

EchoMLLM: Incentivizing Echocardiographic Video Understanding with Keyframe Grounding and Report Generation

Heyu Huang^{1*}, Wanran Sun^{3*}, Chi Chen^{2†}, Bo Chen⁴, Zonghao Guo², Yuhua Li¹,
Ruixuan Li¹, Kunlun He⁴, Maosong Sun²

¹Huazhong University of Science and Technology, Wuhan, China

²Tsinghua University, Beijing, China

³Beijing University of Posts and Telecommunications, Beijing, China

⁴Chinese People’s Liberation Army General Hospital, Beijing, China

Abstract

Echocardiography analysis demands a dual capability: rigorous quantitative keyframe localization for evidence verification and comprehensive qualitative synthesis for diagnostic reporting. However, current Multi-Modal Large Language Models (MLLMs) struggle to meet these clinical requirements due to a misalignment with diagnostic workflows, a scarcity of video instruction data, and the critical challenge of cyclic temporal ambiguity—where the repetitive nature of cardiac cycles renders standard single-frame supervision ill-posed. To bridge this gap, we introduce EchoMLLM, a unified framework designed for real-world echocardiography video understanding. First, we align model capabilities with clinical needs by defining two fine-grained tasks: cycle- and pathology-conditioned keyframe grounding and video report generation. To facilitate this, we curate EchoMM-120k, a large-scale instruction dataset specifically constructed to support temporal localization and professional reporting. Furthermore, to resolve the cyclic ambiguity, we propose a multi-stage training paradigm incorporating a novel cycle-aware Reinforcement Learning (RL) strategy. By prioritizing logical consistency over rigid index matching, our approach moves beyond rote memorization to elicit invariant reasoning. Extensive experiments demonstrate that EchoMLLM reduces temporal grounding errors by up to 76% and improves report generation quality by 65% over its backbone, achieving state-of-the-art performance against both generalist and medical baselines.

1 Introduction

Echocardiography serves as a cornerstone in cardiology, capturing rapid temporal dynamics of the heart through ultrasound videos. Unlike general video understanding, clinical diagnosis imposes a

dual requirement: rigorous quantitative analysis and comprehensive qualitative synthesis. Specifically, clinicians must not only scrutinize the full video to generate diagnostic reports but also precisely isolate keyframes at specific cardiac phases (e.g., end-diastole) to serve as verifiable visual evidence. This standard procedure, as illustrated in Fig. 1(a), ensures diagnostic rigor but remains prohibitively labor-intensive, necessitating the development of efficient and automated frameworks.

Recently, Multimodal Large Language Models (MLLMs) have demonstrated substantial progress in general video understanding (Bai et al., 2025a; Yu et al., 2025; Li et al., 2025a; Luo et al., 2025). However, adaptation to the medical domain remains predominantly confined to static imaging. Current approaches primarily align 2D images with text (Li et al., 2023a; Chen et al., 2024; Team et al., 2025), neglecting critical temporal dynamics. While exceptions like Hulu-Med (Jiang et al., 2025) explore video understanding, they are limited to surgical contexts and focus on holistic VQA tasks. Such methods are incompatible with real-world echocardiography workflows. Consequently, the application of MLLMs to echocardiography video understanding remains underexplored.

Generally, the advancement of this field is impeded by three primary challenges: (i) **Task misalignment**: Existing VQA designs fail to mirror the real-world diagnostic workflow, preventing models from learning the coarse-to-fine association between temporal visual evidence and pathological semantics; (ii) **Data scarcity**: There is a lack of echocardiography video instruction corpora encompassing both fine-grained temporal localization and professional report generation; and (iii) **Cyclic Temporal Ambiguity**: The inherent repetitiveness of cardiac cycles creates substantial visual redundancy, where semantically identical phases (e.g., end-diastole) recur across the video. Under such conditions, the standard paradigm of learning from

*Equal contribution.

†Corresponding author: chenchithu@gmail.com.

