

# ReXSonoVQA: A Video QA Benchmark for Procedure-Centric Ultrasound Understanding

Xucheng Wang, BS<sup>1</sup>  
 Xiaoman Zhang, PhD<sup>1</sup>  
 Sung Eun Kim, MD<sup>1</sup>  
 Ankit Pal, BS<sup>1</sup>  
 Pranav Rajpurkar, PhD<sup>1</sup>

DAVIDX\_WANG@HMS.HARVARD.EDU  
 XIAOMAN\_ZHANG@HMS.HARVARD.EDU  
 SUNGEUN\_KIM2@HMS.HARVARD.EDU  
 ANKIT\_PAL@FAS.HARVARD.EDU

<sup>1</sup>*Department of Biomedical Informatics, Harvard Medical School, Boston, MA*

## Abstract

Ultrasound acquisition requires skilled probe manipulation and real-time adjustments. Vision-language models (VLMs) could enable autonomous ultrasound systems, but existing benchmarks evaluate only static images, not dynamic procedural understanding. We introduce **ReXSonoVQA**, a video QA benchmark with **514** video clips and **514** questions (249 MCQ, 265 free-response) targeting three competencies: Action-Goal Reasoning, Artifact Resolution & Optimization, and Procedure Context & Planning. Zero-shot evaluation of Gemini 3 Pro, Qwen3.5-397B, LLaVA-Video-72B, and Seed 2.0 Pro shows VLMs can extract some procedural information, but troubleshooting questions remain challenging with minimal gains over text-only baselines, exposing limitations in causal reasoning. ReXSonoVQA enables developing perception systems for ultrasound training, guidance, and robotic automation.

**Data and Code Availability** The dataset comprises publicly available ultrasound demonstration videos sourced from YouTube. No video files are redistributed; we provide only YouTube URLs with timestamps and regions of interest, following established practices in video benchmark curation (e.g., HowTo100M (Miech et al., 2019), AVOS (Goodman et al., 2024)). All video frames shown in figures throughout this paper correspond to entries in the dataset; source URLs are provided in the repository. The ReXSonoVQA’s dataset and code are publicly available at <https://github.com/rajpurkarlab/RexSonoVQA>.

**Institutional Review Board (IRB)** IRB was not required for this study as the data consist exclusively of publicly available demonstration videos sourced from YouTube.

## 1. Introduction

Ultrasound is widely used because it is portable, safe, and can provide real time imaging, but its acquisition is operator-dependent: obtaining clinically usable views requires skilled, adaptive control of probe position, orientation, pressure, and device settings (Sharma et al., 2021). This dependence limits both the scalability of ultrasound services and the consistency of scan quality, motivating growing interest in intelligent ultrasound assistance: from real-time guidance systems that support novice operators to fully autonomous robotic scanning (Jiang et al., 2025; Chen et al., 2025; Jiang et al., 2023; Guo et al., 2025). For such systems to succeed, their perception modules must understand not just *what* is visible in static frames, but *how* to obtain and refine it: how probe maneuvers map to imaging goals, how operators troubleshoot artifacts, and how protocols progress.

Vision Language Models (VLMs) have emerged as a promising foundation for such perception systems (Fung et al., 2025; Salimpour et al., 2025) and are increasingly evaluated on medical imaging tasks (Hu et al., 2024; Chen et al., 2024; Maani et al., 2025). However, existing evaluation has focused almost exclusively on static interpretation spanning from identifying anatomy, recognizing views to detecting pathology from individual images (Le et al., 2025). This leaves a critical gap: automation systems require temporal and causal reasoning about actions,

troubleshooting, and protocol execution that static benchmarks cannot assess (Munir et al., 2025).

We introduce **ReXSonoVQA**, the first video-based benchmark for procedure-centric ultrasound understanding, comprising 514 video clips paired with **249** multiple choice (MCQ) and **265** free-response questions spanning six clinical categories. ReXSonoVQA evaluates dynamic procedural reasoning via three cognitive tasks: (1) *Action-Goal Reasoning*, (2) *Artifact Resolution & Optimization*, and (3) *Procedure Context & Planning*. (Fig. 1) Our benchmark’s construction pipeline leverages instructional videos with timestamped transcripts, applying quality control of QA including blind(text-only) solvability screening and distractor refinement to ensure video is required for answering the question. (Fig. 2)

While existing VLMs excel at static frame interpretation, dynamic procedural understanding requires frontier video-language models. We evaluate Gemini 3 Pro, Qwen3.5-397B, LLaVA-Video-72B, and Seed 2.0 Pro, all supporting native video input. Other leading commercial VLMs (GPT-5.2 (OpenAI), Claude Opus 4.5 (Anthropic)) lack native video support and showed poor temporal reasoning with image sequences in exploratory tests. Evaluation under paired conditions (with video vs. text-only) reveals that even with frontier video-language models, substantial gaps remain: Type 2 questions prove most challenging, exposing critical limitations in causal troubleshooting reasoning essential for real-world ultrasound automation.

## 2. Related Work

**Vision-Language Models in Medical Imaging.** Medical VLMs have advanced medical image interpretation by integrating visual and textual understanding into unified frameworks (Li et al., 2023; Moor et al., 2023; Nath et al., 2025; Zhang et al., 2024; Wu et al., 2023). Recent ultrasound-specific models, including FetalCLIP (Maani et al., 2025) and LLAUS (Guo et al., 2025), have demonstrated strong performance in aligning ultrasound features with clinical text and performing static anatomy identification. However, existing medical VLMs focus primarily on static, single-frame interpretation, such as identifying anatomy, recognizing views, or detecting pathology from individual images, rather than reasoning about dynamic procedural content. Moreover, most leading commercial VLMs including GPT-5.2 (OpenAI) and Claude Opus 4.5 (Anthropic) do

not support native video input. Exploratory evaluation using sequential image inputs showed these models failed to perform temporal reasoning effectively (Fig. A6, Appendix B), motivating our focus on models with native video understanding (Gemini 3 Pro, Qwen3.5-397B, LLaVA-Video-72B, and Seed 2.0 Pro) and extended context windows necessary for analyzing procedural ultrasound sequences.

**Ultrasound Automation and Robotics.** Ultrasound’s operator dependence has driven research toward robotic standardization (Munir et al., 2025). Recent work has shifted from rule-based visual servoing toward learning-based autonomy: UltraBot employs imitation learning for autonomous carotid scanning (Jiang et al., 2025), USPilot uses LLM-enhanced graph planning to translate natural language into robotic actions (Chen et al., 2025), and Sonomate provides real-time scanning guidance through visually grounded language models (Guo et al., 2026). Despite this progress, a fundamental challenge persists: current perception systems lack the deep procedural understanding required for adaptive, real-world deployment. Successful automation demands more than anatomy recognition but requires models reason about probe manipulation strategies, diagnose and resolve image quality issues in real-time, adapt protocols to patient-specific constraints, and make decisions about scan progression. These capabilities remain largely unevaluated in existing benchmarks.

**Medical AI Benchmarks.** Current medical AI benchmarks fall into two main categories: identification/classification and reasoning/generation tasks. Text-centric benchmarks like MedMCQA evaluate professional text-based medical knowledge (Pal et al., 2022), while visual benchmarks such as OmniMedVQA (Hu et al., 2024), ReXVQA (Pal et al., 2025), GMAI-MMBench (Chen et al., 2024), and VQA-RAD (Lau et al., 2018) test fine-grained recognition across imaging modalities and anatomical regions. For deeper clinical understanding, benchmarks like ReXrank assess radiology report generation (Zhang et al., 2025), and U2-Bench evaluates ultrasound image comprehension (Le et al., 2025). However, as shown in Table 1, almost all existing benchmarks focus on static image understanding, identifying what is visible in single frames, rather than dynamic reasoning about the acquisition process itself. This limitation is particularly problematic for ultrasound automation, where perception systems must understand not just anatomical content but the

