of information: presence, category, location, and size. For example, the model may generate *"A visible kidney target is present in the middle-right region, with an approximate size of 77 $\times$ 45 pixels."* The key information contained in the report is therefore directly supported by the ground-truth masks.

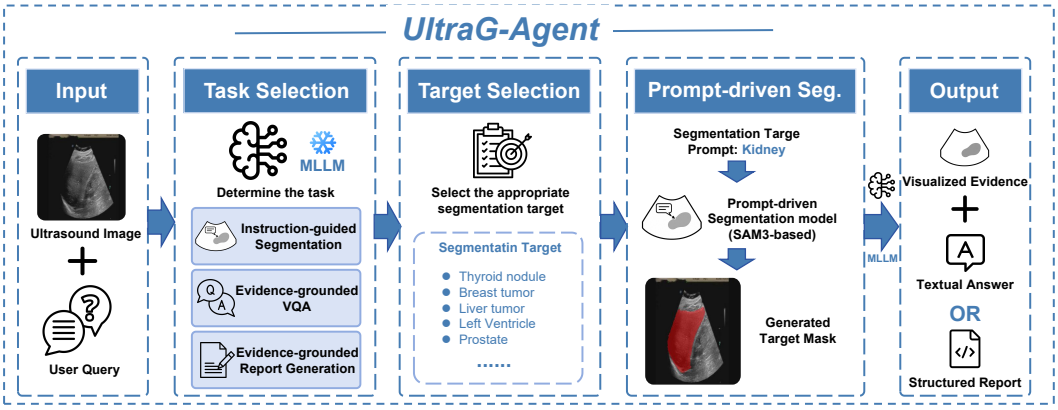


Figure 3: Overview of UltraG-Agent.

The entire annotation process follows a strict evidence constraint. The annotation model is not allowed to introduce diagnoses, pathological attributes, treatment recommendations, or other medical conclusions that cannot be supported by the available pixel-level annotations. In other words, segmentation masks are primarily used to support visual facts such as target presence, spatial location, and extent, rather than serving as evidence for additional clinical attributes. In this way, the annotations are not generated independently of the segmentation labels, but remain traceably connected to the original pixel-level annotations, ensuring that every grounded prediction in UltraG-Bench can be traced back to explicit visual evidence.

3.2.3 Dataset Validation

To ensure annotation reliability, we adopt a two-stage validation pipeline consisting of automated validation and expert review. Automated checks assess data completeness, formatting consistency, image–annotation correspondence, and task-specific alignment between textual annotations and pixel-level evidence. Samples that fail these checks are removed or regenerated.

For manual review, two reviewers with training in ultrasound imaging independently assess 10% samples from each anatomical structure, with representative coverage across categories and annotation types. They verify the consistency between annotations and their corresponding pixel-level evidence across all three tasks. Disagreements are jointly reviewed to reach consensus.

3.3 Statistics

UltraG-Bench contains **331,125** instruction-guided segmentation instructions, **666,779** evidence-grounded VQA pairs, and **138,832** evidence-grounded reports, forming a multi-level evaluation framework that ranges from target segmentation and visual question answering to report generation. Detailed statistics are provided in Appendix A.1.

4 UltraG-Agent

SAM3 Carion et al. (2026) is a prompt-driven segmentation model that supports open-vocabulary text prompts. It can perform segmentation according to text prompts, and further employs an MLLM-based agent to translate complex user instructions into executable segmentation targets. Inspired by this framework, we further propose **UltraG-Agent**, which combines the semantic understanding capability of a multimodal large language model with the fine-grained pixel-level localization capability of a prompt-driven ultrasound segmentation model to achieve more reliable Pixel-level Evidence Grounding. UltraG-Agent does not require additional end-to-end training. Instead, it consists of a pretrained MLLM without task-specific fine-tuning and UltraSAM3 Xu et al. (2026a), a prompt-driven segmentation model adapted to the ultrasound domain. The overall framework is illustrated in Figure 3.

Given an ultrasound image and a user instruction, the MLLM first performs **task selection** to determine whether the current request corresponds to Instruction-guided Segmentation, Evidence-grounded VQA, or Evidence-grounded Report Generation. It then selects the **segmentation target** required for the task based on the image, the user instruction, and a predefined set of ultrasound target categories, and converts the selected target into a concise text prompt for segmentation. Depending on the task, the MLLM can select one or multiple targets and generate a corresponding UltraSAM3 prompt for each target.

After target selection, UltraSAM3 performs **prompt-driven segmentation** using the generated text prompts and produces pixel-level masks for the selected targets. For segmentation tasks, the predicted masks are directly returned as the output. For VQA and report generation, the system further extracts visual information from the predicted masks, including target presence, category, spatial location, and extent, and uses these mask-derived facts as reliable visual evidence to constrain the final text generation. UltraG-Agent ultimately outputs the visualized pixel-level evidence together with the corresponding textual answer or structured report, thereby combining the high-level semantic understanding capability of the MLLM with the fine-grained spatial localization capability of UltraSAM3. Detailed formulations and prompting strategies are provided in the Appendix A.4.

5 Experiment

5.1 Experimental Setting

We evaluate **UltraG-Bench** on 14 representative MLLMs, including closed-source models, open-source general-purpose models, medical MLLMs, and pixel-level grounded medical MLLMs. To examine whether UltraG-Agent can extend prompt-driven segmentation models with VQA and report generation capabilities while preserving or improving their segmentation performance, we additionally evaluate BiomedParse Zhao et al. (2025) and UltraSAM3 Xu et al. (2026a) as prompt-driven segmentation baselines. We further evaluate four UltraG-Agent variants with different MLLMs and prompt-driven segmentation models. All models are evaluated under a unified protocol across the three tasks of UltraG-Bench. Detailed baseline introduction and the prompts used for evaluation are given in Appendix A.2 and Appendix A.5, respectively.

5.2 Evaluation Metric

Instruction-guided Segmentation. We use **Intersection over Union (IoU)** and **Dice coefficient (Dice)** to evaluate the overlap between the predicted masks and the ground-truth masks.

Evidence-grounded VQA. We report **Answer Accuracy (Acc.)** to evaluate the semantic correctness of the generated answers. Since a correct answer does not necessarily imply that the model relies on the correct visual evidence, we further introduce **Grounded Accuracy (G-Acc.)**. A prediction is counted as correct by G-Acc. only when both the textual answer is correct and the predicted evidence region is sufficiently consistent with the corresponding ground-truth mask. Specifically,

$$\text{G-Acc.} = \frac{1}{N}\sum_{i=1}^{N}\mathbb{I}\left(a_i = \hat{a}_i \wedge \text{IoU}(\hat{M}_i, M_i) \geq \tau\right), \tag{1}$$

where τ is the predefined grounding **threshold 0.5**.

Evidence-grounded Report Generation. We use **ROUGE-L** to measure the lexical similarity between the generated and reference reports. To further evaluate whether the generated report correctly captures the key visual evidence, we introduce **Semantic Accuracy (SemAcc)**. Specifically, we assess four semantic attributes: **presence**, which indicates whether the target is correctly identified as present or absent; **category**, which evaluates whether the target structure is correctly recognized; **location**, which measures whether the spatial position of the target is correctly described; and **size**, which evaluates whether the reported target extent is consistent with the ground-truth mask. These attributes correspond to binary indicators presence_ok, category_ok, location_ok, and size_ok. The report-level semantic accuracy is defined as

$$\text{SemAcc} = \frac{1}{4}\left(\text{presence_ok} + \text{category_ok} + \text{location_ok} + \text{size_ok}\right), \tag{2}$$

Table 2: Experimental Results on UltraG-Bench. The best, second-best, and third-best results among all models are highlighted in green, blue, and yellow, respectively. For UltraG-Agent, I, Q, and L refer to InternVL3.5-8B, Qwen3-VL-8B, and Lingshu-7B, while $US3$ and $S3$ refer to UltraSAM3 and SAM3, respectively.

Model	Instruction-guided Seg.		Evidence-grounded VQA		Evidence-grounded Report Gen.	
	IoU↑	Dice↑	Acc.↑	G-Acc.↑	ROUGE-L↑	SemAcc↑
Closed-source MLLMs						
Claude-4.5	0.1544	0.2180	0.3327	0.0214	0.1334	0.3281
GPT-5.4	0.2653	0.3718	0.3631	0.0609	0.2264	0.3577
Open-source General MLLMs						
Qwen3-VL-4B	0.1069	0.1537	0.2197	0.0137	0.0793	0.1899
MiniCPM-V 2.6	0.0642	0.0854	0.0254	0.0000	0.1554	0.2592
Qwen3-VL-8B	0.1061	0.1614	0.1938	0.0071	0.1274	0.2930
Qwen3-VL-32B	0.1136	0.1595	0.3095	0.0113	0.2211	0.3025
InternVL3.5-8B	0.0974	0.1462	0.3678	0.0081	0.1903	0.3113
InternVL3.5-14B	0.1162	0.1755	0.3528	0.0176	0.2408	0.3426
Open-source Medical MLLMs						
LLaVA-Med	0.1369	0.2006	0.2290	0.0000	0.1690	0.2822
MedGemma-4B	0.1081	0.1702	0.3207	0.0025	0.1355	0.2938
Lingshu-7B	0.0715	0.1114	0.3292	0.0000	0.1666	0.2975
Lingshu-32B	0.0920	0.1413	0.2949	0.0046	0.1940	0.3000
MedGemma-27B	0.1616	0.2424	0.3668	0.0056	0.2158	0.3436
Pixel-level Grounded Medical MLLMs						
UniBiomed	0.1550	0.2068	0.0267	0.0002	0.1217	0.0567
Prompt-driven Segmentation Models						
BiomedParse	0.2645	0.3303	- -	- -	- -	- -
UltraSAM3	0.5478	0.6190	- -	- -	- -	- -
Ours						
UltraG-Agent$_{I+S3}$	0.0432	0.0483	0.0851	0.0000	0.1751	0.1568
UltraG-Agent$_{L+US3}$	0.5292	0.6114	0.5418	0.3351	0.3805	0.6225
UltraG-Agent$_{Q+US3}$	0.5435	0.6228	0.5684	0.4222	0.3300	0.5766
UltraG-Agent$_{I+US3}$	0.5564	0.6384	0.5970	0.4531	0.3971	0.6687

and the final score is averaged over all evaluated samples. Detailed definitions and evaluation of the four attributes are provided in the Appendix A.6 and Appendix A.7, respectively.

5.3 Evaluation Results

5.3.1 Main Results

Table 2 summarizes the overall performance of different models on the three tasks of UltraG-Bench. General-purpose MLLMs show relatively strong semantic understanding but limited pixel-level grounding. For example, InternVL3.5-8B and GPT-5.4 achieve VQA accuracies of 0.3678 and 0.3631, while their G-Acc. scores remain only 0.0081 and 0.0609, respectively. Medical MLLMs exhibit a similar pattern: models such as MedGemma-27B achieve competitive VQA and report-generation performance, but all obtain G-Acc. below 0.006 and relatively low segmentation IoU. In contrast, the pixel-level grounded medical model UniBiomed shows stronger segmentation capability but substantially weaker semantic performance. Prompt-driven segmentation models exhibit much stronger localization, with BiomedParse and UltraSAM3 achieving IoU/Dice scores of 0.2645/0.3303 and 0.5478/0.6190, respectively, but they cannot perform VQA or report generation.