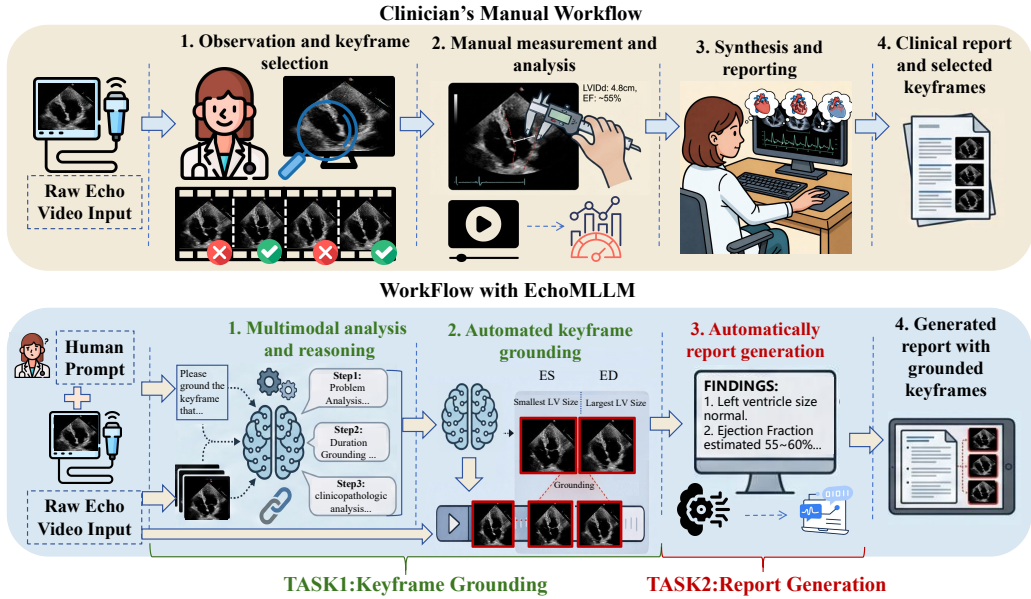


Figure 1: Illustration of how EchoMLLM aligns with the real-world clinical ultrasound diagnostic pipeline. It automates labor-intensive manual steps by unifying **keyframe grounding** and **report generation** to produce diagnostic reports with visual evidence.

a single deterministic ground truth becomes fundamentally ill-posed, leading to rote memorization rather than invariant reasoning.

To address these challenges, we first bridge the clinical gap by introducing two tasks aligned with real-world practice: cycle- and pathology-conditioned **keyframe grounding** and video **report generation**. As illustrated in Fig. 1, these tasks mirror the diagnostic pipeline of human experts, allowing seamless integration into routine clinical workflows. Second, to facilitate these tasks, we curate **EchoMM-120k**, a large-scale instruction dataset specifically constructed for fine-grained localization and report generation through rigorous synthesis and data-mining from diverse echocardiography sources. Finally, to conquer the cyclic temporal ambiguity, we devise a multi-stage training paradigm, combining multi-task Supervised Fine-Tuning (SFT) for both tasks with a specialized cycle-aware Reinforcement Learning (RL) strategy for keyframe grounding. By rewarding logical consistency over rigid index matching, this strategy incentivizes the model to learn invariant reasoning amidst visual redundancy, effectively resolving the ill-posed nature of single-frame supervision. The resulting model, **EchoMLLM**, demonstrates robust capabilities in these tasks, significantly outperforming both generalist and medical-specific MLLMs, paving the way for helpful and trustworthy automated cardiac diagnosis assistants.

Overall, our contributions are threefold:

- We align MLLMs with clinical echocardiographic diagnosis by introducing two fine-grained tasks: cycle- and pathology-conditioned keyframe grounding and video report generation, bridging the gap to real-world workflows.
- We mitigate data scarcity by curating EchoMM-120k, a large-scale instruction dataset specifically constructed to support both fine-grained temporal localization and professional report generation.
- We propose EchoMLLM, an MLLM for echocardiography video understanding. By integrating a cycle-aware temporal reward, it elicits explicit reasoning and achieves state-of-the-art performance on both keyframe grounding and report generation tasks.

2 Related Work

Medical Vision-Language Models. Most medical MLLMs focus on 2D image-text alignment (Li et al., 2023b; Chen et al., 2024) or general video QA (Jiang et al., 2025). In cardiology, existing frameworks (Ouyang et al., 2019; Duffy et al., 2022; Vukadinovic et al., 2024) generally rely on global classification or regression metrics. In contrast, our work shifts the paradigm from holistic prediction to fine-grained temporal reasoning, unifying keyframe grounding and report generation

within a cycle-conditioned framework.