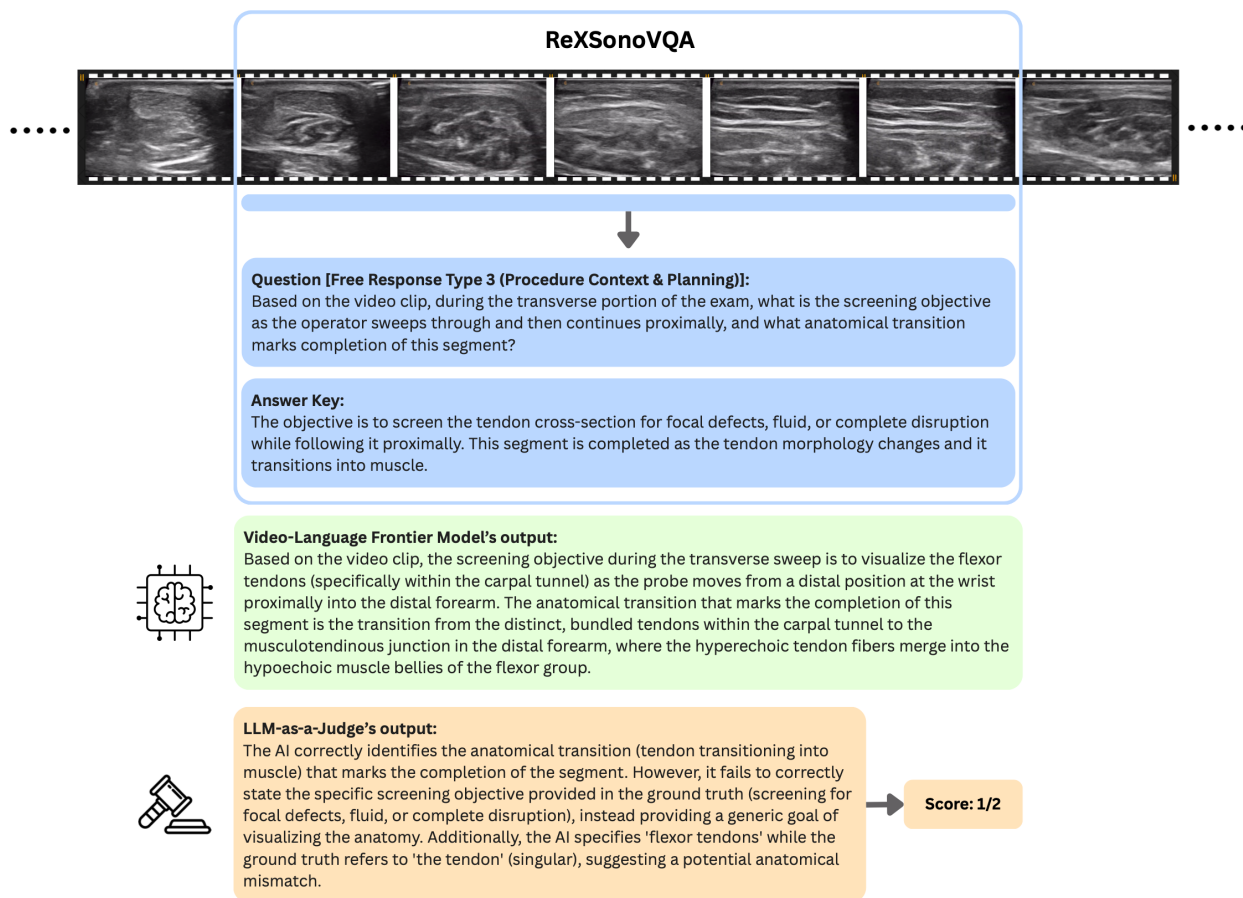


Figure 1: A ReXSonoVQA example: Type 3 (Procedure Context & Planning, Free-Response) question requiring identification of the screening objective and anatomical transition during a transverse sweep. Gemini 3 Pro correctly identifies the anatomical transition (tendon to muscle) but fails to specify the correct screening objective, receiving a partial score (1/2). More MCQs and free-responses examples see Appendix B.

real-time procedural logic of obtaining and optimizing views. ReXSonoVQA addresses this gap by evaluating procedure-centric reasoning of video sequences. We assess how models understand scanning maneuvers, troubleshooting decisions, protocol progression, and the causal relationships between actions and image outcomes.

### 3. Method

We present a pipeline for constructing a procedure-centric ultrasound video question-answering benchmark that evaluates vision-language models' ability to reason about scan execution: probe maneuvers,

acquisition goals, troubleshooting adjustments, and protocol progression. As summarized in Fig. 2, the pipeline comprises four stages: (i) task definition, (ii) data curation and ground truth construction, (iii) prompt refinement and quality control loop, and (iv) benchmark construction.

#### 3.1. Task Definition

Ultrasound acquisition follows a recurring *scan-adjust-progress* loop: operators execute maneuvers toward imaging goals, troubleshoot when quality degrades, and advance through protocol steps. We define three task types that mirror this loop and cap-

Table 1: Comparison of ReXSonoVQA with existing medical vision-language benchmarks. *Dynamic Reasoning*: Requires understanding of temporal processes, actions, and causal relationships. *Diagnosis*: Evaluates pathology identification or diagnostic interpretation. *Blind Baseline*: Includes text-only evaluation to verify questions require visual input. ✓ indicates present; ✗ indicates absent.

Benchmark	Modality	# Items	Diagnosis	US-Specific	Multi-Organ	Dynamic Reasoning	MCQ	Free-Response	Blind Baseline
MedMCQA	Text	193,155	✓	✗	✓	✗	✓	✗	✗
GMAI-MMBench	Image	26,000	✓	✗	✓	✗	✓	✗	✗
OmniMedVQA	Image	127,995	✓	✗	✓	✗	✓	✗	✗
VQA-RAD	Image	3,515	✓	✗	✓	✗	✓	✓	✗
ReXVQA	Image	696,000	✓	✗	✗	✗	✓	✗	✗
ReXrank	Image	600,000+	✓	✗	✗	✗	✗	✓	✗
U2-Bench	Image	7,241	✓	✓	✓	✗	✓	✓	✗
<b>ReXSonoVQA</b>	<b>Video</b>	<b>514</b>	<b>✗</b>	✓	✓	✓	✓	✓	✓

ture the core competencies required for autonomous ultrasound systems.

#### Type 1: Action–Goal Reasoning

**Task:** Identify the probe maneuver and its immediate acquisition goal (target view/region).

**Motivation:** Autonomous systems must map visual observations to operator intent, the foundation for imitation learning and real-time guidance.

**Example:** “Based on the clip, what maneuver is performed, and what view/region is the operator trying to acquire?”

#### Type 2: Artifact Resolution & Optimization

**Task:** Identify what limits image quality (artifact, occlusion, or ambiguity) and what corrective adjustment resolves it.

**Motivation:** Real-world scanning requires continuous troubleshooting; systems must recognize degradation and select appropriate corrections.

**Example:** “The clip shows reduced clarity or ambiguity. What adjustment is made, and what issue does it resolve?”

#### Type 3: Procedure Context & Planning

**Task:** Infer the current position within the protocol and predict the next step.

**Motivation:** Complete autonomy requires situating local actions within global workflow—knowing what was done and what comes next.

**Example:** “Based on this segment, what protocol window/view has just been completed, and what should the operator move to next?”

### 3.2. Data Curation and Ground Truth Construction

**Data Source and Inclusion Criteria.** We curate ultrasound instructional videos from public sources (e.g., YouTube) based on four criteria:

- **Procedure-rich Narration.** Clear voice-over explaining probe maneuvers, view targets, and troubleshooting rationale.
- **Continuous Scanning Segments.** Sustained scanning sequences rather than static frames.
- **Procedure Relevance.** Focus on acquisition workflow (probe movements, patient positioning, device adjustments, protocol progression) rather than diagnostic interpretation.
- **No On-screen Annotations.** Videos with instructional overlays (e.g., labeled anatomy) are excluded to avoid visual confounding.

Selected videos are manually trimmed to retain only active scanning segments and cropped to preserve only the ultrasound image stream, excluding interface elements (Fig. 3).

**Transcription and Time Alignment.** We transcribe videos using WhisperX (Bain et al., 2023) to produce word-level timestamps. We apply light normalization using an LLM (GPT 5.2) to remove filler words and standardize terminology while preserving

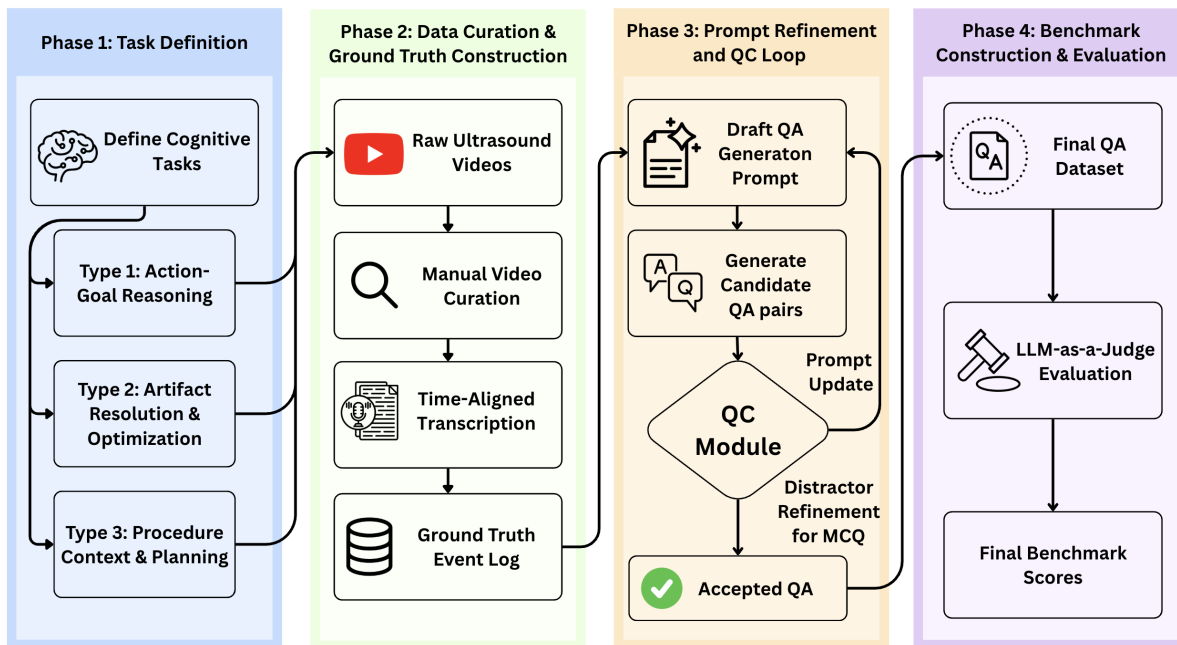


Figure 2: End-to-end pipeline for constructing ReXSonoVQA: (1) Task Definition, (2) Data Curation & Ground Truth Construction, (3) Prompt Refinement and Quality Control Loop, and (4) Benchmark Construction & Evaluation.