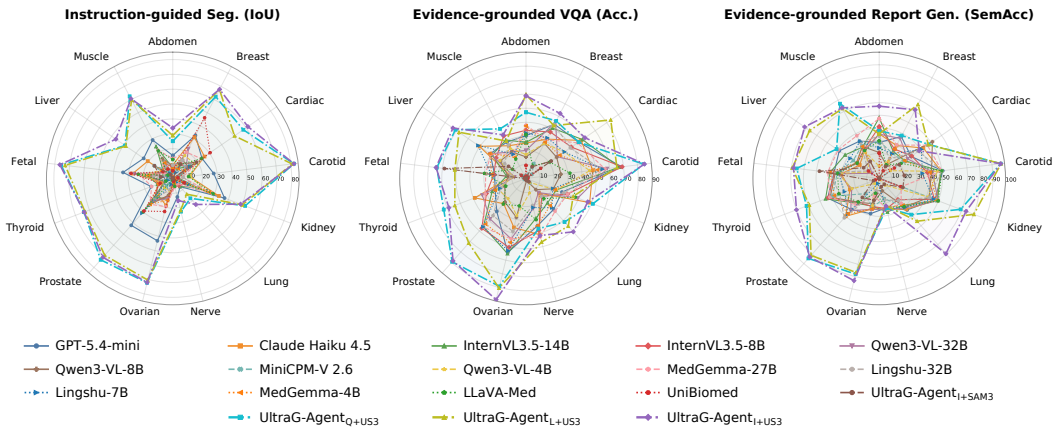


Figure 4: Fine-grained organ-level comparison across 13 anatomical categories for the three tasks in UltraG-Bench. UltraG-Agent variants with UltraSAM3 show consistently strong performance across diverse organs.

In contrast, **UltraG-Agent** achieves substantial improvements across all three tasks, demonstrating the effectiveness of combining MLLM-based semantic understanding with ultrasound-specific pixel-level localization. Among all variants, UltraG-Agent$_{I+US3}$ achieves the best performance on all six metrics, outperforming the best baseline by 0.2911, 0.2666, 0.2292, 0.3922, 0.1563, and 0.3110 in IoU, Dice, Acc., G-Acc., ROUGE-L, and SemAcc, respectively. Notably, compared with UltraSAM3 alone, UltraG-Agent$_{I+US3}$ further improves IoU/Dice from 0.5478/0.6190 to 0.5564/0.6384, while additionally enabling VQA and report generation. Further comparison shows that UltraSAM3 is a key factor behind these gains: replacing it with the original SAM3 leads to substantial performance drops across segmentation, VQA, and report generation. Meanwhile, different MLLM backbones yield only modest differences and all maintain strong performance. Overall, these results indicate that reliable Pixel-level Evidence Grounding depends on both high-level semantic reasoning and specialized ultrasound pixel-level localization, with UltraSAM3 serving as a critical foundation.

5.3.2 Fine-grained Organ-level Analysis

Figure 4 presents the fine-grained performance of different models across 13 anatomical categories. Clear performance variations can be observed across organs, indicating that the difficulty of Pixel-level Evidence Grounding is jointly affected by anatomical structure and task type. In addition, both general-purpose and medical MLLMs exhibit substantial performance fluctuations across organs, particularly on Instruction-guided Segmentation, where their pixel-level localization ability is generally limited. Detailed case study and error analysis are provided in Appendix A.8 and A.9.

In contrast, UltraG-Agent variants with UltraSAM3 achieve broader radar coverage across most anatomical categories and show relatively consistent trends across different MLLM backbones. This indicates that the advantage of UltraG-Agent is not driven by only a few specific organs, but generalizes more consistently across diverse ultrasound anatomical structures.

6 Conclusion

In this work, we construct **UltraG-Bench**, a multi-task benchmark for evaluating Pixel-level Evidence Grounding in ultrasound. By annotating 40 public ultrasound segmentation datasets, UltraG-Bench provides unified evaluation across Instruction-guided Segmentation, Evidence-grounded VQA, and Evidence-grounded Report Generation. Our experiments reveal a clear gap between semantic understanding and pixel-level localization in existing MLLMs. To bridge this gap, we further propose **UltraG-Agent**. Extensive experiments show that UltraG-Agent consistently improves both semantic prediction and visual grounding performance.

AI USE STATEMENT

We used ChatGPT only for language refinement and writing assistance. All generated contents were carefully reviewed and revised.

REPRODUCIBILITY STATEMENT

To support reproducibility, detailed dataset information and baseline descriptions are provided in Appendix A.1 and A.2. Benchmark construction details, including annotation prompts and quality validation, are described in Appendix A.3. The implementation details of UltraG-Agent and the complete evaluation protocols are provided in Appendix A.4 and A.5, respectively. Detailed definitions and attribute-wise analyses of Semantic Accuracy are given in Appendix A.6 and A.7. Additional case studies and error analyses are provided in Appendix A.8 and A.9. Our benchmark and associated code have been released anonymously at `https://anonymous.4open.science/r/UltraG-Bench-F908`.

REFERENCES

Walid Al-Dhabyani, Mohammed Gomaa, Hussien Khaled, and Aly Fahmy. Dataset of breast ultrasound images. *Data in brief*, 28:104863, 2020.

AI Anthropic. System card: Claude opus 4 & claude sonnet 4. *Claude-4 Model Card*, 2025.

Ali Abbasian Ardakani, Afshin Mohammadi, Mohammad Mirza-Aghazadeh-Attari, and U Rajendra Acharya. An open-access breast lesion ultrasound image database: Applicable in artificial intelligence studies. *Computers in Biology and Medicine*, 152:106438, 2023.

Vahid Ashkani Chenarlogh, Mostafa Ghelich Oghli, Ali Shabanzadeh, Nasim Sirjani, Ardavan Akhavan, Isaac Shiri, Hossein Arabi, Morteza Sanei Taheri, and Mohammad Kazem Tarzamni. Fast and accurate u-net model for fetal ultrasound image segmentation. *Ultrasonic imaging*, 44 (1):25–38, 2022.

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.

Zachary Baum, Shaheer Saeed, Zhe Min, Yipeng Hu, and Dean Barratt. Mr to ultrasound registration for prostate challenge-dataset. In *Medical Image Computing and Computer Assisted Intervention–MICCAI*, volume 2023, 2023.

Clyde J Belasso, Bahareh Behboodi, Habib Benali, Mathieu Boily, Hassan Rivaz, and Maryse Fortin. Luminous database: lumbar multifidus muscle segmentation from ultrasound images. *BMC Musculoskeletal Disorders*, 21(1):703, 2020.

Yuan Bi, Zhongliang Jiang, Ricarda Clarenbach, Reza Ghotbi, Angelos Karlas, and Nassir Navab. Mi-segnet: Mutual information-based us segmentation for unseen domain generalization. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 130–140. Springer, 2023.

Benedikt Boecking, Naoto Usuyama, Shruthi Bannur, Daniel C Castro, Anton Schwaighofer, Stephanie Hyland, Maria Wetscherek, Tristan Naumann, Aditya Nori, Javier Alvarez-Valle, et al. Making the most of text semantics to improve biomedical vision–language processing. In *European conference on computer vision*, pp. 1–21. Springer, 2022.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris Coll-Vinent, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, et al. Sam 3: Segment anything with concepts. In *International conference on learning representations*, volume 2026, pp. 138846–138923, 2026.

Geon Choi, Hangyul Yoon, Hyunju Shin, Hyunki Park, Sang Hoon Seo, Eunho Yang, and Edward Choi. Instruction-guided lesion segmentation for chest x-rays with automatically generated large-scale dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1482–1492, 2026.

Rebecca Cunningham and Neil Cronin. FALLMUD: FAscicle Lower Leg Muscle Ultrasound Dataset. https://kalisteo.cea.fr/index.php/fallmud/, 2020.

Karine Souza Da Correggio, R Noya Galluzzo, Luıs Otávio Santos, F Soares Muylaert Barroso, T Zimmermann Loureiro Chaves, A Sherlley Casimiro Onofre, and Aldo von Wangenheim. Fetal abdominal structures segmentation dataset using ultrasonic images. *Mendeley Data*, 1:1, 2023.

Daniel Coelho De Castro, Aurelia Bustos, Shruthi Bannur, Stephanie L Hyland, Kenza Bouzid, Maria Teodora Wetscherek, Maria Dolores Sánchez-Valverde, Lara Jaques-Pérez, Lourdes Pérez-Rodríguez, Kenji Takeda, et al. Padchest-gr: A bilingual chest x-ray dataset for grounded radiology report generation. *NEJM AI*, 2(7):AIdbp2401120, 2025.

Yi Ding, IEEE Member, Qiqi Yang, Yiqian Wang, Dajiang Chen, Zhiguang Qin, and Jian Zhang. Mallesnet: A multi-object assistance based network for brachial plexus segmentation in ultrasound images. *Medical Image Analysis*, 80:102511, 2022.

Kenneth A Fleming, Susan Horton, Michael L Wilson, Rifat Atun, Kristen DeStigter, John Flanigan, Shahin Sayed, Pierrick Adam, Bertha Aguilar, Savvas Andronikou, et al. The lancet commission on diagnostics: transforming access to diagnostics. *The Lancet*, 398(10315):1997–2050, 2021.

Wilfrido Gómez-Flores, Maria Julia Gregorio-Calas, and Wagner Coelho de Albuquerque Pereira. Bus-bra: A breast ultrasound dataset for assessing computer-aided diagnosis systems. *Medical physics*, 51(4):3110–3123, 2024.

Haifan Gong, Guanqi Chen, Ranran Wang, Xiang Xie, Mingzhi Mao, Yizhou Yu, Fei Chen, and Guanbin Li. Multi-task learning for thyroid nodule segmentation with thyroid region prior. In *2021 IEEE 18th international symposium on biomedical imaging (ISBI)*, pp. 257–261. IEEE, 2021.

Haifan Gong, Jiaxin Chen, Guanqi Chen, Haofeng Li, Guanbin Li, and Fei Chen. Thyroid region prior guided attention for ultrasound segmentation of thyroid nodules. *Computers in biology and medicine*, 155:106389, 2023.

Xiaoqing Guo, Mohammad Alsharid, He Zhao, Yipei Wang, Jayne Lander, Aris T Papageorghiou, and J Alison Noble. A visually grounded language model for fetal ultrasound understanding. *Nature Biomedical Engineering*, pp. 1–17, 2026.

Yanjun Guo, Xingguang Duan, Chengyi Wang, and Huiqin Guo. Segmentation and recognition of breast ultrasound images based on an expanded u-net. *Plos one*, 16(6):e0253202, 2021.

Xuehai He, Yichen Zhang, Luntian Mou, Eric Xing, and Pengtao Xie. Pathvqa: 30000+ questions for medical visual question answering. *arXiv preprint arXiv:2003.10286*, 2020.

Yutao Hu, Tianbin Li, Quanfeng Lu, Wenqi Shao, Junjun He, Yu Qiao, and Ping Luo. Omnimedvqa: A new large-scale comprehensive evaluation benchmark for medical lvlm. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22170–22183. IEEE, 2024.

Ahmed Iqbal and Muhammad Sharif. Memory-efficient transformer network with feature fusion for breast tumor segmentation and classification task. *Engineering Applications of Artificial Intelligence*, 127:107292, 2024.

JeevaWS. Ultrasound Normal Kidney Image Dataset. https://universe.roboflow.com/jeevaws/ultrasound-normal-kidney-image, 2025.