RL-based Medical Reasoning. Recent works leverage RL (e.g., GRPO) to enhance clinical chain-of-thought capabilities (Pan et al., 2025; Lai et al., 2025; Team et al., 2025) or improve spatial visual grounding (Li et al., 2025b; Xu et al., 2025). However, applying RL to resolve *temporal* ambiguity remains unexplored. We fill this gap by proposing a cycle-aware RL strategy that rewards logical consistency amidst the repetitive visual redundancy of echocardiography.

3 Methodology

3.1 Task Definition

To bridge the gap between general video understanding and real-world echocardiographic diagnosis, we formulate two tasks that mirror the cognitive workflow of cardiac sonographers.

Cycle- and Pathology-Conditioned Keyframe Grounding (KG). Unlike standard video grounding which relies on generic text queries, this task explicitly tackles the cyclic temporal ambiguity inherent in echocardiography. Given an ultrasound video and a specific cycle or pathology query (e.g., "locate a representative end-diastolic frame"), the model is required to output *a coarse-to-fine triplet*: (1) the relevant temporal segment (Cycle-level), (2) the precise timestamp and frame index (Frame-level), and (3) the associated semantic justification (Evidence-level). This hierarchy forces the model to comprehend temporal dynamics throughout the cardiac cycle, capturing transient visual cues (e.g., valve closure, myocardial thickening) and aligning them with traceable visual evidence, rather than simply memorizing frame indices. This aligns with the routine of the expert to find a suitable frame that satisfies a specific need (e.g., measuring LVIDd).

Given the periodic nature of cardiac cycles, **multiple valid frames may satisfy the same query** (e.g., end-diastole in Cycle 1 vs. Cycle 2). Thus, the grounding objective is relaxed to identifying *one* representative instance. To resolve this inherent one-to-many mapping, we enforce a consistent cycle-aware paradigm across data construction, reward formulation, and evaluation protocols.

Video Report Generation (RG). Complementary to fine-grained grounding, this task demands a holistic synthesis of diagnostic information. Given an ultrasound video sequence, the model is tasked with generating a structured, clinically coherent

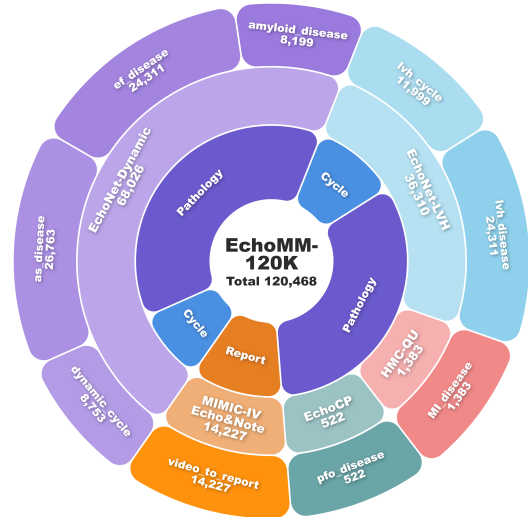


Figure 2: Task-wise data distribution in the proposed EchoMM-120K.

report. This requires the model to perform comprehensive diagnostic assessment, progressing from a holistic scan of the view to an in-depth analysis of cardiac structures, functions, and quantitative metrics. By enforcing the generation of key findings and conclusions, this task elicits the model’s ability to interpret complex pathological patterns and produce professional-grade diagnostic summaries. This emulates the final documentation phase, where the expert synthesizes scattered visual observations into a medical record for patient management. Further details are provided in the Appendix.

3.2 Construction of EchoMM-120k

The advancement of echocardiographic MLLMs is currently constrained by a critical scarcity of video instruction-following data, particularly for the proposed tasks of fine-grained grounding and professional report generation. Existing resources are predominantly limited to two forms: they either provide discrete labels for classification and regression (e.g., EchoNet-Dynamic (Ouyang et al., 2019)) or consist of large-scale, unannotated video archives (e.g., MIMIC-IV-ECHO (Gow et al., 2023)). Consequently, there is a lack of high-quality, multi-modal video instruction corpora required to train models for complex temporal understanding and clinically coherent generation.