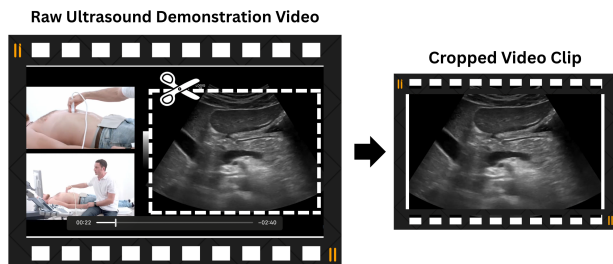


Figure 3: Example of video preprocessing. We crop the original video frame to retain only the ultrasound image stream, excluding surrounding content.

all scanning-relevant content: maneuver descriptions, view targets, and troubleshooting instructions.

**Ground Truth Event Log Construction.** Each video captures a single continuous ultrasound scan, which we represent as a standardized event log: a sequence of time-aligned procedural events (Fig. 4). Each event is anchored to a temporal window ( $time\_start$ ,  $time\_end$ ) and captures a single coherent

procedural unit with two components: ((i) *action*, what the operator does (probe motion, device adjustments, patient repositioning), and (ii) *intent*, the operator’s goal (acquiring a target view, optimizing an acoustic window, confirming orientation, resolving ambiguity). To construct the event log, we use an LLM (GPT 5.2) to convert the time-aligned transcript into structured, timestamped events. LLM-assisted extraction is effective here because instructional narration often describes procedural intent spanning multiple low-level actions (e.g., “rotate to long axis,” “use the liver as an acoustic window,” “reduce sector width”), and the LLM distills these into structured representations directly usable for QA generation.

### 3.3. Prompt Refinement and Quality Control Loop

**Initial Question Generation.** For each video, we generate candidate question-answer (QA) items by prompting an LLM (GPT 5.2) with a full time-segmented event log as input. The model produces a mixture of MCQs and free-response questions, each aligned to a clip window via  $time\_start$  and  $time\_end$ .

Ground-Truth Event Log		Question-Answer Item	
<b>Event 1</b>	<i>0.03–7.64 s</i>	<b>Type:</b> Type 1 (Action–Goal)	
<b>Action:</b>	Place transducer in the midline near the xiphoid; orient indicator cranially.	<b>Format:</b> free-response	<i>Clip: 71.96–82.57 s</i>
<b>Intent:</b>	Identify the liver edge in sagittal view.	<b>Q:</b>	Based on the video clip, what sweep direction is used for the next segment and what is the imaging objective?
<b>Event 2</b>	<i>8.48–17.60 s</i>	<b>A:</b>	The operator begins superiorly and sweeps inferiorly through the remaining segment, then returns to the starting position. The goal is to fully visualize the target region.
<b>Action:</b>	Slide transducer to the patient’s right while maintaining plane.		
<b>Intent:</b>	Bring the portal vein into view.		

Figure 4: Examples from ReXSonoVQA. **Left:** Time-aligned procedural events derived from instructional narration. **Right:** A clip-grounded question-answer item targeting action-goal reasoning.

Questions may derive from a single event or from adjacent event spans to form longer coherent procedural units (e.g., setup → maneuver → confirmation) (see Appendix A, Fig. A1 and A2). Each QA item follows a standard format: (i) question (tagged as MCQ or free-response), (ii) answer (single correct option for MCQ; brief explanatory response for free-response), (iii) question type (Type 1/2/3), (iv) groundtruth (concise rephrasing fusing action and intent), and (v) *time\_start*, *time\_end* timestamps for clip extraction (Fig. 4). To control redundancy and maintain difficulty, the number of QA items generated scales with video duration and event density. Shorter videos yield proportionally fewer questions, reducing trivial or repetitive coverage of the same event.

**Quality Control Loop.** Unlike simple identification tasks, questions for procedural reasoning are more complex and can inadvertently contain answer cues in their phrasing, allowing models to guess without viewing the video. To ensure questions genuinely require video-based understanding, we apply an iterative quality control and prompt refinement loop.

- **Free-Response QC.** We perform blind solvability screening by submitting each question to a VLM without the video. Items answered correctly are flagged as potentially solvable from generic clinical priors or answer-leaking phrasing. Flagged items are manually reviewed to distinguish questions answerable from wording alone from those answered correctly by chance. Trivial items are either (i) rewritten to remove leading cues, (ii) regenerated from the same

event span with stricter prompt constraints, or (iii) removed from the benchmark.

- **MCQ QC.** We run the same text-only evaluation for MCQs, flagging items solved correctly without video. MCQs introduce an additional confound: the correct option may be identifiable through elimination if distractors are implausible or mismatched in abstraction level. During initial generation, we constrain the LLM to produce distractors that are (i) medically plausible within the ultrasound procedure context, (ii) at the same abstraction level as the correct answer (e.g., all maneuvers, all acquisition goals, or all troubleshooting actions), and (iii) compatible with the local anatomy or protocol rather than mixing unrelated organ workflows.

To further improve distractor quality, we implement a post-processing step. For MCQs with weak incorrect options, we sample correct answers from other questions as replacement candidates, which are inherently more procedurally plausible. An LLM (GPT-5.2) then adapts these candidates to the local question context while keeping them incorrect (see Appendix A, Fig. A4). We also shuffle option positions as the LLM disproportionately places correct answers.

### 3.4. Benchmark Construction & Evaluation

The benchmark contains 514 time-localized clips/questions after QC, comprising 249 MCQs and 265 free-response items. The questions span a diverse set

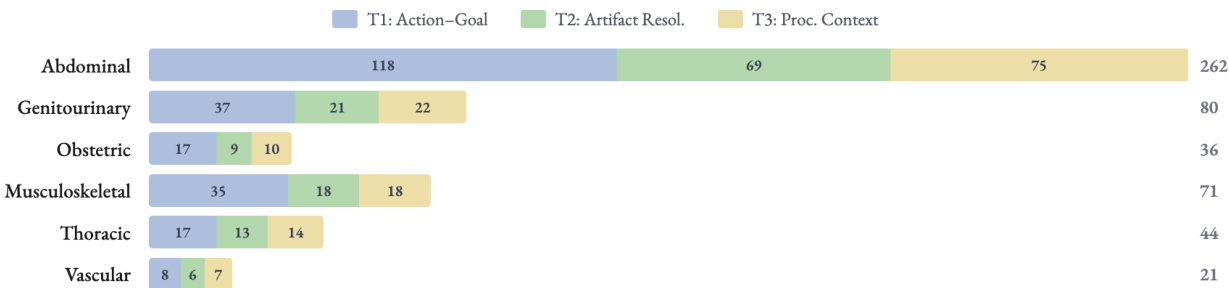


Figure 5: ReXSonoVQA dataset composition (514 items). Distribution of questions across clinical categories by task type.

of ultrasound categories and scanning purposes, including abdominal, genitourinary, musculoskeletal, thoracic, and vascular protocols (Fig. 5). The current video collection contains 115.1 minutes of scanning footage after cropping. Ground truth annotations derive from expert-narrated demonstration videos, and QA generation involved direct input from a board-certified clinician who participated in manual review of flagged items for medical accuracy. Free-response predictions are scored using an LLM-as-a-judge approach, while MCQ performance is evaluated using standard accuracy against the ground-truth option.

## 4. Experiments

### 4.1. Evaluation Method

For evaluation, we perform zero-shot assessment using Gemini 3 Pro, Qwen3.5-397B, LLaVA-Video-72B, and Seed 2.0 Pro by submitting each QA item along with its corresponding video clip with audio track removed. MCQ items are scored using accuracy by comparing the selected option against the ground-truth option. Free-response items are graded using an LLM-as-a-judge protocol with a three-level rubric:

- **Score 2 (Correct):** Both visual evidence and procedural reasoning/conclusion are correct.
- **Score 1 (Partially Correct):** Either the visual evidence is correct but the reasoning/conclusion is wrong, or the reasoning/conclusion is correct but the visual evidence is incorrect.
- **Score 0 (Incorrect):** Both visual evidence and reasoning/conclusion are incorrect.

