Multimodal Vision–Language Models for Automated FAST Ultrasound Instruction Generation

Ashley Anne Ho[1], Yuxiao Yang[2], Rebecca Tamrat Melaku[1], Anubhav Shankar[1], Sophia Yiwei Li[1], Annie J Bright[1], Jeremy Jose[1], Matthew Joseph Eckert[3], Arian Azarang[1], Yueh Z. Lee[1]

[1]*Department of Radiology, University of North Carolina School of Medicine, Chapel Hill, NC*

[2]*Department of Computer Science, University of North Carolina School of Medicine, Chapel Hill, NC*

[3]*Division of Acute Care & Trauma Surgery, University of North Carolina School of Medicine, Chapel Hill, NC*

UNC SCHOOL OF MEDICINE

Background

- Focused Assessment with Sonography in Trauma (FAST) ultrasound (US) is widely used to detect free fluid in trauma and critical care settings [1].
- Acquiring reliable FAST examinations requires extensive operator training, limiting use outside expert clinical environments.
- This training barrier restricts use in prehospital care, remote locations, disaster response, and home monitoring [2].
- Recent advances in artificial intelligence (AI) and multimodal large language models (MLLMs) enable innovative approaches for guided US acquisition [3].
- We developed a real-time multimodal AI coaching system to assist novice scanners in performing complete FAST examinations without expert supervision.

Methods

- Produced a synchronized multimodal acquisition pipeline integrating US video, expert audio guidance, and probe motion tracking, ***Figure 1***.

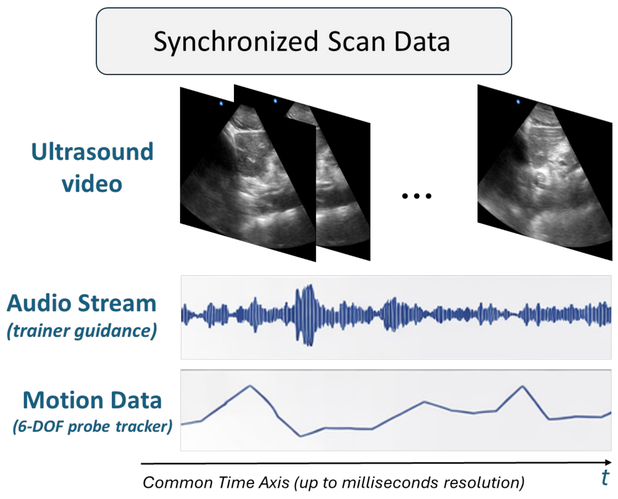


Figure 1. Collected data modalities

- Recruited novice US scanners with minimal or no prior US experience (approved IRB 24-2379).
- Three human subjects participated as standardized patients in this pilot study after providing written informed consent.
- Collected 46 FAST examination recordings including 23 right upper quadrant (RUQ) and 23 left upper quadrant (LUQ) scans following step-by-step predefined protocol.

Methods (Cont.)

- Temporally synchronized all modalities to align US frames, probe movements, and instructional guidance.
- Implemented zero-shot and LoRA fine-tuned frameworks to evaluate multimodal large language models for US instruction generation.
- Input RUQ and LUQ FAST US videos as inputs to GPT-4.1 and Qwen2.5-VL-7B (and 3B light version).
- Applied custom prompts to generate step-by-step scanning instructions from US video alone.
- Compared AI-generated instructions with expert ground-truth guidance using text embedding similarity metrics [4] and instruction-response annotation confirmed by abdominal radiologist (JJ).
- The LoRA fine-tuning stage was performed at the video-level and instruction-response segment-level.

Results

- LoRA-adapted Qwen achieved the highest instruction similarity scores across all test splits and acquisition systems (≈0.67–0.88), ***Figure 2(A)***.
- LoRA fine-tuning outperformed both GPT-4.1 (≈0.67–0.82) and zero-shot Qwen (≈0.38–0.66), ***Figure 2(A)***.
- Statistical analysis showed significant improvements for LoRA Qwen compared with GPT-4.1 ($t = -3.90$, $p = 0.0018$) and its base model.
- LoRA-adapted consistently improved instruction similarity at the segment-level (+0.1 for 3B, +0.2 for 7B; both statistically significant), ***Figure 2(B)***.
- The 7B model showed stronger and more consistent segment-level improvements than the 3B model.