Bai Jieyun and Ou ZhanHong. Pubic symphysis-fetal head segmentation and angle of progression, March 2024. URL https://doi.org/10.5281/zenodo.10829116.

Kai Jin, Qixuan Sun, Daohuan Kang, Ziyao Luo, Tao Yu, Wenzheng Han, Yi Zhang, Meng Wang, Danli Shi, and Andrzej Grzybowski. Grounded report generation for enhancing ophthalmic ultrasound interpretation using vision-language segmentation models. *npj Digital Medicine*, 9(1):99, 2026.

Markus Krönke, Christine Eilers, Desislava Dimova, Melanie Köhler, Gabriel Buschner, Lilit Schweiger, Lemonia Konstantinidou, Marcus Makowski, James Nagarajah, Nassir Navab, et al. Tracked 3d ultrasound and deep neural network-based thyroid segmentation reduce interobserver variability in thyroid volumetry. *Plos one*, 17(7):e0268550, 2022.

Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. Lisa: Reasoning segmentation via large language model. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9579–9589. IEEE, 2024.

Jason J Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. A dataset of clinically generated visual questions and answers about radiology images. *Scientific data*, 5(1):180251, 2018.

Sarah Leclerc, Erik Smistad, Joao Pedrosa, Andreas Østvik, Frederic Cervenansky, Florian Espinosa, Torvald Espeland, Erik Andreas Rye Berg, Pierre-Marc Jodoin, Thomas Grenier, et al. Deep learning for segmentation using an open large-scale dataset in 2d echocardiography. *IEEE transactions on medical imaging*, 38(9):2198–2210, 2019.

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.

Bo Liu, Li-Ming Zhan, Li Xu, Lin Ma, Yan Yang, and Xiao-Ming Wu. Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. In *2021 IEEE 18th international symposium on biomedical imaging (ISBI)*, pp. 1650–1654. IEEE, 2021.

Bo Liu, Xiangyu Zhao, Along He, Yidi Chen, Huazhu Fu, and Xiao-Ming Wu. Gemex-rmcot: An enhanced med-vqa dataset for region-aware multimodal chain-of-thought reasoning. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 13213–13220, 2025a.

Bo Liu, Ke Zou, Li-Ming Zhan, Zexin Lu, Xiaoyu Dong, Yidi Chen, Chengqiang Xie, Jiannong Cao, Xiao-Ming Wu, and Huazhu Fu. Gemex: A large-scale, groundable, and explainable medical vqa benchmark for chest x-ray diagnosis. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 21310–21320. IEEE, 2025b.

Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang. Segment anything in medical images. *Nature communications*, 15(1):654, 2024.

Francesco Marzola, Nens Van Alfen, Jonne Doorduin, and Kristen M Meiburger. Deep learning segmentation of transverse musculoskeletal ultrasound images for neuromuscular disease assessment. *Computers in biology and medicine*, 135:104623, 2021.

James R McLaughlan, Lewis Howell, and Nicola Ingram. Lung ultrasound covid phantom dataset used for training machine learning model. 2024.

Peder L Myhre, Bjørnar Grenne, Federico M Asch, Victoria Delgado, Rohan Khera, Stéphane Lafitte, Roberto M Lang, Patricia A Pellikka, Partho P Sengupta, Sreekanth Vemulapalli, et al. Artificial intelligence-enhanced echocardiography in cardiovascular disease management. *Nature Reviews Cardiology*, 23(3):164–182, 2026.

José Ignacio Orlando. AbdomenUS: US Simulation and Segmentation. `https://www.kaggle.com/datasets/ignaciorlando/ussimandsegm`, 2018.

David Ouyang, Bryan He, Amirata Ghorbani, Neal Yuan, Joseph Ebinger, Curtis P Langlotz, Paul A Heidenreich, Robert A Harrington, David H Liang, Euan A Ashley, et al. Video-based ai for beat-to-beat assessment of cardiac function. *Nature*, 580(7802):252–256, 2020.

Anna Pawłowska, Anna Ćwierz-Pieńkowska, Agnieszka Domalik, Dominika Jaguś, Piotr Kasprzak, Rafał Matkowski, Łukasz Fura, Andrzej Nowicki, and Norbert Żołek. Curated benchmark dataset for ultrasound based breast lesion analysis. *Scientific Data*, 11(1):148, 2024.

Lina Pedraza, Carlos Vargas, Fabián Narváez, Oscar Durán, Emma Muñoz, and Eduardo Romero. An open access thyroid ultrasound image database. In *10th International symposium on medical information processing and analysis*, volume 9287, pp. 188–193. SPIE, 2015.

Charitha D Reddy, Leo Lopez, David Ouyang, James Y Zou, and Bryan He. Video-based deep learning for automated assessment of left ventricular ejection fraction in pediatric patients. *Journal of the American Society of Echocardiography*, 36(5):482–489, 2023.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241. Springer, 2015.

M Sofia Sappia, Chris L de Korte, Bram van Ginneken, Dean Ninalga, Satoshi Kondo, Satoshi Kasai, Kousuke Hirasawa, Tanya Akumu, Carlos Martín-Isla, Karim Lekadir, et al. Acouslic-ai challenge report: Fetal abdominal circumference measurement on blind-sweep ultrasound data from low-income countries. *Medical image analysis*, 105:103640, 2025.

Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, et al. Medgemma technical report. *arXiv preprint arXiv:2507.05201*, 2025.

W Shao and W Brisbane. Micro-ultrasound prostate segmentation dataset. *URL: https://doi.org/10.5281/zenodo*, 10475293, 2024.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, et al. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*, 2025.

Rohit Singla, Cailin Ringstrom, Grace Hu, Victoria Lessoway, John Reid, Christopher Nguan, and Robert Rohling. The open kidney ultrasound data set. In *International Workshop on Advances in Simplifying Medical Ultrasound*, pp. 155–164. Springer Nature Switzerland, 2023.

Wu Songxiong, Zhang Hongyuan, Ye Tingting, X Haoyu, Z Ping, S Qingjun, W Panying, H Bingsheng, D Lei, and W Guangyao. Focus: four-chamber ultrasound image dataset for fetal cardiac biometric measurement. *Zenodo10*, 5281, 2025.

Stanford AIMI. Thyroid ultrasound cine-clip, April 2026. URL `https://stanford.redivis.com/datasets/2n5z-4m99zdpyp?v=1.0`.

Unity Imaging Collaborative. Unity Imaging Echocardiography Datasets. `https://data.unityimaging.net/`, 2024.

Noelia Vallez, Gloria Bueno, Oscar Deniz, Miguel Angel Rienda, and Carlos Pastor. Bus-uclm: Breast ultrasound lesion segmentation dataset. *Scientific Data*, 12(1):242, 2025.

Thomas LA van den Heuvel, Dagmar de Bruijn, Chris L de Korte, and Bram van Ginneken. Automated measurement of fetal head circumference using 2d ultrasound images. *PloS one*, 13(8): e0200412, 2018.

Tianchen Wang, Zhihe Li, Meiping Huang, Jian Zhuang, Shanshan Bi, Jiawei Zhang, Yiyu Shi, Hongwen Fei, and Xiaowei Xu. Echocp: An echocardiography dataset in contrast transthoracic echocardiography for patent foramen ovale diagnosis. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 506–515. Springer, 2021.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025.

Yuchen Wang, Zhongyu Li, Xiangxiang Cui, Liangliang Zhang, Xiang Luo, Meng Yang, and Shi Chang. Key-frame guided network for thyroid nodule recognition using ultrasound videos. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 238–247. Springer, 2022.

Joy T Wu, Nkechinyere N Agu, Ismini Lourentzou, Arjun Sharma, Joseph A Paguio, Jasper S Yao, Edward C Dee, William Mitchell, Satyananda Kashyap, Andrea Giovannini, et al. Chest imagenome dataset for clinical reasoning. *arXiv preprint arXiv:2108.00316*, 2021.

Linshan Wu, Yuxiang Nie, Sunan He, Jiaxin Zhuang, Luyang Luo, Tao Li, Zhuoyao Xie, Dexuan Chen, Yinghua Zhao, Neeraj Mahboobani, et al. A universal foundation model for grounded biomedical image interpretation. *Nature Communications*, 2026.

Bo Xu, Quanhao Zhu, Rui Lin, Boling Zhu, Chenyuan Wang, Hongfei Lin, Feng Xia, and Chenhua Ji. Ultrasam3: A concept-driven foundation model for universal ultrasound image segmentation. *arXiv preprint arXiv:2607.29200*, 2026a.

Weiwen Xu, Hou Pong Chan, Long Li, Mahani Aljunied, Ruifeng Yuan, Jianyu Wang, Chenghao Xiao, Guizhen Chen, Chaoqun Liu, Zhaodonghui Li, et al. Lingshu: Generalist foundation model for unified multimodal medical understanding and reasoning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026b.

Jiewen Yang, Xinpeng Ding, Ziyang Zheng, Xiaowei Xu, and Xiaomeng Li. Graphecho: Graph-driven unsupervised domain adaptation for echocardiogram video segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 11878–11887, 2023.

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*, 2024.

Moi Hoon Yap, Gerard Pons, Joan Marti, Sergi Ganau, Melcior Sentis, Reyer Zwiggelaar, Adrian K Davison, and Robert Marti. Automated breast ultrasound lesions detection using convolutional neural networks. *IEEE journal of biomedical and health informatics*, 22(4):1218–1226, 2017.

Xu Yiming, Zheng Bowen, Liu Xiaohong, Wu Tao, Ju Jinxiu, Wang Shijie, Lian Yufan, Zhang Hongjun, Liang Tong, Sang Ye, Jiang Rui, Wang Guangyu, Ren Jie, and Chen Ting. Annotated ultrasound liver images, November 2022. URL `https://doi.org/10.5281/zenodo.7272660`.

Qi Zhao, Shuchang Lyu, Wenpei Bai, Linghan Cai, Binghao Liu, Guangliang Cheng, Meijing Wu, Xiubo Sang, Min Yang, and Lijiang Chen. Mmotu: A multi-modality ovarian tumor ultrasound image dataset for unsupervised cross-domain semantic segmentation. *arXiv preprint arXiv:2207.06799*, 2022.

Theodore Zhao, Yu Gu, Jianwei Yang, Naoto Usuyama, Ho Hin Lee, Sid Kiblawi, Tristan Naumann, Jianfeng Gao, Angela Crabtree, Jacob Abel, et al. A foundation model for joint segmentation, detection and recognition of biomedical objects across nine modalities. *Nature methods*, 22(1): 166–176, 2025.

A Appendix

A.1 Datasets Details

Source Dataset Statistics. Table 3 summarizes the 40 ultrasound segmentation datasets spanning 13 organs that serve as sources for UltraG-Bench. The selected source data comprises 138,832 images and 181,952 segmentation masks across the training and test splits.

Table 3: Statistics of the 40 ultrasound datasets retained in the Grounded-US annotations.