This scheme captures procedure-relevant failure modes: misinterpreting visual evidence, incorrect causal reasoning, or both. The judge receives the ground truth as context and is explicitly instructed to penalize hallucinations while applying criteria consistently across all questions and models.

### 4.2. Evaluation Settings

To quantify the contribution of visual evidence and identify questions answerable without video, we evaluate the model under two matched input settings. Concretely, each QA item is tested under: (i) *video-informed*, where the model receives the clip corresponding to  $[time\_start, time\_end]$ , and (ii) *text-only*, where the model receives only the question (and MCQ options, if applicable) but no visual input. The text-only setting serves two purposes: it provides a lower-bound reference to quantify the marginal value of visual cues, and it acts as a quality diagnostic, items solvable without video indicate potential issues with question design (generic priors, answer-leaking phrasing, or MCQ option artifacts) and are flagged for revision during QC.

## 5. Results

We evaluate Gemini 3 Pro, Qwen3.5-397B, LLaVA-Video-72B, and Seed 2.0 Pro under both video-informed and text-only (blind) settings to assess whether the benchmark meaningfully requires visual understanding. Results are summarized in Figs. 6 and 7, with cross-setting outcome tables in Appendix C (Tables A1–A8). We additionally include qualitative case studies (Fig. A7 and A10, Ap-

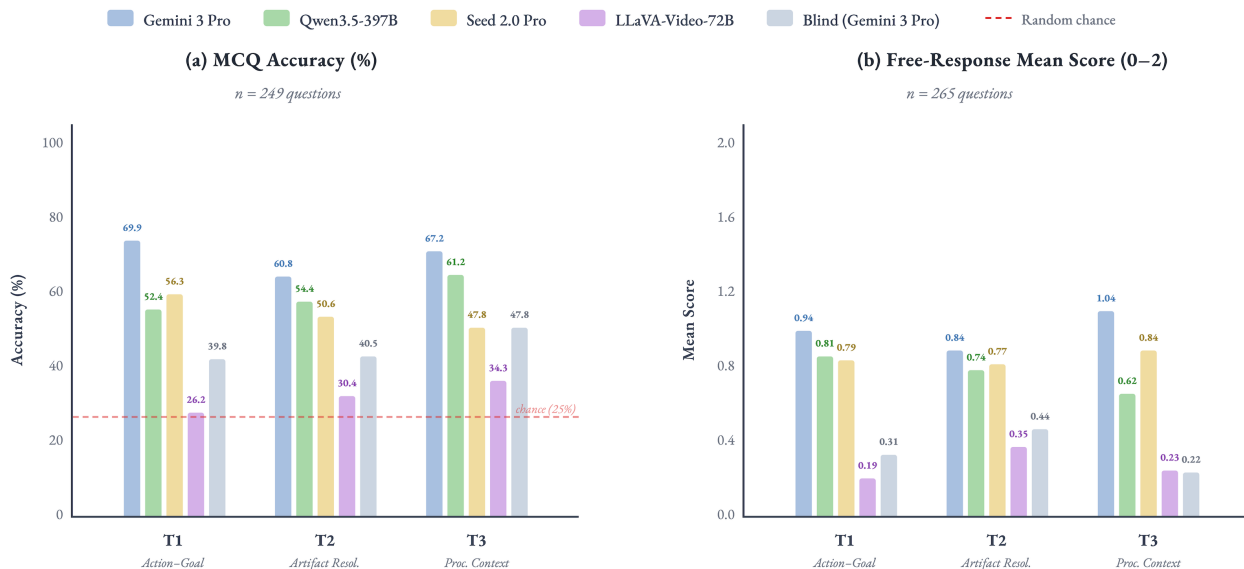


Figure 6: Multi-model zero-shot MCQ accuracy (%), left) and free-response mean score (0–2 rubric, right) by task type under paired evaluation settings. Dashed red line marks 25% random chance for MCQ.

pendix B) that highlight typical failure modes and serve as concrete examples of benchmark items.

#### Task-wise Trends and the Role of Duration.

We present detailed analysis for Gemini 3 Pro, the best-performing model, followed by a multi-model comparison. Across both MCQ and free-response evaluations, Gemini 3 Pro struggles most on Type 2 (Artifact Resolution & Optimization) (Fig. 6). For MCQ, Type 1 reaches the highest accuracy (69.9%), while Type 2 has the lowest (60.8%), indicating that troubleshooting remains challenging even given the visual evidence is available. For free-response (0–2 rubric), Type 3 attains the highest mean score with video (1.04), whereas Type 2 again yields the lowest (0.84), reinforcing that resolving artifacts and ambiguity requires tighter causal reasoning beyond recognizing salient motion cues.

Performance improves with clip duration across both formats: MCQ accuracy ranges from 61.0% (0–5s) to 75.8% (>20s), and free-response scores vary from 0.89 (0–5s) to 1.11 (>20s) (Fig. 7). This pattern suggests longer clips provide richer temporal context that aids procedural reasoning. Both factors influence difficulty, with Type 2 tasks and short clips presenting the greatest challenges.

**Multi-Model Comparison.** Figs. 6 and 7 compare all four evaluated models. Gemini 3 Pro achieves the highest overall performance across both formats. Qwen3.5-397B and Seed 2.0 Pro show meaningful gains over the blind baseline, while LLaVA-Video-72B performs near or below blind levels, suggesting limited procedural video understanding. Across all models, Type 2 remains the most challenging on average on all tasks and question types, confirming that troubleshooting reasoning represents a fundamental challenge rather than a model-specific weakness.

**Visual Input Helps, but Not Uniformly.** Visual input provides a clear overall benefit, but the margin depends on question type and format. For MCQ, overall accuracy improves from 42.2% in the text-only setting to 66.3% with video (+24.1 points), with the largest gain in Type 1 (+30.1) and the smallest in Type 3 (+19.4). For free-response, the overall mean score increases from 0.31 (text-only) to 0.95 (with video) (+0.64), with the weakest improvement for Type 2 (+0.40).

Duration-stratified results reveal distinct patterns between formats. For MCQ, gains from video fluctuate across duration bins, while the text-only baseline consistently rises with clip length (from 26.8% to 51.5%). This pattern suggests that longer MCQ questions carry more contextual cues enabling par-

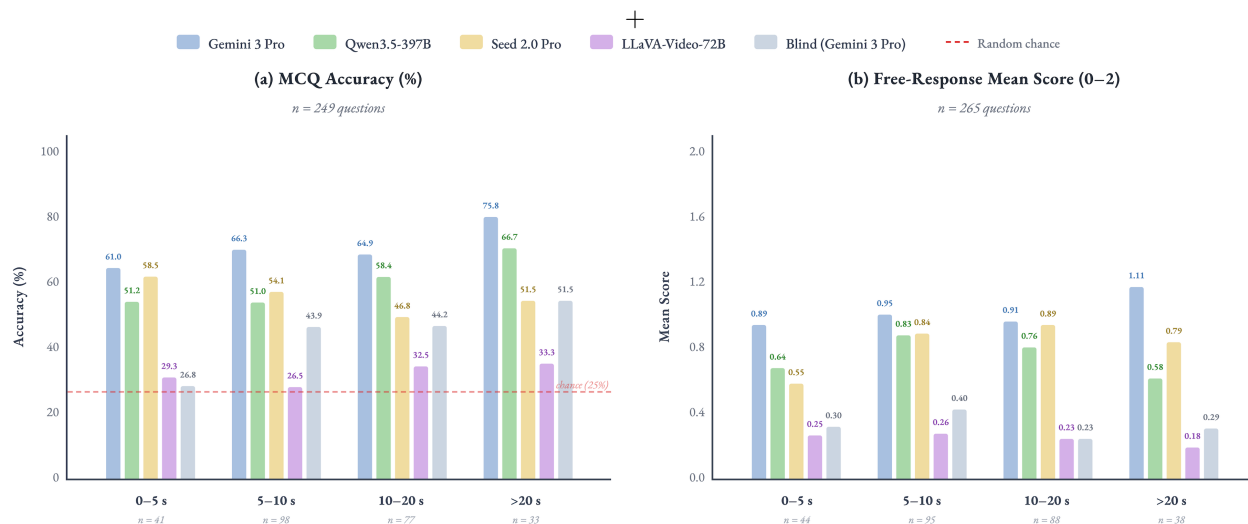


Figure 7: Multi-model zero-shot MCQ accuracy (%), left) and free-response mean score (0–2 rubric, right) by video duration under paired evaluation settings. Dashed red line marks 25% random chance for MCQ.

tial answering without visual evidence. In contrast, free-response shows increasing gains from video with longer durations (from +0.59 for 0–5 s to +0.82 for >20 s), while the text-only baseline fluctuates rather than increasing monotonically. This indicates that free-response questions benefit more from extended temporal context and are less confounded by question length than MCQs, demonstrating stronger dependence on genuine visual understanding.