To bridge this gap, we transform and synthesize these disparate resources to construct EchoMM-120K, a comprehensive training dataset illustrated in Fig. 2. For keyframe grounding, we harmonize four public datasets (EchoNet-LVH (Duffy et al.,

2022), EchoNet-Dynamic (Ouyang et al., 2019), EchoCP (Wang et al., 2021), HMC-QU (Degerli et al., 2024)) into a standardized instruction corpus containing 105K samples. Ground-truth labels are derived via rigorous physiological rules: for EchoNet-LVH, we define cycle-aware ED/ES keyframes based on LVID/wall-thickness extrema; for EchoNet-Dynamic, we utilize LV pixel area limits for ED/ES and stratify EF values to define pathology-aware frames (e.g., for CAD or aortic stenosis). to accommodate cardiac periodicity, we not only identify the most representative keyframe as the primary target but also explicitly annotate the corresponding frames in all other valid cycles as a candidate set. This multi-cycle annotation is essential for supporting our subsequent cycle-aware reward mechanism. Furthermore, we leverage grayscale intensity variations in EchoCP for PFO localization and expert segmentations in HMC-QU for myocardial infarction grounding. We implement a bi-directional instruction construction strategy, formatting samples as both "Query \rightarrow Grounding" and "Grounding \rightarrow Description" pairs to enhance model robustness and diversity.

For report generation, we curated 14,227 video-report pairs by aligning ultrasound videos from MIMIC-IV-ECHO with discharge summaries from MIMIC-IV-Note (Johnson et al., 2023). The alignment relies on Subject-ID and a 7-day temporal window, with redundant studies (within 2 days) filtered. We employ Qwen3-32B to extract clean, exam-relevant content from the raw discharge records. To expand coverage, we integrate MIMICEchoQA (Thapa et al., 2025) samples and utilize EchoPrime (Vukadinovic et al., 2024) to distill pseudo-reports for unlabelled videos, applying strict quality filtering. Finally, the processed reports are paired with multiple corresponding view videos to facilitate robust multi-view learning. For more details please refer to the Appendix A.

3.3 Stage 1: Synergistic Multi-Task SFT

We initiate the training with a Supervised Fine-Tuning (SFT) stage to equip the model with foundational capabilities for both proposed tasks. We posit that a synergistic relationship exists between the two proposed tasks: keyframes provide implicit visual anchors that ground the generation process, while professional reports offer explicit semantic details that enrich the feature for grounding.

To exploit this synergy, we construct a mixed

dataset comprising 85K keyframe grounding samples and 14K report generation samples. We optimize the model using the standard auto-regressive cross-entropy loss. Upon completion of this stage, the model achieves a baseline proficiency, capable of generating coherent clinical reports and performing basic keyframe localization based on visual features. This establishes a robust starting point for subsequent reasoning optimization.

3.4 Stage 2: Cycle-Aware Reinforcement Learning

Although the SFT model is functional, its fine-grained grounding capability is constrained by the cyclic temporal ambiguity inherent in echocardiography. Due to the high redundancy between consecutive frames, the SFT model often relies on rote memorization of frame indices rather than learning invariant physiological features, leading to instability in complex cyclic scenarios. To further elevate the model’s performance and elicit explicit reasoning, we introduce a second stage utilizing Group Relative Policy Optimization (GRPO) (Shao et al., 2024). This stage specifically targets the keyframe grounding task to resolve subtle visual ambiguities.

CoT-based Initialization. Prior to RL, we perform a "cold-start" using 3,500 high-quality samples derived from the SFT set. We employ the advanced Qwen2.5-VL-72B to distill Chain-of-Thought (CoT) rationales, guiding the model to generate step-by-step analysis (e.g., "*The valve is starting to open...*") before outputting the final timestamp. This initialization primes the model for reasoning-based optimization.

GRPO Optimization Objective. We employ GRPO to estimate gradients based on group-relative advantages, which avoids the need for a separate value network and reduces computational overhead. Formally, for each query q , we sample a group of outputs $\{o_1, \dots, o_G\}$ from the old policy $\pi_{\theta_{\text{old}}}$. The objective function is formulated as:

$$\mathcal{J}_{\text{GRPO}}(\theta) = \mathbb{E}_{q \sim P(Q), \{o_i\} \sim \pi_{\theta_{\text{old}}}} \left[\frac{1}{G} \sum_{i=1}^G \left(\frac{\pi_{\theta}(o_i | q)}{\pi_{\theta_{\text{old}}}(o_i | q)} A_i - \beta \mathbb{D}_{\text{KL}}(\pi_{\theta} \| \pi_{\text{ref}}) \right) \right] \quad (1)$$

where β is the KL-divergence penalty coefficient to prevent reward hacking. A_i is the advantage