Organ	Dataset	Train Images	Train Masks	Test Images	Test Masks
Abdomen	AbdomenUS Orlando (2018)	784	3,079	198	781
Breast	BrEast Pawłowska et al. (2024)	204	402	52	102
Breast	BUID Ardakani et al. (2023)	185	370	47	94
Breast	BUSI Al-Dhabyani et al. (2020)	624	1,056	156	274
Breast	BUS_BRA Gómez-Flores et al. (2024)	1,500	3,000	375	750
Breast	BUS_DatasetB Yap et al. (2017)	130	260	33	66
Breast	BUS_UC Iqbal & Sharif (2024)	648	1,294	163	326
Breast	BUS_UCLM Vallez et al. (2025)	546	408	137	120
Breast	S1 Guo et al. (2021)	192	388	9	18
Cardiac	CAMUS Leclerc et al. (2019)	1,600	4,800	400	1,200
Cardiac	CardiacUDC Yang et al. (2023)	1,717	5,892	533	1,754
Cardiac	EchoCP Wang et al. (2021)	434	1,696	132	521
Cardiac	EchoNet_Dynamic Ouyang et al. (2020)	17,476	17,476	2,550	2,550
Cardiac	EchoNet_Pediatric Reddy et al. (2023)	12,071	12,071	3,300	3,300
Cardiac	Unity Unity Imaging Collaborative (2024)	13,136	5,675	3,283	1,367
Carotid artery	CCA Bi et al. (2023)	2,307	2,195	540	540
Fetal	ACOUSLIC Sappia et al. (2025)	5,327	5,327	1,293	1,293
Fetal	FASS Da Correggio et al. (2023)	1,270	10,012	318	2,506
Fetal	Fast_UNet Ashkani Chenarlogh et al. (2022)	1,128	1,128	283	283
Fetal	HC van den Heuvel et al. (2018)	799	799	200	200
Fetal	fh_ps Jieyun & ZhanHong (2024)	3,200	6,400	800	1,600
Fetal	focus Songxiong et al. (2025)	250	500	50	100
Kidney	KidneyUS Singla et al. (2023)	388	388	98	98
Kidney	Ultrasound_Normal_Kidney JeevaWS (2025)	864	1,728	216	432
Liver	Annotated_Ultrasound_Liver Yiming et al. (2022)	588	510	147	125
Lung	LUSS McLaughlan et al. (2024)	451	2,977	113	775
Muscle	FALLMUD Cunningham & Cronin (2020)	650	655	163	326
Muscle	LUMINOUS Belasso et al. (2020)	269	302	68	78
Muscle	STMUS_NDA Marzola et al. (2021)	6,529	6,530	1,632	1,632
Nerve	UPBD Ding et al. (2022)	764	3,441	191	935
Ovarian	OTU_2d Zhao et al. (2022)	1,175	1,191	294	298
Ovarian	OTU_3d Zhao et al. (2022)	136	146	34	40
Prostate	MicroSeg Shao & Brisbane (2024)	2,078	2,085	542	542
Prostate	RegPro Baum et al. (2023)	3,516	3,786	367	396
Thyroid	DDTI Pedraza et al. (2015)	509	510	128	128
Thyroid	KFGNet Wang et al. (2022)	131	161	33	37
Thyroid	Segthy Krönke et al. (2022)	9,938	19,187	1,952	3,482
Thyroid	TG3K Gong et al. (2023)	2,868	3,056	717	772
Thyroid	TN3K Gong et al. (2021)	2,794	3,063	699	757
Thyroid	Thyroid_US_Cineclip Stanford AIMI (2026)	13,565	13,565	3,845	3,845
Total	–	**112,741**	**147,509**	**26,091**	**34,443**

Task-wise Anatomical Distribution. Figure 5 shows the anatomical-category distributions of the three tasks in UltraG-Bench. Overall, the three tasks exhibit similar distributions across the 13 anatomical categories. Cardiac and thyroid samples constitute the largest proportions, accounting for 38%–41% and 26%–27% of the annotations, respectively, followed by fetal, muscle, and prostate categories. The remaining anatomical categories occupy smaller proportions. Despite differences in the total number of annotations across tasks, their anatomical compositions remain broadly consistent, providing comparable category coverage for evaluating pixel-level evidence grounding from segmentation to VQA and report generation.

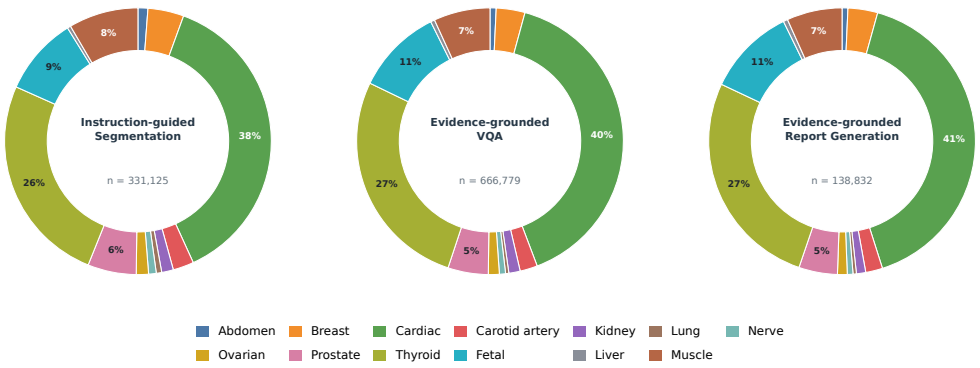


Figure 5: Anatomical-category distributions of the three tasks in UltraG-Bench across 13 anatomical categories.

A.2 Baseline Introduction

We evaluate a diverse set of baseline models covering closed-source general-purpose MLLMs, open-source general-purpose MLLMs, medical MLLMs, pixel-level grounded medical MLLMs, and prompt-driven segmentation models. The evaluated baselines are briefly introduced below.

Claude-4.5. Claude-4.5 Anthropic (2025) is a proprietary multimodal large language model developed by Anthropic, with strong general-purpose capabilities in visual understanding and multimodal reasoning. We include it as a representative closed-source MLLM for evaluating semantic understanding and visual grounding in ultrasound images.

GPT-5.4. GPT-5.4 Singh et al. (2025) is a proprietary frontier model developed by OpenAI with advanced multimodal understanding and reasoning capabilities. We evaluate GPT-5.4 as a strong closed-source general-purpose baseline across all three tasks of UltraG-Bench.

Qwen3-VL Series. Qwen3-VL Bai et al. (2025) is an open-source vision-language model family designed for general multimodal perception and reasoning, supporting diverse visual inputs and fine-grained visual understanding. We evaluate the 4B, 8B, and 32B variants to investigate the effect of model scale on ultrasound Pixel-level Evidence Grounding.

MiniCPM-V 2.6. MiniCPM-V 2.6 Yao et al. (2024) is a compact 8B multimodal model designed for efficient image, multi-image, and video understanding. Despite its relatively small model size, it provides strong general-purpose visual-language capabilities and serves as an efficient open-source MLLM baseline.

InternVL3.5 Series. InternVL3.5Wang et al. (2025) is an open-source multimodal model family that improves visual-language understanding, reasoning, and inference efficiency through enhanced multimodal training and reinforcement learning. We evaluate its 8B and 14B variants as representative general-purpose MLLMs with strong multimodal reasoning capabilities.

LLaVA-Med. LLaVA-Med Li et al. (2023) adapts a general-purpose vision-language assistant to the biomedical domain using large-scale biomedical image-text data and instruction tuning. It is designed for open-ended conversational understanding of biomedical images and serves as a representative early medical MLLM.

MedGemma Series. MedGemma Sellergren et al. (2025) is a family of medical vision-language foundation models built upon Gemma and specifically adapted for medical image and text understanding. We evaluate the 4B and 27B variants to assess whether medical-domain adaptation and increased model capacity improve ultrasound evidence grounding.

Lingshu Series. Lingshu Xu et al. (2026b) is a generalist medical MLLM designed for unified multimodal medical understanding and reasoning through large-scale medical data curation and multi-stage training. We evaluate the 7B and 32B variants, which support a broad range of medical imaging modalities and tasks including medical VQA and report generation.

UniBiomed. UniBiomed Wu et al. (2026) is a grounded biomedical vision-language foundation model that jointly performs semantic interpretation and pixel-level segmentation of biomedical targets. Unlike conventional medical MLLMs that primarily generate textual responses, UniBiomed explicitly connects generated medical findings with localized visual regions, making it particularly relevant to Pixel-level Evidence Grounding.

BiomedParse. BiomedParse Zhao et al. (2025) is a biomedical image parsing foundation model that unifies segmentation, detection, and recognition across diverse biomedical imaging modalities. It supports text-prompted segmentation of biomedical objects and is therefore included as a representative prompt-driven biomedical segmentation model.

UltraSAM3. UltraSAM3 Xu et al. (2026a) is a concept-driven foundation model specifically developed for universal ultrasound image segmentation. By adapting SAM3 to ultrasound-specific image–mask–concept triplets, it enables text-based target specification and provides strong prompt-driven pixel-level localization across diverse ultrasound anatomical structures.

A.3 Benchmark Construction Details

Annotation Details. For each ultrasound image, we first extract evidence facts from the original COCO-format segmentation annotations, including target presence, category, mask identity, coarse image location, pixel extent, mask area ratio, boundary information, and the number of distinct targets. These mask-derived facts serve as the only visual evidence available during annotation. Based on them, we construct three types of annotations: instruction-guided segmentation, evidence-grounded VQA, and evidence-grounded report generation. The annotation process follows strict evidence constraints: generated text is restricted to information directly supported by the supplied labels and mask geometry, while unsupported diagnostic, pathological, treatment-related, or metadata-related statements are prohibited. After generation, all annotations are automatically normalized and validated to ensure valid evidence references and task-format consistency. Segmentation instructions and grounded reports are deterministically constructed from mask-derived facts, while VQA pairs are further filtered according to clinical relevance and evidence validity; invalid outputs are replaced by rule-based fallback annotations. The same annotation pipeline is applied to all 13 anatomical categories by replacing the organ-specific target with {`organ`}.

Prompt

Create a grounded multi-task ultrasound sample for {`organ`}. Return valid JSON only.
Use only the supplied facts. The {`organ`} label identifies the segmentation target. Do not introduce any clinical attribute that is not explicitly supported by the supplied labels or mask geometry.
Do not invent diagnoses, pathological findings, prognosis, treatment, recommendations, or other unsupported clinical conclusions.
Never ask about or mention file names, case numbers, slice numbers, image IDs, dataset/source names, split names, paths, dates, annotations, mask IDs, or metadata in user-facing text.
Every VQA question must be directly supported by the available {`organ`} evidence, such as:

- target presence or visibility;
- image-region location;
- pixel extent or size;
- area ratio;
- boundary availability;
- number of visible targets.

Use `[SEG]` only when `target_mask_ids` is non-empty. Mask IDs may appear only in ID fields.
The supplied masks support target presence, approximate image location, pixel-level size and extent, target count, and boundary information.