The cross-setting outcome tables (Tables A1–A2, Appendix C) reveal additional nuance for Gemini 3 Pro. While video context generally improves performance, it is not without noise: a small number of cases transition from correct to incorrect, suggesting that visual features can sometimes be misleading. More significantly, the bulk of previously incorrect responses remained unsolved even with video, indicating that the visual stream did not always provide the information needed to correct text-only guessing failures. Specifically, in the MCQ cross-setting table, 68 items transition from incorrect (blind) to correct (with video), yielding a 47% recovery rate (68/144); the 76 items remaining incorrect confirm substantial residual difficulty. The free-response cross-setting table reveals that video input selectively resolves perception gaps (64 items improve to full credit), while 98 items remain at 0 points and 22 items show correct perception but persistent reasoning errors, demon-

strating that procedural understanding requires both visual perception and causal clinical reasoning.

Taken together, these results confirm that the benchmark meaningfully relies on video, with free-response questions showing particularly strong dependence on visual evidence. However, troubleshooting questions (Type 2) remain the least reliably solved in terms of marginal gains and likely demand stronger causal grounding than current zero-shot VLMs provide.

## 6. Limitations and Future Work

**MCQ blind accuracy exceeds random guessing.** Blind accuracy substantially exceeds random chance (25%), indicating exploitable patterns in question phrasing or option structure that models can leverage without viewing the video. This can arise from distractors that are not equally plausible, mismatched abstraction levels across options, or subtle linguistic cues favoring certain options. While our distractor polishing and blinding QC reduce this effect, further improvements include harder distractor generation, adversarial option auditing, and automated elimination-bias checks could push blind performance closer to chance. Notably, free-response questions exhibit much lower blind baselines and stronger dependence on visual evidence, suggest-

ing they may be better suited for evaluating video-grounded procedural reasoning.

**Annotation scope and domain coverage.** Our ground-truth answers may under-specify clinically correct details that are visually plausible from the clip, which can complicate evaluation for open-ended questions where multiple specificity levels are acceptable. For example (Fig. 1), an answer key may refer to “the tendon,” while a model might correctly identify a more specific structure (e.g., within the carpal tunnel). We mitigate this through rubric-based grading with partial credit and plan to expand annotations with multiple acceptable reference answers. Additionally, we deliberately selected instructional videos because expert narration provides natural supervision labels enabling scalable ground truth construction. Extending to in-hospital clinical footage is part of our proposed roadmap as clinical partnerships are established; the pipeline design allows integration of diverse video sources with minimal modification.

**Model coverage.** We evaluate Gemini 3 Pro, Qwen3.5-397B, LLaVA-Video-72B, and Seed 2.0 Pro, all supporting native video input (Figs. 6 and 7). Existing medical VLMs rely solely on image-to-language architectures without video capability, and other leading commercial VLMs (GPT-5.2, Claude Opus 4.5) lack native video support. As video-language models with medical-specific pretraining emerge, future work will extend evaluation to more models.

## 7. Conclusion

We introduce ReXSonoVQA, a procedure-centric ultrasound video QA benchmark evaluating VLMs’ understanding of ultrasound scanning technique. Built from instructional videos with time-aligned event logs, it comprises 514 clip-grounded questions (249 MCQ, 265 free-response) across six clinical categories and three procedural competencies: Action-Goal Reasoning, Artifact Resolution & Optimization, and Procedure Context & Planning. An iterative quality control loop ensures questions require genuine video-based reasoning through text-only screening and distractor refinement.

Zero-shot evaluation across four models (Gemini 3 Pro, Qwen3.5-397B, LLaVA-Video-72B, and Seed 2.0 Pro) reveals that while VLMs can extract procedural information from video, substantial gaps remain. Type 1 (Action-Goal) achieves highest MCQ

performance while Type 3 (Procedure Context) excels in free-response, but Type 2 (Artifact Resolution) shows minimal improvement over text-only baseline, exposing limitations in causal troubleshooting reasoning. Performance improves with clip duration, with free-response questions demonstrating stronger dependence on visual evidence. However, most text-only failures remain unsolved even with video, indicating fundamental procedural understanding gaps.

REXSonoVQA establishes the first evaluation framework for video-based procedural ultrasound understanding, providing a foundation for developing VLM perception systems critical to ultrasound training, real-time clinical guidance, and robotic automation.

## References

- Anthropic. Claude API Documentation. <https://docs.anthropic.com/en/api/>.
- Max Bain, Jaesung Huh, Tengda Han, and Andrew Zisserman. Whisperx: Time-accurate speech transcription of long-form audio. *INTERSPEECH 2023*, 2023.
- Mingcong Chen, Siqi Fan, Guanglin Cao, Yun-hui Liu, and Hongbin Liu. Uspilot: An embodied robotic assistant ultrasound system with a large language model enhanced graph planner. *IEEE Robotics and Automation Letters*, 10(10):10027–10034, 2025. doi: 10.1109/LRA.2025.3598625.
- Pengcheng Chen, Jin Ye, Guoan Wang, Yanjun Li, Zhongying Deng, Wei Li, Tianbin Li, Haodong Duan, Ziyang Huang, Yanzhou Su, Benyou Wang, Shaoting Zhang, Bin Fu, Jianfei Cai, Bohan Zhuang, Eric J Seibel, Yu Qiao, and Junjun He. Gmai-mmbench: A comprehensive multimodal evaluation benchmark towards general medical ai. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 94327–94427. Curran Associates, Inc., 2024. doi: 10.52202/079017-2992.
- Pascale Fung, Yoram Bachrach, Asli Celikyilmaz, Kamalika Chaudhuri, Delong Chen, Willy Chung, Emmanuel Dupoux, Hongyu Gong, Hervé Jégou, Alessandro Lazaric, Arjun Majumdar, Andrea Madotto, Franziska Meier, Florian Metze, Louis-Philippe Morency, Théo Moutakanni, Juan Pino,

- Basile Terver, Joseph Tighe, Paden Tomasello, and Jitendra Malik. Embodied ai agents: Modeling the world, 2025. URL <https://arxiv.org/abs/2506.22355>.
- Emmett D. Goodman, Krishna K. Patel, Yilun Zhang, William Locke, Chris J. Kennedy, Rohan Mehrotra, Stephen Ren, Melody Guan, Orr Zohar, Maren Downing, Hao Wei Chen, Jevin Z. Clark, Margaret T. Berrigan, Gabriel A. Brat, and Serena Yeung-Levy. Analyzing surgical technique in diverse open surgical videos with multitask machine learning. *JAMA Surgery*, 159(2):185–192, 02 2024. ISSN 2168-6254. doi: 10.1001/jamasurg.2023.6262.
- Junhao Guo, XueFeng Shan, Guoming Wang, Dong Chen, Rongxing Lu, and Siliang Tang. Llaus: A high-quality instruction-tuned large vision language assistant for ultrasound. In *Proceedings of the 2025 International Conference on Multimedia Retrieval, ICMR '25*, page 398–406, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400718779. doi: 10.1145/3731715.3733374. URL <https://doi.org/10.1145/3731715.3733374>.
- Xiaoqing Guo, Mohammad Alsharid, He Zhao, Yipei Wang, Jayne Lander, Aris T. Papa-georghiou, and J. Alison Noble. A visually grounded language model for fetal ultrasound understanding. *Nature Biomedical Engineering*, 2026. doi: <https://doi.org/10.1038/s41551-025-01578-3>. URL <https://www.nature.com/articles/s41551-025-01578-3>.
- Yutao Hu, Tianbin Li, Quanfeng Lu, Wenqi Shao, Junjun He, Yu Qiao, and Ping Luo. Omnimed-vqa: A new large-scale comprehensive evaluation benchmark for medical lvlm. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22170–22183, June 2024.
- Haojun Jiang, Andrew Zhao, Qian Yang, Xiangjie Yan, Teng Wang, Yulin Wang, Ning Jia, Jiangshan Wang, Guokun Wu, Yang Yue, Shai Luo, Huanqian Wang, Ling Ren, Siming Chen, Pan Liu, Guocai Yao, Wenming Yang, Shiji Song, Xiang Li, Kunlun He, and Gao Huang. Towards expert-level autonomous carotid ultrasonography with large-scale learning-based robotic system. *Nature Communication*, 16:7893, 2025. doi: <https://doi.org/10.1038/s41467-025-62865-w>. URL <https://www.nature.com/articles/s41551-025-01546-x>.
- Zhongliang Jiang, Septimiu E. Salcudean, and Nasir Navab. Robotic ultrasound imaging: State-of-the-art and future perspectives. *Medical Image Analysis*, 89:102878, 2023. ISSN 1361-8415. doi: <https://doi.org/10.1016/j.media.2023.102878>. URL <https://www.sciencedirect.com/science/article/pii/S136184152300138X>.
- 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):1–10, 2018.
- Anjie Le, Henan Liu, Yue Wang, Zhenyu Liu, Rongkun Zhu, Taohan Weng, Jinze Yu, Boyang Wang, Yalun Wu, Kaiwen Yan, Quanlin Sun, Meirui Jiang, Jialun Pei, Siya Liu, Haoyun Zheng, Zhoujun Li, Alison Noble, Jacques Souquet, Xiaoqing Guo, Manxi Lin, and Hongcheng Guo. U2-bench: Benchmarking large vision-language models on ultrasound understanding, 2025. URL <https://arxiv.org/abs/2505.17779>.
- 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. *arXiv preprint arXiv:2306.00890*, 2023.
- Fadillah Maani, Numan Saeed, Tausifa Saleem, Zaid Farooq, Hussain Alasmawi, Werner Diehl, Ameera Mohammad, Gareth Waring, Saudabi Valappi, Leanne Bricker, and Mohammad Yaqub. Fetal-clip: A visual-language foundation model for fetal ultrasound image analysis, 2025. URL <https://arxiv.org/abs/2502.14807>.
- Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips, 2019.
- Michael Moor, Qian Huang, Shirley Wu, Michihiro Yasunaga, Yash Dalmia, Jure Leskovec, Cyril Zakkka, Eduardo Pontes Reis, and Pranav Rajpurkar. Med-flamingo: a multimodal medical few-shot learner. In Stefan Hegselmann, Antonio Parziale, Divya Shanmugam, Shengpu Tang, Mercy Nyame-waa Asiedu, Serina Chang, Tom Hartvigsen, and