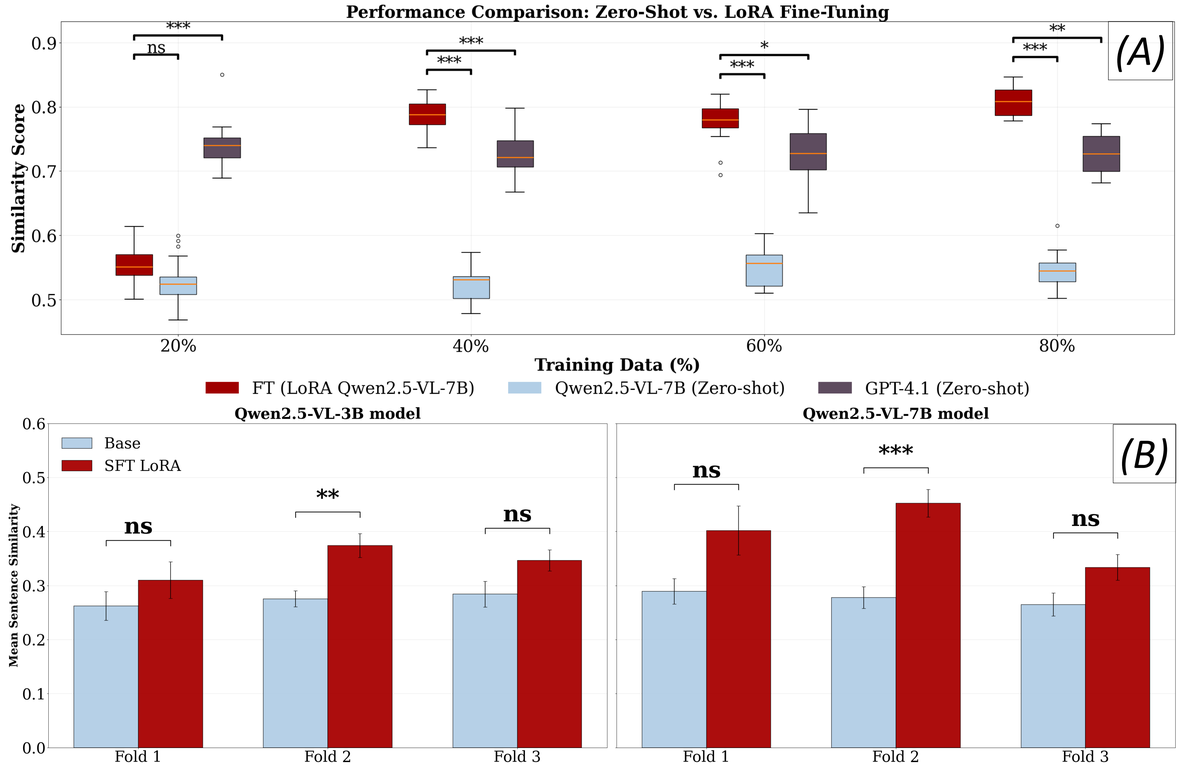
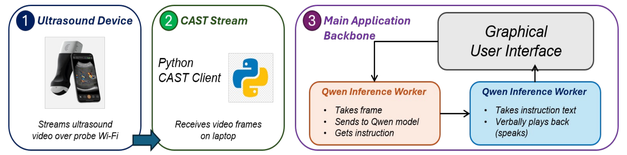


Figure 2. A) Video-level comparison for base and LoRA fine-tuned selected models. B) Bars show mean sentence similarity for Base and SFT LoRA within each cross-brand fold. Significance is based on a two-sided exact sign test over paired SFT – Base deltas.

Results (Cont.)



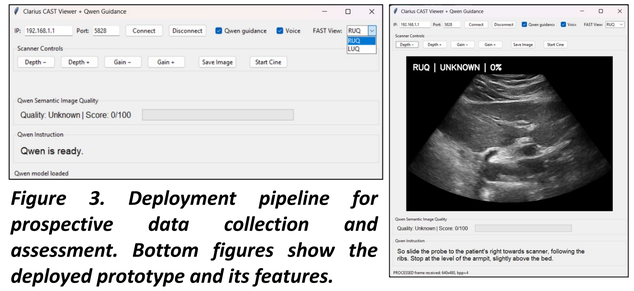


Figure 3. Deployment pipeline for prospective data collection and assessment. Bottom figures show the deployed prototype and its features.

Discussions and Conclusions

- Results indicate that multimodal large language models can interpret US videos and generate clinically meaningful scanning guidance.
- Domain-specific fine-tuning substantially improved instruction generation performance.
- Findings support the feasibility of real-time AI-assisted US coaching systems for non-expert users and a program was developed for future prospective data collection with our trained models, Figure 3.
- AI-guided US acquisition has the potential to improve US accessibility, user confidence, and timely clinical decision-making in settings with limited expert availability.

Acknowledgements

This work was supported by Civil Military Innovation Institute (CMI²) with award number W911QX23C0011-RFP-026 (PI: Yueh Z. Lee).

References

1. Arnold, M.J., Jonas, C.E. and Carter, R.E., 2020. Point-of-care ultrasonography. American family physician, 101(5), pp.275-285.
2. Anderson, A. and Theophanous, R.G., 2024. Point-of-care ultrasound use in austere environments: a scoping review. Plos one, 19(12), p.e0312017.
3. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., Zhong, H., Zhu, Y., Yang, M., Li, Z., Wan, J., Wang, P., Ding, W., Fu, Z., Xu, Y., … Lin, J. (2025). Qwen2.5-VL Technical Report. arXiv preprint arXiv:2502.13923.
4. Sentence-Transformers. (2021). all-MiniLM-L6-v2 [Machine learning model]. Hugging Face. https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2