For questions about a visible {organ} target, set requires_grounding=true and evidence_source=mask_geometry.
If no target mask is available, generate only absence or not-applicable questions with requires_grounding=false.
Return exactly the following top-level keys:
segmentation_tasks: 2–4 objects with task_id, instruction, answer, target_mask_ids, and answerability.
grounded_vqa: 4–6 objects with qa_id, question, answer, answer_type, requires_grounding, evidence_mask_ids, reasoning_type, and evidence_source.
grounded_report: an object with a report prompt and 5–7 sentence objects containing section, text, evidence_mask_ids, and finding_type.
Facts:
{mask-derived evidence facts}

Automated Quality Validation. We perform rule-based automatic quality validation for every generated sample before manual review. The validator checks five aspects. First, masks must have unique identifiers, valid anatomical categories, positive geometry and area, valid area ratios, coordinates within image boundaries, and no duplicated geometry. Second, each sample must contain 2–4 instruction-guided segmentation annotations, where every referenced target mask exists and the [SEG] token is used if and only if a valid mask is referenced. Third, each sample must contain 4–6 evidence-grounded VQA pairs with non-empty question–answer content, valid evidence masks, and consistent requires_grounding flags. Fourth, grounded reports must contain 5–7 sentences with at least *Findings* and *Impression*, and all statements about target presence, category, location, extent, and count must agree with the referenced masks. Finally, we reject unsupported content beyond the available pixel-level evidence, including unannotated anatomical attributes, unsupported laterality or physical measurements, diagnostic or pathological claims, treatment recommendations, and metadata-related statements. Organ-specific constraints are further defined for each {organ}. Issues are categorized as critical, major, minor, or informational, with penalties of 25, 10, 3, and 0 points from an initial score of 100; samples containing any critical or major issue are flagged as failing automatic validation.

A.4 UltraG-Agent Implementation Details

UltraG-Agent follows a modular inference pipeline that separates high-level instruction understanding from pixel-level visual grounding. Given an ultrasound image I, a user instruction q, and a predefined target-category set $\mathcal{C}$, an MLLM first acts as a task planner to identify the task type and the anatomical targets required to answer the instruction:

$$\mathcal{P} = \mathcal{M}_{\text{plan}}(I, q, \mathcal{C}) = \left(\tau, \{(c_k, p_k)\}_{k=1}^{K}\right), \tag{3}$$

where $\tau \in \{\text{Segmentation}, \text{VQA}, \text{Report}\}$ denotes the inferred task, $c_k \in \mathcal{C}$ is the selected target category, and p_k is a concise text prompt generated for UltraSAM3. Multiple targets are selected only when required by the instruction.

For each planned target, UltraSAM3 independently performs prompt-driven segmentation:

$$\{(m_{k,j}, s_{k,j})\}_{j=1}^{N_k} = \mathcal{S}_{\text{US3}}(I, p_k), \tag{4}$$

where $m_{k,j}$ and $s_{k,j}$ denote the j-th candidate mask and its confidence score, respectively. After filtering invalid candidates, the highest-scoring mask is selected as

$$\hat{m}_k = \arg\max_{m_{k,j}} s_{k,j}. \tag{5}$$

We then extract mask-derived evidence

$$g_k = \mathcal{G}(\hat{m}_k) = \{b_k,\, l_k,\, a_k,\, r_k,\, d_k^{\text{long}},\, d_k^{\text{short}}\}, \tag{6}$$

including the bounding box b_k, image-region location l_k, pixel area a_k, area ratio r_k, and long- and short-axis pixel extents. These quantities form the authoritative visual evidence used by the subsequent response generation stage.

The final response is conditioned on the original instruction, the task plan, and the mask-derived evidence:

$$y = \mathcal{M}_{\text{resp}}\left(I, q, \mathcal{P}, \{(c_k, g_k)\}_{k=1}^{K}\right). \tag{7}$$

For instruction-guided segmentation, the output consists of the textual response and the corresponding predicted mask. For evidence-grounded VQA, the answer is paired with the union of the masks referenced by the response:

$$M_{\text{VQA}} = \bigcup_{k \in \mathcal{T}_a} \hat{m}_k, \tag{8}$$

where $\mathcal{T}_a$ denotes the targets supporting the answer. For evidence-grounded report generation, each report statement y_i is associated with its own evidence mask,

$$M_i = \bigcup_{k \in \mathcal{T}_i} \hat{m}_k, \tag{9}$$

thereby preserving fine-grained statement–evidence correspondence throughout the generated report. Mask-derived facts are treated as authoritative during response generation, preventing the MLLM from modifying target presence, location, pixel extent, or area measurements inferred from UltraSAM3.

A.5 Evaluation Details

Prompt-driven Segmentation Models Evaluation Details. For prompt-driven segmentation models, the ultrasound image and its segmentation instruction are directly provided as input. The predicted mask is evaluated against the ground-truth segmentation using IoU and Dice, following the same protocol as Instruction-guided Segmentation.

UltraG-Agent Evaluation Details. The same evaluation samples, reference annotations, mask decoding procedure, and metric implementation are used for UltraG-Agent to ensure direct comparability with the MLLM baselines.

MLLM Evaluation Details. We evaluate all MLLMs under a unified zero-shot protocol across the three tasks of UltraG-Bench. For each sample, the model receives the original ultrasound image together with a task-specific textual prompt and is required to return a structured JSON response. For spatial predictions, models may output either a polygon or a bounding box in the original-image pixel coordinate system; these predictions are rasterized into binary masks and compared with the corresponding ground-truth segmentation masks. All API-based models are evaluated with temperature 0 to reduce sampling variance, and the same task prompts and output constraints are used across different models. No ground-truth mask coordinates or reference answers are provided to the model during inference. For report generation, only the required finding-type structure is specified to standardize the output format, while the reference report content and evidence masks remain hidden.

For Instruction-guided Segmentation, we report IoU and Dice between the predicted and ground-truth masks. For Evidence-grounded VQA, Answer Accuracy measures semantic correctness, while Grounded Accuracy further requires the answer to be correct and the predicted evidence region to achieve an IoU above 0.5 with the ground-truth evidence mask. For Evidence-grounded Report Generation, ROUGE-L evaluates textual similarity, while Semantic Accuracy measures the correctness of four evidence-supported attributes: target presence, category, location, and size. Location is evaluated using the predefined 3×3 image-region grid, and predicted pixel extent is considered correct when both dimensions are within a 20% tolerance of the reference values.

Evaluation Prompts. The following prompts are used consistently for all evaluated MLLMs. Here, {`organ`}, {`W`}, and {`H`} denote the anatomical category and the original image width and height, respectively.

Instruction-guided Segmentation Prompt
You are given a clinical ultrasound image and a segmentation instruction.
Return valid JSON only with keys: `answer`, `predicted_mask`.

`predicted_mask` must be {`"type":"polygon","points":[[x,y],...]`}, {`"type":"bbox","bbox_xywh":[x,y,w,h]`}, or `null`.
Use original-image pixel coordinates. Include `[SEG]` in `answer` if a target is present.
Do not use annotation, annotated, dataset, mask id, or metadata.
Input:

```
{
  "image_width": {W},
  "image_height": {H},
  "instruction": "{instruction}"
}
```

Evidence-grounded VQA Prompt
You are given a clinical ultrasound image and a question.
Return valid JSON only with keys: `answer`, `predicted_mask`.
If the answer depends on a visible target, organ, or structure, provide `predicted_mask` as a polygon or bbox in original-image pixel coordinates; otherwise use `null`.
Do not use annotation, annotated, dataset, mask id, or metadata.
Input:

```
{
  "organ": "{organ}",
  "image_width": {W},
  "image_height": {H},
  "question": "{question}"
}
```

Evidence-grounded Report Generation Prompt
You are given a clinical ultrasound image.
Return valid JSON only with key `report`.
`report` must be a list of sentence objects with `section`, `text`, `finding_type`, and `predicted_mask`.
For spatial `finding_type` values, `predicted_mask` must localize the supporting target, organ, or structure as a polygon or bbox in original-image pixel coordinates.
Use `null` predicted mask for non-spatial statements. Use the provided finding-type structure when applicable.
Use concise benchmark-style wording. Do not invent ultrasound descriptors if uncertain. Do not invent diagnostic grades, pathology classes, or treatment recommendations, and do not use cm unless a scale marker is visible. If estimating size, use pixel coordinates or pixel extent only. Do not use annotation, annotated, dataset, mask id, or metadata.
Input:

```
{
  "organ": "{organ}",
  "image_width": {W},
  "image_height": {H},
  "prompt": "{report_instruction}"
}
```

A.6 Detailed Definition of Semantic Accuracy

To complement lexical similarity metrics, we evaluate whether a generated report preserves the key visual evidence represented by the reference annotation. Specifically, Semantic Accuracy (**SemAcc**) consists of four binary attributes: target presence, anatomical category, spatial location, and target size. Let R and $\hat{R}$ denote the reference and generated reports, respectively. Each attribute produces a binary indicator in $\{0, 1\}$, as defined below.

Target Presence. The presence attribute evaluates whether the generated report correctly describes the target as present or absent. We extract a binary target-presence state from both the reference and generated reports:

$$p(R),\, p(\hat{R}) \in \{0, 1\}, \tag{10}$$

where 1 denotes that the target is present and 0 denotes that the target is absent. Presence correctness is defined as

$$\text{presence_ok} = \mathbb{I}\left[p(\hat{R}) = p(R)\right], \tag{11}$$

where $\mathbb{I}[\cdot]$ is the indicator function. In practice, explicit absence expressions (e.g., "no target", "target is absent", or "target is not visible") are interpreted as negative evidence, whereas explicit mentions of the target structure indicate target presence.

Anatomical Category. The category attribute measures whether the generated report correctly identifies the anatomical target specified by the reference annotation. Let $\mathcal{C}(R)$ and $\mathcal{C}(\hat{R})$ denote the sets of target categories mentioned in the reference and generated reports, respectively. We define

$$\text{category_ok} = \mathbb{I}\left[\mathcal{C}(R) \subseteq \mathcal{C}(\hat{R})\right]. \tag{12}$$

Thus, all target categories required by the reference report must be correctly identified in the generated report.

Spatial Location. We represent target location using a 3×3 spatial grid over the original image. Given an image of width W and height H, and the center of the reference target mask (x_c, y_c), its horizontal position is defined as

$$h(x_c) = \begin{cases} \text{left}, & x_c < W/3, \\ \text{central}, & W/3 \le x_c < 2W/3, \\ \text{right}, & x_c \ge 2W/3, \end{cases} \tag{13}$$

and its vertical position is

$$v(y_c) = \begin{cases} \text{upper}, & y_c < H/3, \\ \text{middle}, & H/3 \le y_c < 2H/3, \\ \text{lower}, & y_c \ge 2H/3. \end{cases} \tag{14}$$

The resulting location label is

$$\ell = (v(y_c), h(x_c)), \tag{15}$$

corresponding to one of nine regions such as *upper-left*, *middle-central*, or *lower-right*. Let $\mathcal{L}(R)$ and $\mathcal{L}(\hat{R})$ denote the spatial-location labels extracted from the reference and generated reports. Location correctness is defined as

$$\text{location_ok} = \mathbb{I}\left[\mathcal{L}(\hat{R}) = \mathcal{L}(R)\right]. \tag{16}$$

Target Size. The size attribute evaluates whether the pixel-level extent reported by the model is consistent with that derived from the ground-truth mask. Let (d_1, d_2) denote the two reference pixel extents and $(\hat{d}_1, \hat{d}_2)$ denote the corresponding values extracted from the generated report. Since the ordering of the long and short axes may vary in natural-language generation, we consider both direct and reversed matching. For each reference dimension d, the allowable tolerance is

$$\delta(d) = \max(1,\ 0.2d), \tag{17}$$

i.e., either one pixel or 20% of the reference dimension, whichever is larger. We define

$$\text{Match}\big((\hat{d}_1, \hat{d}_2), (d_1, d_2)\big) = \bigwedge_{i=1}^{2} \left(|\hat{d}_i - d_i| \le \delta(d_i)\right). \tag{18}$$

The final size indicator is

$$\begin{aligned} \text{size_ok} = \mathbb{I}\Big[&\text{Match}\big((\hat{d}_1, \hat{d}_2), (d_1, d_2)\big) \\ &\vee \text{Match}\big((\hat{d}_1, \hat{d}_2), (d_2, d_1)\big)\Big]. \end{aligned} \tag{19}$$

Semantic Accuracy. The four indicators are equally weighted to obtain the report-level semantic accuracy:

$$\mathrm{SemAcc} = \frac{1}{4}\left(\mathrm{presence_ok} + \mathrm{category_ok} + \mathrm{location_ok} + \mathrm{size_ok}\right). \tag{20}$$

The final SemAcc reported for a model is the average over all evaluated report samples:

$$\mathrm{SemAcc}_{\mathrm{final}} = \frac{1}{N}\sum_{n=1}^{N} \mathrm{SemAcc}^{(n)}, \tag{21}$$

where N denotes the number of evaluated reports.