- Harvineet Singh, editors, *Proceedings of the 3rd Machine Learning for Health Symposium*, volume 225 of *Proceedings of Machine Learning Research*, pages 353–367. PMLR, 10 Dec 2023. URL <https://proceedings.mlr.press/v225/moor23a.html>.
- Khushboo Munir, Abdullah F. Al-Battal, Ammar Alsheghri, Harald Becher, Michelle Noga, and Kumaradevan Punithakumar. A survey of autonomous robotic ultrasound scanning systems. *IEEE Access*, 13:103178–103197, 2025. doi: 10.1109/ACCESS.2025.3574464.
- Vishwesh Nath, Wenqi Li, Dong Yang, Andriy Myronenko, Mingxin Zheng, Yao Lu, Zhijian Liu, Hongxu Yin, Yee Man Law, Yucheng Tang, Pengfei Guo, Can Zhao, Ziyue Xu, Yufan He, Stephanie Harmon, Benjamin Simon, Greg Heinrich, Stephen Aylward, Marc Edgar, Michael Zephyr, Pavlo Molchanov, Baris Turkbey, Holger Roth, and Daguang Xu. Vila-m3: Enhancing vision-language models with medical expert knowledge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14788–14798, June 2025.
- OpenAI. OpenAI Platform Documentation: Models. <https://platform.openai.com/docs/models>.
- Ankit Pal, Logesh Kumar Umaphathi, and Malaikanan Sankarasubbu. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In Gerardo Flores, George H Chen, Tom Pollard, Joyce C Ho, and Tristan Naumann, editors, *Proceedings of the Conference on Health, Inference, and Learning*, volume 174 of *Proceedings of Machine Learning Research*, pages 248–260. PMLR, 07–08 Apr 2022. URL <https://proceedings.mlr.press/v174/pal22a.html>.
- Ankit Pal, Jung-Oh Lee, Xiaoman Zhang, Malaikanan Sankarasubbu, Seunghyeon Roh, Won Jung Kim, Meesun Lee, and Pranav Rajpurkar. Rextvqa: A large-scale visual question answering benchmark for generalist chest x-ray understanding, 2025.
- Sahar Salimpour, Lei Fu, Kajetan Rachwał, Pascal Bertrand, Kevin O’Sullivan, Robert Jakob, Farhad Keramat, Leonardo Militano, Giovanni Toffetti, Harry Edelman, and Jorge Peña Queraltá. Towards embodied agentic ai: Review and classification of llm- and vlm-driven robot autonomy and interaction, 2025. URL <https://arxiv.org/abs/2508.05294>.
- Harshita Sharma, Lior Drukker, Aris T. Papa-georghiou, and J. Alison Noble. Machine learning-based analysis of operator pupillary response to assess cognitive workload in clinical ultrasound imaging. *Computers in Biology and Medicine*, 135:104589, 2021. ISSN 0010-4825. doi: <https://doi.org/10.1016/j.compbiomed.2021.104589>. URL <https://www.sciencedirect.com/science/article/pii/S0010482521003838>.
- Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. Towards generalist foundation model for radiology by leveraging web-scale 2d&3d medical data, 2023. URL <https://arxiv.org/abs/2308.02463>.
- Kai Zhang, Rong Zhou, Eashan Adhikarla, Zhiling Yan, Yixin Liu, Jun Yu, Zhengliang Liu, Xun Chen, Brian D. Davison, Hui Ren, Jing Huang, Chen Chen, Yuyin Zhou, Sunyang Fu, Wei Liu, Tianming Liu, Xiang Li, Yong Chen, Lifang He, James Zou, Quanzheng Li, Hongfang Liu, and Lichao Sun. A generalist vision-language foundation model for diverse biomedical tasks. *Nature Medicine*, 30(11): 3129–3141, August 2024. ISSN 1546-170X. doi: 10.1038/s41591-024-03185-2. URL <http://dx.doi.org/10.1038/s41591-024-03185-2>.
- Xiaoman Zhang, Hong-Yu Zhou, Xiaoli Yang, Oishi Banerjee, Julián N. Acosta, Josh Miller, Ouwen Huang, and Pranav Rajpurkar. Rextrank: A public leaderboard for ai-powered radiology report generation. In Junde Wu, Jiayuan Zhu, Min Xu, and Yueming Jin, editors, *Proceedings of The First AAAI Bridge Program on AI for Medicine and Healthcare*, volume 281 of *Proceedings of Machine Learning Research*, pages 90–99. PMLR, 25 Feb 2025. URL <https://proceedings.mlr.press/v281/zhang25b.html>.

## Appendix A. Prompts

See Fig [A1](#), [A2](#), [A3](#), [A4](#), [A5](#)

## Appendix B. Case Studies

See Fig [A7](#), [A8](#), [A9](#), [A10](#), [A11](#)

## Appendix C. Cross-Setting Outcome Tables

Tables A1–A8 report the cross-tabulation of video-informed vs. text-only (blind) outcomes for all four evaluated models, separately for MCQ and free-response formats. Each cell counts the number of items falling into a given (blind outcome, video-informed outcome) pair, providing a fine-grained view of how visual input changes model predictions.

Table A1: Gemini 3 Pro: MCQ cross-setting outcome table ( $n = 249$ ). Rows = blind (text-only) outcome; columns = video-informed outcome.

	Vid. Correct	Vid. Incorrect	Total
<b>Blind Correct</b>	97	8	105
<b>Blind Incorrect</b>	68	76	144
<b>Total</b>	165	84	249

Table A2: Gemini 3 Pro: Free-response cross-setting outcome table ( $n = 265$ ). Rows = blind outcome; columns = video-informed outcome. W.C. = wrong conclusion; W.V. = wrong visual evidence.

	2 pt	1 pt (W.C.)	1 pt (W.V.)	0 pt	Total
<b>2 pt</b>	15	2	1	1	19
<b>1 pt (W.C.)</b>	7	4	2	8	21
<b>1 pt (W.V.)</b>	13	1	3	6	23
<b>0 pt</b>	64	22	18	98	202
<b>Total</b>	99	29	24	113	265

Table A3: LLaVA-Video-72B: MCQ cross-setting outcome table ( $n = 249$ ).

	Vid. Correct	Vid. Incorrect	Total
<b>Blind Correct</b>	45	60	105
<b>Blind Incorrect</b>	29	115	144
<b>Total</b>	74	175	249

Table A4: LLaVA-Video-72B: Free-response cross-setting outcome table ( $n = 265$ ).

	2 pt	1 pt (W.C.)	1 pt (W.V.)	0 pt	Total
<b>2 pt</b>	0	2	6	11	19
<b>1 pt (W.C.)</b>	1	1	3	16	21
<b>1 pt (W.V.)</b>	4	2	3	14	23
<b>0 pt</b>	2	16	16	168	202
<b>Total</b>	7	21	28	209	265

Table A5: Qwen3.5-397B: MCQ cross-setting outcome table ( $n = 249$ ).

	Vid. Correct	Vid. Incorrect	Total
<b>Blind Correct</b>	66	39	105
<b>Blind Incorrect</b>	72	72	144
<b>Total</b>	138	111	249

Table A6: Qwen3.5-397B: Free-response cross-setting outcome table ( $n = 265$ ).

	2 pt	1 pt (W.C.)	1 pt (W.V.)	0 pt	Total
<b>2 pt</b>	13	1	1	4	19
<b>1 pt (W.C.)</b>	7	0	3	11	21
<b>1 pt (W.V.)</b>	9	2	5	7	23
<b>0 pt</b>	45	13	23	121	202
<b>Total</b>	74	16	32	143	265

Table A7: Seed 2.0 Pro: MCQ cross-setting outcome table ( $n = 249$ ).

	Vid. Correct	Vid. Incorrect	Total
<b>Blind Correct</b>	65	40	105
<b>Blind Incorrect</b>	65	79	144
<b>Total</b>	130	119	249

Table A8: Seed 2.0 Pro: Free-response cross-setting outcome table ( $n = 265$ ).

	2 pt	1 pt (W.C.)	1 pt (W.V.)	0 pt	Total
<b>2 pt</b>	13	1	0	5	19
<b>1 pt (W.C.)</b>	7	4	1	9	21
<b>1 pt (W.V.)</b>	10	1	4	8	23
<b>0 pt</b>	50	18	23	111	202
<b>Total</b>	80	24	28	133	265

**Prompt A1: QA generation**

You are generating Q&A items for a benchmark that evaluates a Vision-Language Model (VLM) on ULTRASOUND DEMONSTRATION VIDEOS.

CRITICAL EVAL SETUP (what the VLM will see at test time):

- Video ONLY (muted). No audio.
- Therefore, every question must be answerable from VISUAL cues only:
  - probe motion (slide/sweep/rock/rotate/tilt, directionality, marker orientation when visible),
  - patient positioning/motion if visually observable,
  - ultrasound image dynamics (structures entering/leaving frame, plane changes, motion, shadowing),
  - mode changes that are visually obvious (e.g., Color Doppler overlay appears/disappears),
  - imaging parameter changes if visible in the image effect (e.g., depth scale changes, focal zone marker moves, sector narrows).

GROUND TRUTH YOU WILL RECEIVE:

- A sequence of time-stamped events extracted from transcript, each with:
  - action (what operator did)
  - interpretation (why / what they were looking for)

These are NOT available to the VLM at test time; they are for YOU to craft correct QA.

THE MOST CRITICAL QUESTION CONSTRUCTION RULE:

Each question must be answerable only by watching the video. Do not include visual clues, descriptions, or contextual hints that would allow someone to answer or guess correctly without viewing the video. Do not contain specific visual details, structures, context, artifacts, or outcomes in the question itself.

BAD (includes visual clues): "When the operator sweeps inferiorly and identifies the bifurcation, what anatomical landmark is being visualized?"

GOOD (requires watching): "Based on the clip, what acquisition goal is being pursued during the continuous inferior sweep, and what key anatomic endpoint indicates success?"

BAD (describes what's visible): "As bowel gas obscures the vessel view and the operator applies firm pressure..."

GOOD (asks about action/goal): "The clip shows loss/obscuration of the target vessel. What optimization maneuver is performed, and what problem is it addressing to restore the view?"

Always phrase questions as:

- "Based on the video/clip, what..."
- "What [action/strategy/technique] is shown, and what is the goal/objective?"
- "This clip shows [general/non-specific observation]. What is being done and why?"

You must generate items in exactly 3 clinically-meaningful question types:

- 1) Type1\_ActionGoalReasoning
  - Tests: action reasoning + goal inference.
  - Ask: what maneuver is being performed AND what imaging goal / target view it serves.
  - Phrase questions naturally and adaptively based on the content.
  - Do NOT describe the specific anatomical structures or visual details in the question.
- 2) Type2\_ArtifactResolutionOptimization
  - Tests: overcoming artifacts or ambiguity + optimization/disambiguation logic.
  - Ask: what (probe maneuver, patient management, or knobology) has changed AND why it resolves an artifact or ambiguity / improves image quality.
  - IMPORTANT: Do not explicitly describe the artifact or ambiguity in the question.
  - Phrase questions naturally - you can reference general observations like "loss of view" or "poor quality" without describing specific details.

Figure A1: Prompt template used for QA generation.

**Prompt A1(continued): QA generation**

- 3) Type3\_ProcedureContextPlanning
- Tests: overall context understanding + next-step planning.
  - Ask: what phase/step the operator is in AND what the broader workflow objective or next logical step is.
  - Usually use TWO or more ADJACENT EVENTS to create sufficient context.
  - Vary your phrasing - ask about exam phases, workflow transitions, procedural objectives, or strategies as appropriate.
  - Do not describe the clip's content or sequence and do not name specific anatomy or maneuvers in the question itself.

**QUESTION FORMAT MIX:**

- Produce a mix of MCQ and free response.
- For MCQ:
  - \* Put exactly 4 options (A-D) inside the "question" string. Ensure only ONE is correct; distractors must be plausible.
  - \* Keep options generic and hypothesis-based when possible
  - \* Do NOT make options that give away visual details that should only be known from watching the video
  - \* IMPORTANT: The "answer" field must contain ONLY the letter of the correct option (e.g., "A" or "B" or "C" or "D")
- For Free response:
  - \* The "question" must be answerable concisely; the "answer" should be 1-3 sentences.

**QUESTION PHRASING PRINCIPLES:**

- Vary your question style naturally based on the content and question type
- Ask WHAT is being done and WHY, without describing what is visible

**TIME GROUNDING:**

- Each item must include time\_start and time\_end aligned to the event(s) used:
  - Single event: use its exact start/end
  - Two or more adjacent events: time\_start = min(starts), time\_end = max(ends)

**GROUNDTRUTH FIELD:**

- Rephrase the action + interpretation into ONE integrated entry (1 to 3 sentences) that preserves all key details.
- Do not introduce details not supported by action/interpretation.

**AVOID INVALID ITEMS:**

- Do NOT reference audio, narration, or transcript.
- Do NOT ask for reading on-screen text or exact numeric readouts (e.g., "104 mm") unless it is guaranteed without OCR (assume it is not).
- If an action is a spoken instruction (e.g., "hold breath"), convert it into an EFFECT-based question that could be inferred visually (e.g., reduced respiratory motion / improved acoustic window), not "what instruction was said".

Return JSON only, matching the schema provided.

Figure A2: Prompt template used for QA generation (continued).

**Prompt A2: MCQ distractor refinement**

You are a medical MCQ editor. Improve distractors so they are clinically plausible and not trivially wrong.

Rules:

- Do NOT change the question stem.
- Do NOT change the correct option text.
- Provide only the requested distractors.
- Distractors must be medically plausible and consistent with the exam context, but clearly incorrect.
- Avoid nonsensical or unrelated anatomy/procedures.
- Keep each option concise (1 sentence max).
- Do not include letters like 'A.' in the option text; return raw option text only.

```
"".strip()
```

```
user_payload = {
    "stem": stem,
    "correct_option": correct_text,
    "distractor_letters": distractor_letters,
    "exemplar_options": exemplar_options,
}
```

Figure A3: Prompt template used for MCQ distractor refinement.

**Prompt A3: Inference**

You are an expert ultrasound clinician evaluating video clips from ultrasound examinations.

You will be shown a video clip and asked a question about it. The video contains ultrasound examination footage with the audio muted. You must answer based ONLY on what you observe visually in the video.

Key Guidelines:

- Watch the video carefully, observing probe movements, image changes, and anatomical structures
- Answer based solely on visual evidence from the video
- For multiple choice questions (MCQ): Start your response with "Answer: X" where X is the correct letter (A, B, C, or D), then provide a brief explanation.
- For free response questions: Provide a concise, accurate answer in 1-3 sentences
- If you cannot determine the answer from the video, state that clearly

Your response should be direct and factual, based on what you observe in the video clip.

Figure A4: Prompt template used for model inference.

**Prompt A4: LLM-as-a-judge for Free-Responses****STRICTNESS OVERRIDE:**

- Penalize guessing or vague organ references.
- If the ground truth names a specific organ/structure/feature and the prediction only gives a generic organ reference (e.g., "an organ" or a different unspecified organ), treat this as incorrect visual evidence.
- If the prediction is uncertain or overly general (e.g., "likely an organ" without identification), do NOT award full credit. At most score=1, and classify the error as wrong\_visual\_evidence.
- Non-identification of the required organ/feature should be categorized as visual evidence error.

```
"".strip()
```

```
system_prompt = ""You are an impartial medical expert judge evaluating the quality of an AI assistant's response to a clinical ultrasound question.
```

Compare the AI's Prediction against the Ground Truth Answer.

**Scoring Criteria:**

- 2 - Correct: The conclusion/answer matches the ground truth AND the cited visual evidence aligns with the ground truth. Minor phrasing differences are acceptable.
- 1 - Partially Correct: Only one of the two is correct (either the conclusion is correct but the visual evidence is wrong, OR the visual evidence is correct but the conclusion is wrong).
- 0 - Incorrect: The prediction is wrong, irrelevant, hallucinated, or contradicts the ground truth.

If score = 1, classify the mistake type:

- wrong\_visual\_evidence: conclusion is correct, but evidence/visual justification is incorrect or mismatched.
- wrong\_conclusion: evidence/visual cues are correct, but the conclusion/answer is wrong.

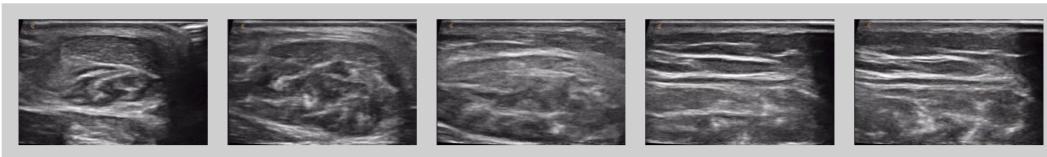
If score = 0, classify the mistake type:

- wrong\_visual\_evidence: evidence/visual justification is incorrect.
- wrong\_conclusion: conclusion/answer is incorrect.
- both\_fail: both evidence and conclusion are incorrect.