A.7 Attribute-wise Analysis of Semantic Accuracy

Table 4: Detailed comparison of the four semantic attributes used in Semantic Accuracy on UltraG-Bench. The best, second-best, and third-best results among all models are highlighted in green, blue, and yellow, respectively. For UltraG-Agent, I, Q, and L refer to InternVL3.5-8B, Qwen3-VL-8B, and Lingshu-7B, while $US3$ and $S3$ refer to UltraSAM3 and SAM3, respectively.

Model	Presence↑	Category↑	Location↑	Size↑
Closed-source MLLMs				
Claude-4.5	0.8789	0.3871	0.0067	0.0001
GPT-5.4	0.8139	0.4824	0.0305	0.0418
Open-source General-purpose MLLMs				
Qwen3-VL-4B	0.5231	0.2328	0.0001	0.0065
MiniCPM-V 2.6	0.7070	0.2978	0.0002	0.0020
Qwen3-VL-8B	0.8173	0.2987	0.0005	0.0000
Qwen3-VL-32B	0.8275	0.3223	0.0181	0.0007
InternVL3.5-8B	0.8508	0.3678	0.0042	0.0025
InternVL3.5-14B	0.9315	0.4092	0.0090	0.0113
Medical MLLMs				
LLaVA-Med	0.8590	0.2416	0.0000	0.0002
MedGemma-4B	0.8932	0.3484	0.0016	0.0003
Lingshu-7B	0.7470	0.4121	0.0000	0.0000
Lingshu-32B	0.6789	0.4836	0.0000	0.0054
MedGemma-27B	0.8647	0.4418	0.0007	0.0037
Pixel-level Grounded Medical MLLMs				
UniBiomed	0.2472	0.0646	0.0000	0.0003
UltraG-Agent (Ours)				
UltraG-Agent$_{I+S3}$	0.1418	0.4068	0.0382	0.0000
UltraG-Agent$_{L+US3}$	0.9614	0.5853	0.5802	0.4335
UltraG-Agent$_{Q+US3}$	0.9500	0.5024	0.5992	0.3997
UltraG-Agent$_{I+US3}$	0.9502	0.7254	0.6462	0.4606

Table 4 provides a detailed breakdown of Semantic Accuracy into its four constituent attributes: presence, category, location, and size. Existing MLLMs generally perform well on target presence, indicating that they can often recognize whether the relevant anatomical structure is visible in an ultrasound image. However, their performance decreases substantially for more fine-grained attributes. In particular, most general-purpose and medical MLLMs achieve very low accuracy on location and

size, suggesting that semantic recognition does not necessarily translate into precise spatial understanding. Category recognition is comparatively stronger, but still remains substantially below the best UltraG-Agent variants.

UltraG-Agent shows clear improvements in the attributes that require pixel-level visual evidence. UltraG-Agent$_{L+US3}$ achieves the highest presence accuracy of 0.9614, while UltraG-Agent$_{I+US3}$ obtains the best results for category, location, and size, reaching 0.7254, 0.6462, and 0.4606, respectively. The gains are particularly pronounced for location and size, where conventional MLLMs achieve only limited performance. Moreover, replacing UltraSAM3 with the original SAM3 leads to substantial degradation across these attributes, further demonstrating the importance of ultrasound-specific pixel-level localization. Overall, the attribute-level results show that the improvement in SemAcc primarily comes from stronger alignment between semantic predictions and fine-grained spatial evidence, rather than from target recognition alone.

A.8 Case Study

Instruction-guided Segmentation Case Study. Figure 6 presents a representative breast lesion segmentation example. UltraG-Agent produces pixel-level masks that closely follow the ground-truth lesion boundary, with UltraG-Agent$_{I+US3}$ showing particularly accurate spatial alignment and boundary delineation. UltraSAM3 and BiomedParse also achieve reasonable localization, reflecting the advantage of dedicated prompt-driven segmentation models for fine-grained spatial prediction. In contrast, most standalone MLLMs struggle to produce precise pixel-level outputs, often returning coarse bounding boxes, incomplete contours, or regions that only partially overlap with the target. This case illustrates the substantial gap between recognizing the requested anatomical target and accurately delineating its spatial extent.

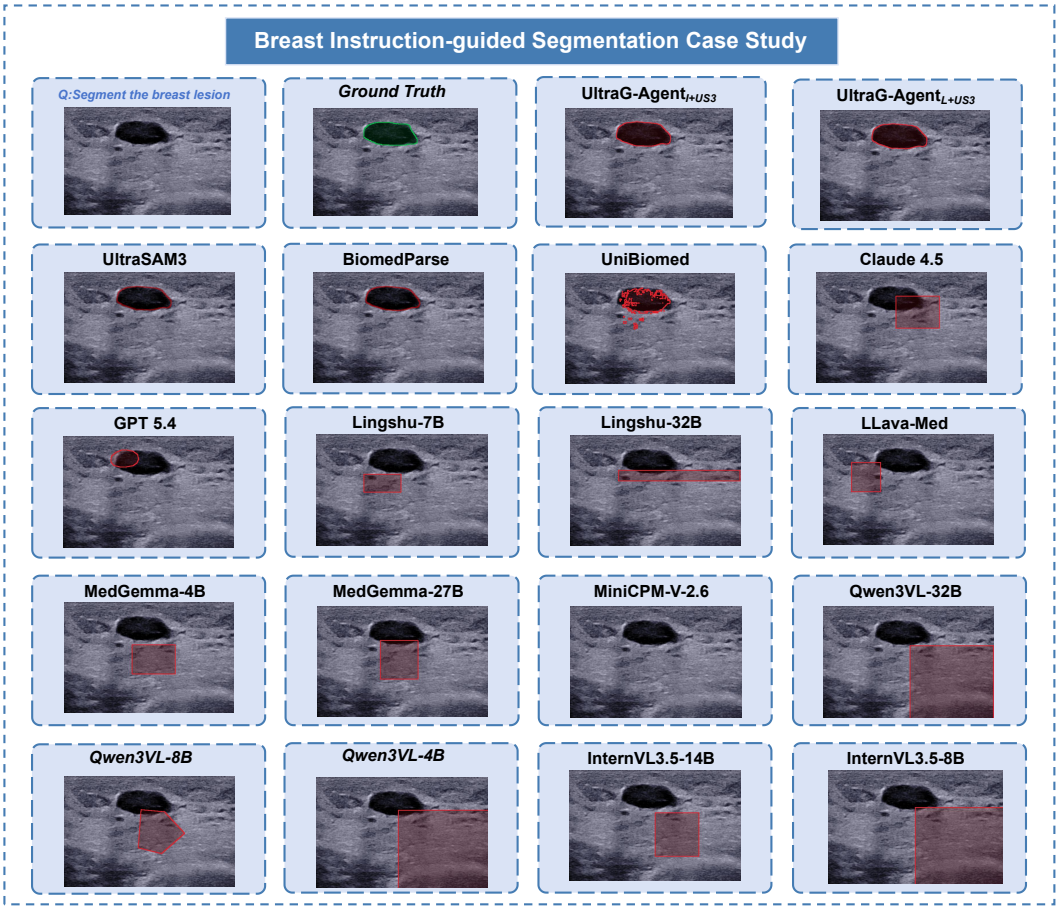


Figure 6: Breast Instruction-guided Segmentation Case Study.



Figure 7: Fetal Instruction-guided Segmentation Case Study.

Figure 7 provides another qualitative comparison under the same instruction-guided segmentation setting. A similar trend can be observed: UltraG-Agent generates a more complete and spatially consistent target mask, while many standalone MLLMs provide only approximate localization or fail to capture the target boundary accurately. Although prompt-driven segmentation models generally exhibit stronger localization ability than conventional MLLMs, UltraG-Agent retains this fine-grained segmentation capability while additionally supporting the semantic understanding required by the other tasks in UltraG-Bench. Together, the two cases demonstrate the complementary benefit of combining MLLM-based instruction understanding with ultrasound-specific pixel-level segmentation.

Evidence-grounded VQA Case Study. Figure 8 presents a representative thyroid VQA example. UltraG-Agent correctly identifies that a single thyroid nodule is present and provides a well-aligned segmentation mask as pixel-level evidence. In contrast, several compared MLLMs either return an answer without valid grounding or generate incorrect and unsupported descriptions. This case demonstrates that a semantically plausible answer alone is insufficient without spatial evidence aligned with the target.

Figure 9 shows a representative ovarian VQA example. UltraG-Agent correctly localizes the ovarian tumor target to the middle-central image region and grounds the answer with the corresponding segmentation mask. By comparison, the other models produce inaccurate spatial descriptions, unsupported anatomical references, or responses without valid pixel-level grounding. This case highlights the importance of explicitly aligning semantic answers with fine-grained visual evidence.

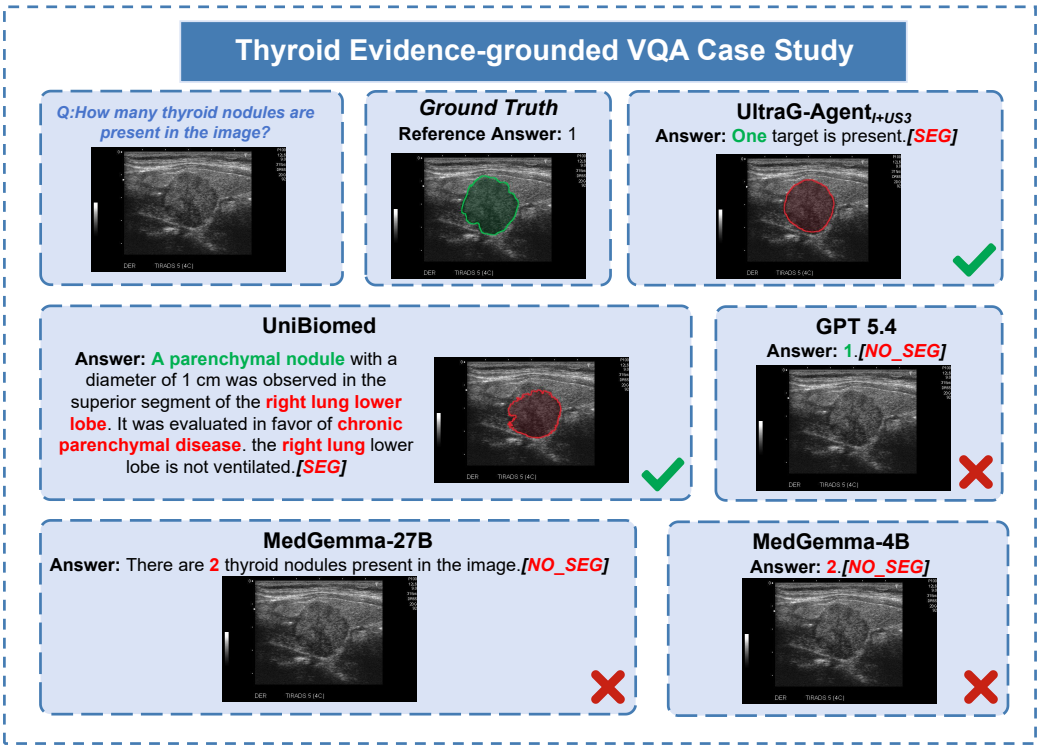


Figure 8: Thyroid Evidence-grounded VQA Case Study.