Output exactly and ONLY valid JSON in this format:

```
{"score": 0, "explanation": "Brief reasoning", "error_type": "none"}
```

Figure A5: Prompt template used for LLM-as-a-judge.



**Q:** Based on the video clip, during the transverse portion of the exam, what is the screening objective as the operator sweeps through and then continues proximally, and what anatomical transition marks completion of this segment?"


**GT:** The objective is to screen the tendon cross-section for focal defects, fluid, or complete disruption while following it proximally. This segment is completed as the tendon morphology changes and it transitions into muscle.

**Gemini 3 Pro (Score: 1/2):** Based on the video clip, the screening objective during the transverse sweep is to visualize the flexor tendons (specifically within the carpal tunnel) as the probe moves from a distal position at the wrist proximally into the distal forearm. The anatomical transition that marks the completion of this segment is the transition from the distinct, bundled tendons within the carpal tunnel to the musculotendinous junction in the distal forearm, where the hyperechoic tendon fibers merge into the hypoechoic muscle bellies of the flexor group.

**GPT 5.2 (Score: 0/2):** During the transverse sweep, the objective is to screen the colon for an ileocolic intussusception, looking for the concentric "target/donut" bowel-within-bowel appearance, and then follow it proximally to define its extent/location. This transverse screening segment is considered complete when the sweep reaches the right upper quadrant and the ascending colon transitions at the hepatic flexure into the transverse colon (colon seen beneath the liver).

**Claude Opus 4.5 (Score: 0/2):** Based on the video clip, during the transverse portion of the exam, the screening objective is to assess vein compressibility at serial intervals by applying probe pressure to ensure complete coaptation of the vessel walls, thereby ruling out deep vein thrombosis. The operator sweeps through the vessel while performing compression maneuvers to detect any intraluminal thrombus. The segment is completed when reaching the saphenofemoral junction, which marks the anatomical transition point where the great saphenous vein joins the common femoral vein at the proximal thigh.

Figure A6: **Exploratory evaluation.** Demonstrated why GPT-5.2 and Claude Opus 4.5 were excluded from our benchmark's baseline. Without native video support, both models hallucinate completely unrelated anatomical structures (colon/intussusception and vein/thrombosis) when provided with sequential images, scoring 0/2. In contrast, Gemini 3 Pro with native video understanding correctly identifies the anatomical transition, achieving 1/2. This stark difference motivated our focus on models with native video support.



**Q:** Based on the video clip, what probe angulation strategy is shown, and what key vascular landmark view is being sought?"

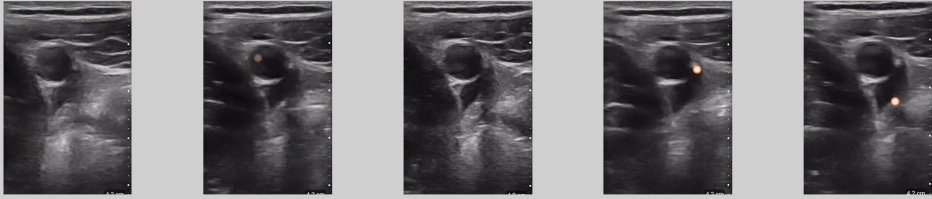
**GT:** The probe is angled cranially to reach the most superior standard cross-section. The goal is to capture the hepatic veins converging toward the IVC (the classic liver vein star appearance) and identify the IVC.

**A:** Based on the video clip, the probe is being **\*\*angled or swept superiorly\*\*** (fanned upwards) from a subcostal position. The key vascular landmark view being sought is the confluence of the **\*\*hepatic veins\*\*** (specifically the middle and left hepatic veins seen joining) as they drain into the inferior vena cava (IVC) near the diaphragm.

**Score:** 2/2

**Judge:** The AI correctly identifies the probe maneuver (angling superiorly/cranially) and the specific vascular landmarks being sought (hepatic veins converging into the IVC), matching the ground truth perfectly." cortical echogenicity

Figure A7: **Case Study 1 (Free-Response, Type 1: Action–Goal Reasoning)**. Gemini 3 Pro correctly identifies the probe maneuver and acquisition target.



**Q:** Based on the video clip, what visual comparison is used to tell two adjacent vessels apart, and what identification ambiguity does this resolve?"


**GT:** They are distinguished by relative depth: the artery is the more superficial vessel and the vein lies deeper. This resolves confusion between the paired vessels when both are seen together.

**A:** Based on the video clip, the visual comparison used is the compressibility of the vessels when pressure is applied with the ultrasound probe. This resolves the ambiguity between the Internal Jugular Vein (IJV), which collapses easily under pressure, and the Carotid Artery (CA), which remains pulsatile and resists compression.

**Score:** 0/2

**Judge:** The AI prediction is incorrect. The ground truth states that the vessels are distinguished by their relative depth (with the artery being superficial to the vein), whereas the AI incorrectly claims the distinction is made via compressibility. Furthermore, the AI assumes the vessels are the IJV and Carotid, but the depth relationship described in the ground truth (artery superficial to vein) contradicts standard IJV/Carotid anatomy (where the vein is typically superficial), suggesting the AI misidentified the anatomy as well.

Figure A8: **Case Study 2 (Free-Response, Type 2: Artifact Resolution & Optimization)**. Gemini 3 Pro hallucinates a compression maneuver instead of the actual optimization action.



Q: Based on the video clip, the operator presses down with the probe and then releases while keeping the target centered. What assessment is being performed and why?

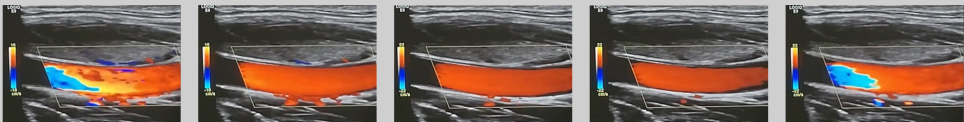
- Venous compression testing to assess patency/non-thrombosis
- Sonopalpation of a suspected soft-tissue abscess to look for internal swirling of debris with gentle compression.
- Internal jugular vein collapsibility assessment to estimate volume status by observing how easily the vein collapses with probe pressure.
- Venous reflux/valvular incompetence testing by applying and releasing compression while assessing for retrograde flow on Doppler.

GT: A

Answer: B. Explanation: The video demonstrates sonopalpation of a superficial, hypoechoic fluid collection (likely a soft tissue abscess). The operator gently compresses and releases the probe over the lesion. This maneuver is used to assess for squishiness or fluctuance and to visualize the swirling motion of internal echoes (pus/debris), which helps distinguish a fluid-filled abscess from a solid mass or cyst. The structure does not have the clear anatomical landmarks of a major vein like the IJV or femoral vein, and the compression is gentle and repetitive rather than a single firm collapse test."

Score: False

Figure A9: **Case Study 3 (MCQ, Type 1: Action–Goal Reasoning)**. Gemini 3 Pro misinterprets compression behavior due to incorrect anatomy identification, leading to the wrong action–goal selection.



Q: The clip shows a refinement of the color overlay to improve interpretability. What is changed, and why?


- Reduce color gain, to prevent color bleeding beyond the vessel
- Narrow the color box, to improve frame rate and spatial resolution of the flow overlay
- Increase color scale (PRF), to reduce aliasing and better display higher-velocity flow
- Raise the wall filter, to suppress low-frequency tissue motion (clutter) and clarify true vascular flow

GT: A

Answer: B. Explanation: Throughout the clip, the width of the color Doppler sampling box (the trapezoidal overlay) is significantly reduced. Narrowing the color box reduces the amount of data the ultrasound machine needs to process per frame, which directly improves the frame rate (temporal resolution) and can also help with the spatial resolution of the color flow overlay, making the hemodynamic display more responsive and accurate.

Score: False

Figure A10: **Case Study 4 (MCQ, Type 2: Artifact Resolution & Optimization)**. Gemini 3 Pro fails to identify the color bleeding artifact and the corresponding optimization cue.



Q:Based on the video clip, what phase of the exam is the operator performing across this segment?

- Performing dynamic compression to distinguish a vein from an adjacent artery and assess for thrombosis
- Transitioning into a long-axis view to align the probe for in-plane needle guidance once the vessel is found
- Confirming venous patency with color Doppler after identifying the target vessel in short axis
- Initial vessel localization followed by confirming depth/orientation using a posterior landmark"

GT: D

Answer: A. Explanation: The video clip demonstrates the operator performing a compression maneuver on the vascular structures. Initially, the vessels (likely the femoral vein and artery, or similar peripheral vessels) are seen in a short-axis view. The operator then applies downward pressure with the probe, causing the thinner-walled vein to collapse completely while the thicker-walled, pulsatile artery remains patent. This dynamic compression is the standard technique used to distinguish veins from arteries and to rule out deep vein thrombosis (DVT), as a thrombosed vein would not fully compress.

Figure A11: **Case Study 5: (MCQ, Type 3: Procedure Context & Planning)** fails to recognize the vertebral acoustic shadow as the key landmark for protocol context.