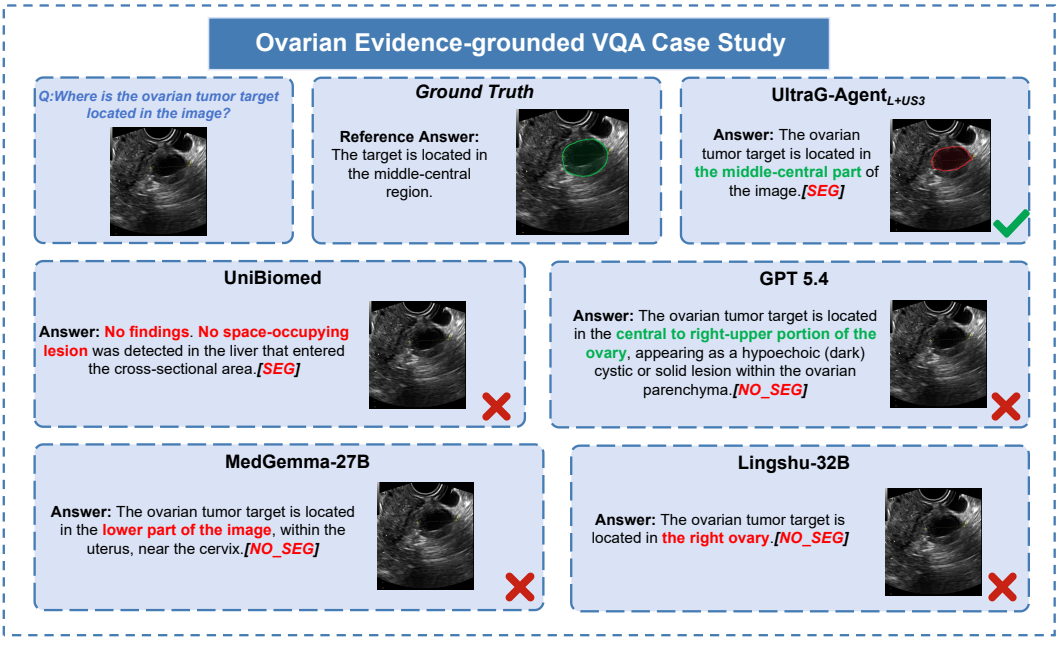


Figure 9: Ovarian Evidence-grounded VQA Case Study.

Evidence-grounded Report Generation Case Study. Figure 10 presents a representative prostate ultrasound report generation example. UltraG-Agent produces a structured report that closely follows the reference in target presence, middle-central location, pixel extent, and visible-area proportion, while providing a corresponding segmentation mask. In contrast, UniBiomed generates largely

nonspecific statements despite producing a plausible mask, Qwen3-VL-32B introduces substantial errors in location and size estimation, and Lingshu-32B further hallucinates unsupported clinical attributes and physical measurements. This example shows that accurate report generation requires not only recognizing the target, but also grounding each spatial description in reliable pixel-level evidence.

Figure 11 shows a liver ultrasound report generation example. UltraG-Agent correctly identifies the liver lesion, localizes it to the middle-central image region, and reports its pixel-level extent and image-area proportion with a mask that closely matches the reference. By comparison, UniBiomed produces semantically incorrect content referring to lung findings, while Claude-4.5 and GPT-5.4 fail to identify the focal liver lesion and instead generate unsupported or contradictory descriptions. The case highlights the benefit of evidence-constrained generation, where report content is explicitly tied to the localized ultrasound target rather than generated from semantic cues alone.

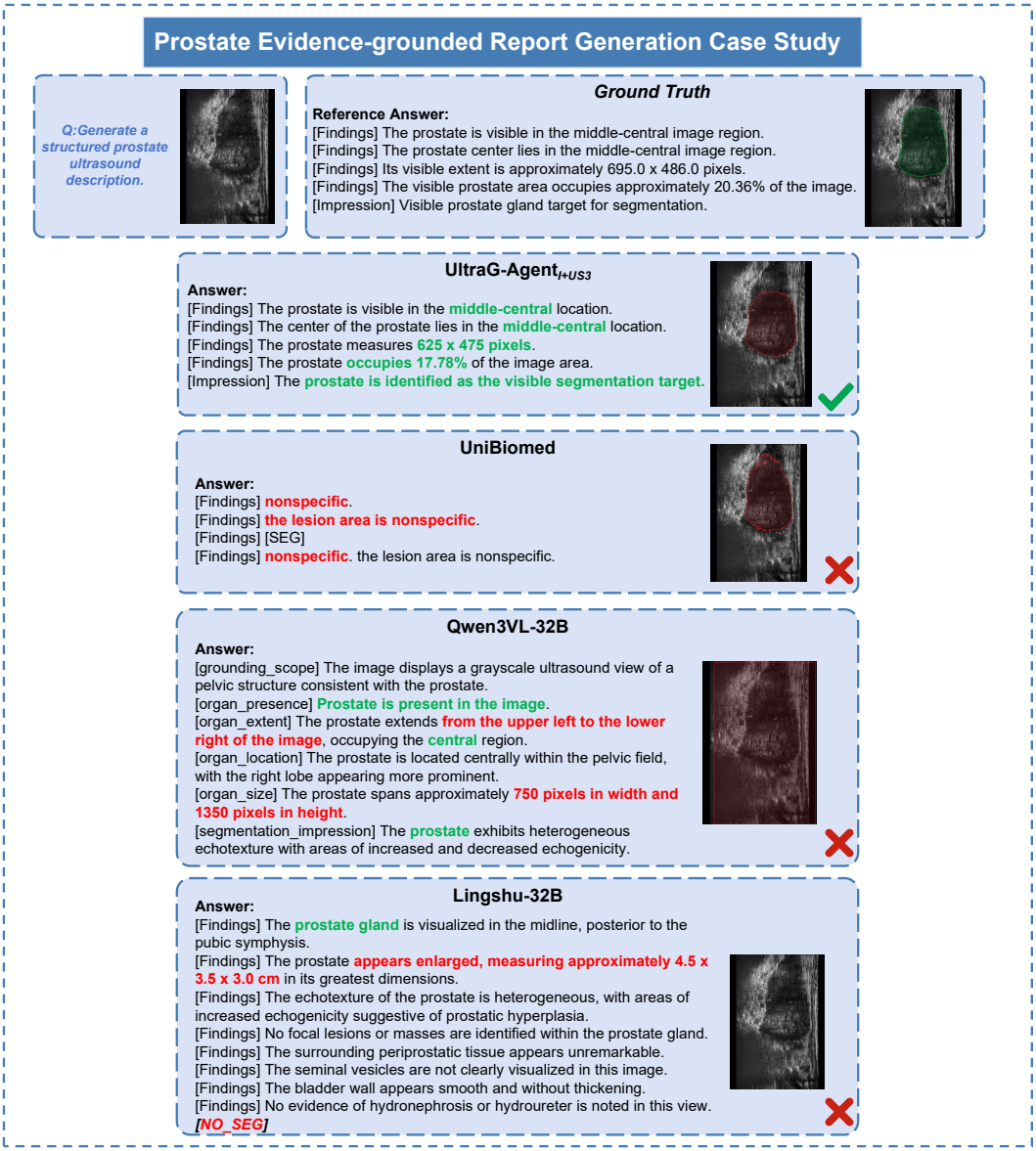


Figure 10: Prostate Evidence-grounded Report Generation Case Study.

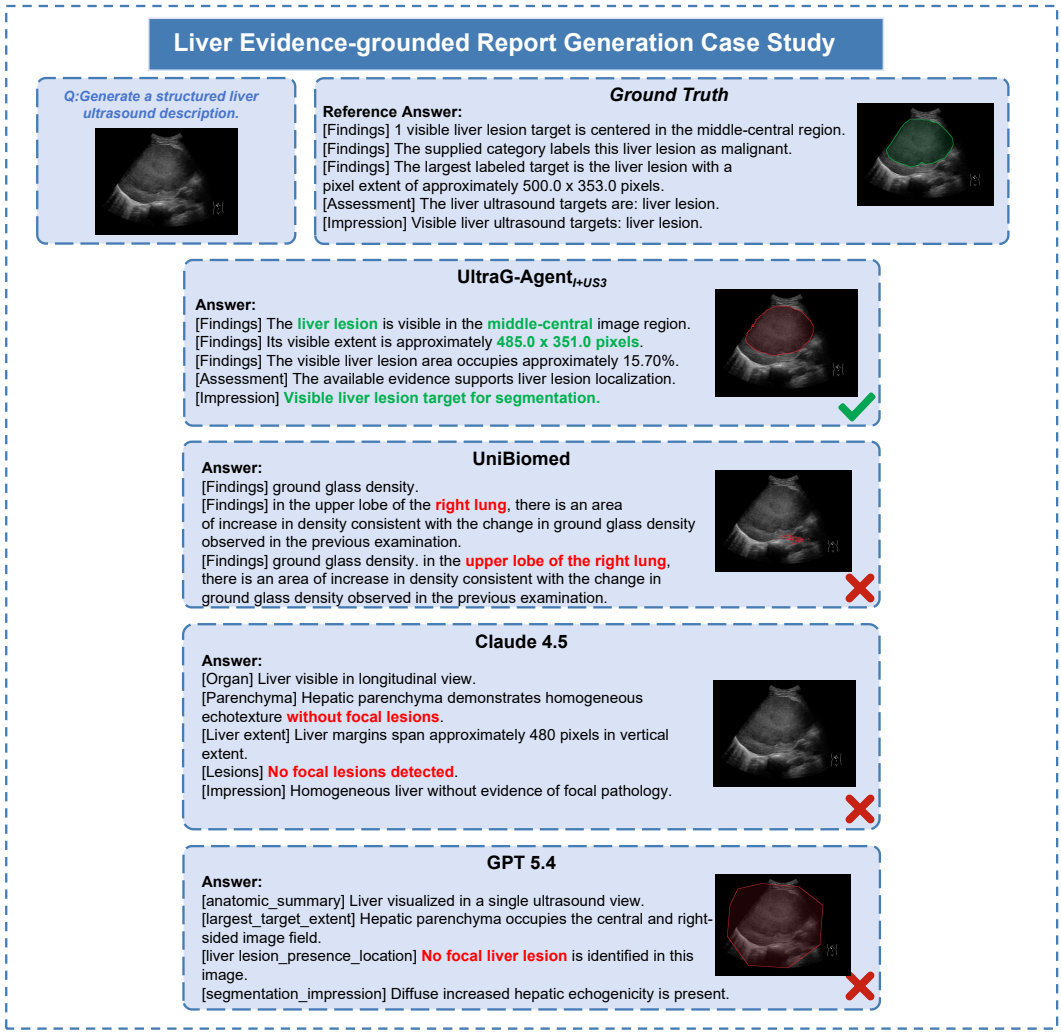


Figure 11: Liver Evidence-grounded Report Generation Case Study.

A.9 Error Analysis

Instruction-guided Segmentation Error Analysis. Figure 12 illustrates a challenging abdominal ultrasound case in which most models fail to accurately delineate the visible liver region. UltraG-Agent and UltraSAM3 localize only a small region near the upper-central part of the image, resulting in a substantial mismatch with the ground-truth mask, while BiomedParse and UniBiomed produce incomplete or fragmented predictions. Standalone MLLMs exhibit more severe failure modes, including coarse bounding boxes, substantial over-segmentation, spatially shifted regions, and empty predictions. This case highlights the difficulty of segmenting targets with weakly defined boundaries and irregular visible extent in ultrasound. It also reveals a limitation of UltraG-Agent: when the underlying prompt-driven segmentation model misidentifies the target region, the localization error is propagated to the final grounded output, indicating that further improvements in ultrasound-specific segmentation remain important.

Evidence-grounded VQA Error Analysis. Figure 13 presents a challenging lung ultrasound VQA case that exposes different failure modes in semantic prediction and visual grounding. UltraG-Agent$_{Q+US3}$ correctly identifies the A-line location as upper-central and provides a corresponding grounded region, while UltraG-Agent$_{L+US3}$ produces an incorrect middle-central localization despite generating a plausible mask. InternVL3.5-14B predicts the correct textual location but does not provide valid pixel-level evidence, illustrating that semantic correctness alone is insufficient for

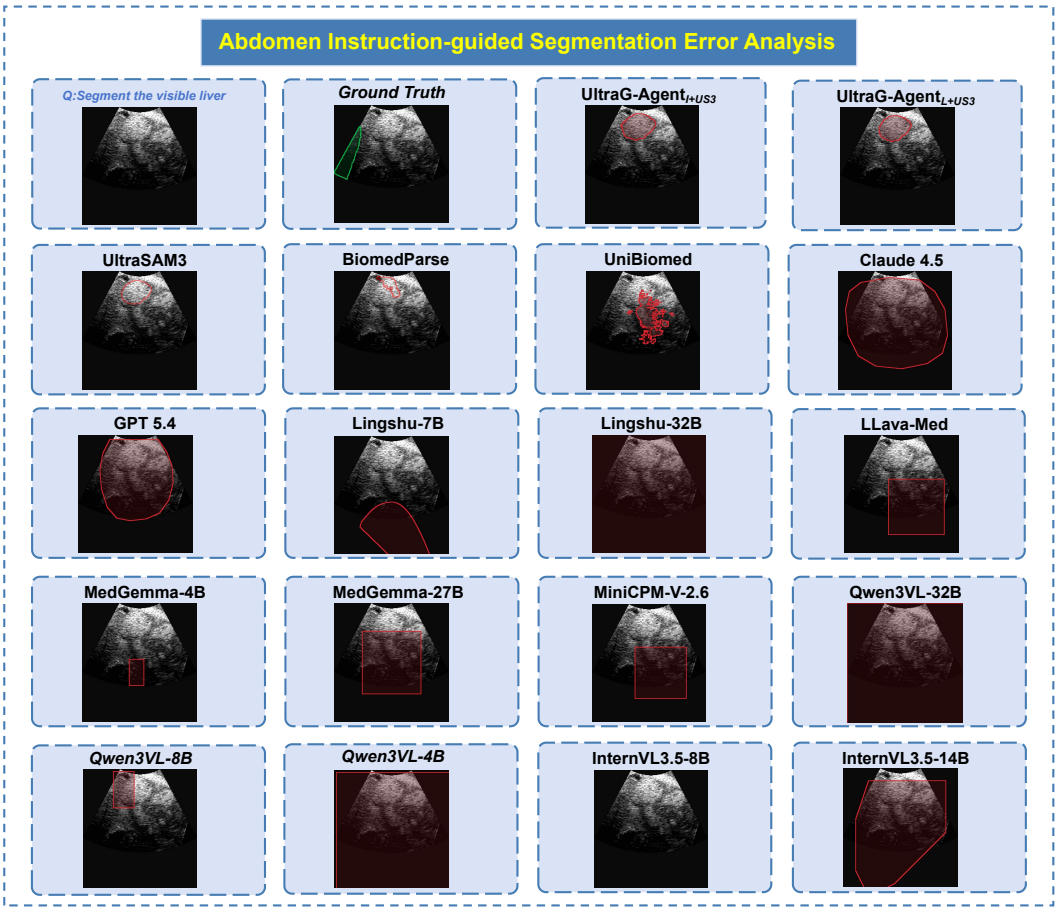


Figure 12: Abdomen Instruction-guided Segmentation Error Analysis.

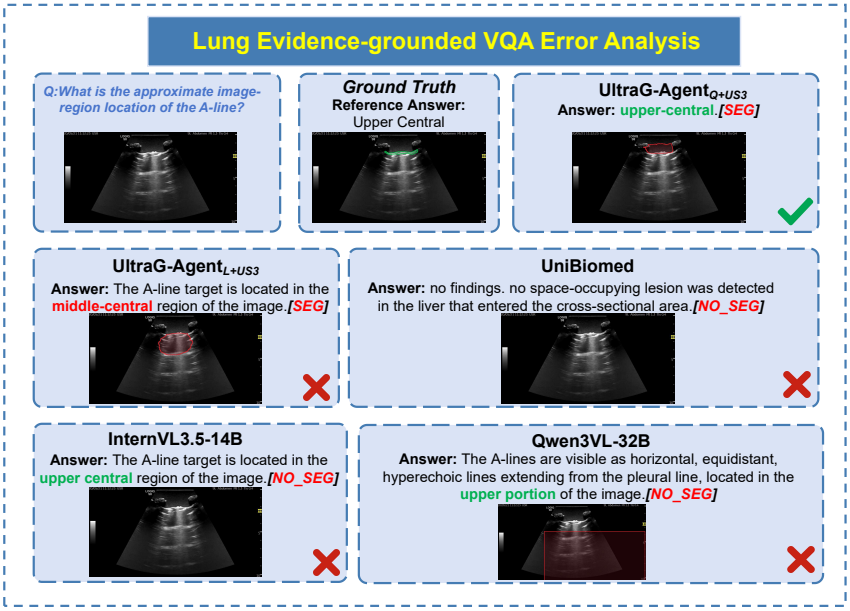


Figure 13: Lung Evidence-grounded VQA Error Analysis.

grounded VQA. UniBiomed fails to identify the relevant target, whereas Qwen3VL-32B generates an over-detailed description with inaccurate localization and an overly broad evidence region. This example highlights that evidence-grounded VQA requires both correct semantic reasoning and spatially consistent localization, and errors in either component lead to failure under the joint evaluation criterion.

Evidence-grounded Report Generation Error Analysis. Figure 14 presents a challenging cardiac ultrasound case involving multiple anatomical structures. The reference report requires grounded descriptions of the left ventricular endocardium, left ventricular epicardium, and left atrium, whereas UltraG-Agent identifies only a single left-ventricular target and further assigns an incorrect upper-central location. This illustrates a limitation of the current agent pipeline when an instruction requires simultaneous planning and grounding of multiple structures. UniBiomed fails to produce relevant cardiac findings, while MedGemma-27B recognizes several cardiac chambers but introduces unsupported structures and provides an imprecise evidence region. GPT-5.4 generates clinically plausible descriptions of cardiac anatomy, but adds unsupported attributes such as normal morphology and valve function without corresponding pixel-level evidence. Overall, this case shows that evidence-grounded report generation remains particularly challenging when multiple anatomical targets must be jointly identified, localized, and described, and that errors in target planning can propagate to the final grounded report.

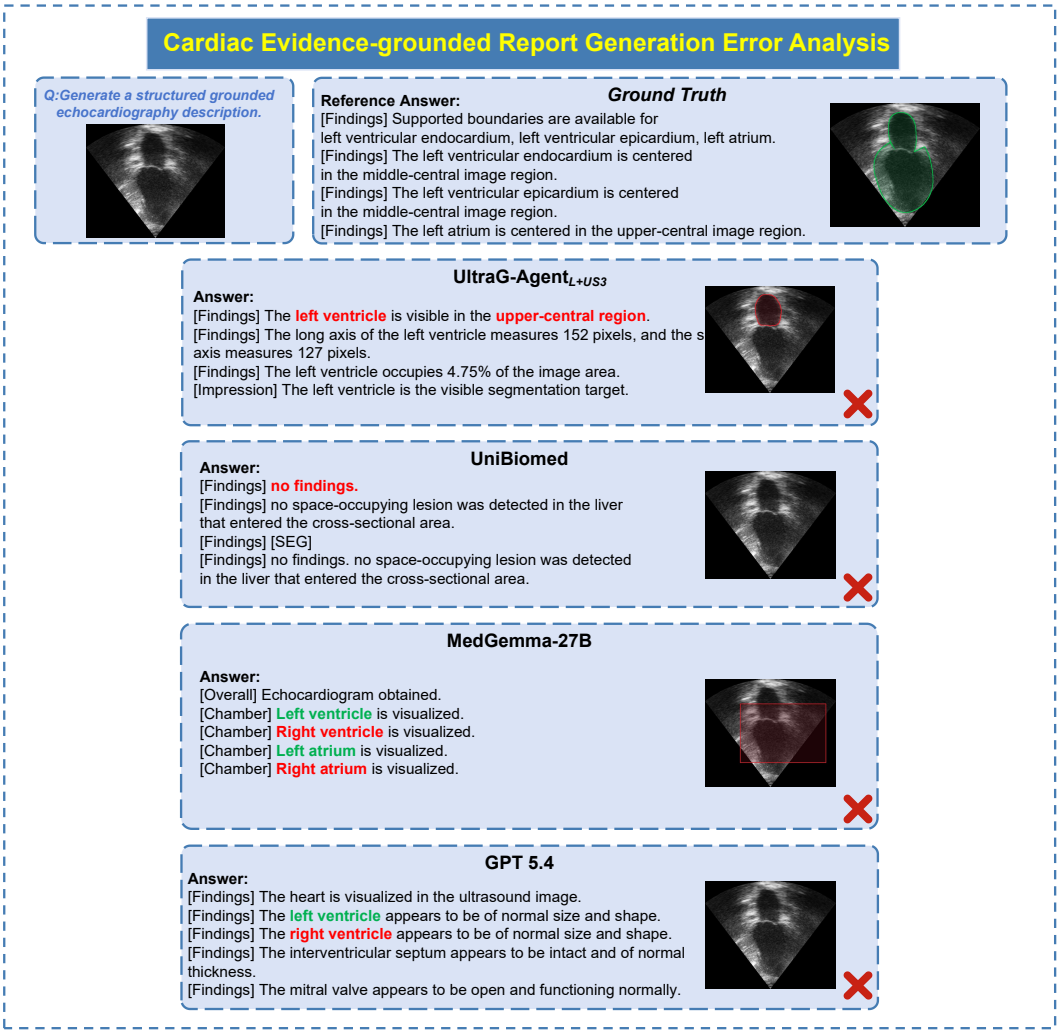


Figure 14: Cardiac Evidence-grounded Report Generation Error Analysis.