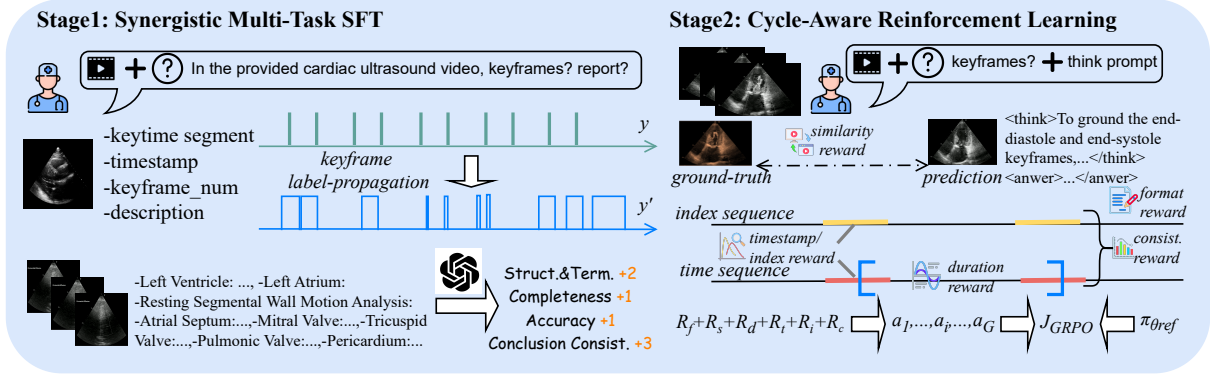


Figure 3: Overview of the proposed EchoMLLM. In stage-1, supervised fine-tuning is performed for keyframe grounding and report generation, enabling initial temporal understanding and perception of ultrasound videos. In stage-2, keyframe reasoning grounding is further optimized using GRPO with a well-designed cycle-aware reward mechanism, improving pathology-conditioned reasoning and refining keyframe-specific features.

standardized from the group rewards $\{r_1, \dots, r_G\}$:

$$A_i = \frac{r_i - \text{mean}(r_1, \dots, r_G)}{\text{std}(r_1, \dots, r_G) + \epsilon} \quad (2)$$

This relative advantage mechanism encourages the model to prioritize answers that are relatively better within the sampled group, effectively guiding it through the ambiguity of visually similar frames. To operationalize this optimization, we formulate a cycle-aware reward mechanism to serve as the evaluation standard for these relative comparisons.

3.5 Cycle-Aware Reward Mechanism

To provide precise guidance signals for the GRPO algorithm, we design a composite reward function specifically tailored for fine-grained echocardiographic grounding. It comprises four components:

Format Reward (R_{fmt}). To enforce structured reasoning, a reward (e.g., 0.25) is assigned only if the output follows the strict format: reasoning process enclosed in `<think>...</think>` tags followed by the answer in `<answer>...</answer>` tags.

Soft-Margin Temporal Precision Reward. Standard exact-match supervision is overly harsh for high-frame-rate videos. We propose a soft-margin mechanism encompassing Duration IoU (R_{IoU}), Timestamp Accuracy (R_t), and Index Accuracy (R_i). For R_{IoU} , we apply a stratified thresholding function with $\tau_h = 0.9$ and $\tau_l = 0.1$ on the IoU between the predicted duration \tilde{B} and

the ground-truth duration B :

$$R_{IoU} = \begin{cases} 1, & \text{if } \text{IoU}(B, \tilde{B}) \geq \tau_h \\ \text{IoU}(B, \tilde{B}), & \text{if } \tau_l < \text{IoU}(B, \tilde{B}) < \tau_h \\ 0, & \text{if } \text{IoU}(B, \tilde{B}) \leq \tau_l \end{cases} \quad (3)$$

For R_t and R_i , to handle temporal redundancy, we define a tolerance radius r (one-sixth of the cycle span). We assign stratified rewards inversely proportional to the normalized distance d_r between the prediction and valid candidate anchors:

$$R_i/R_t = \begin{cases} 1.0, & \text{if } d_r \leq 0.33 \\ 0.6, & \text{if } 0.33 < d_r \leq 0.67 \\ 0.2, & \text{if } 0.67 < d_r < 1.0 \\ 0, & \text{otherwise} \end{cases} \quad (4)$$

where $d_r = \min_{a \in \mathcal{A}} |p - a|/r$, with p being the prediction and \mathcal{A} the set of candidate frames.

Clinical Similarity Reward (R_{sim}). To ensure physiological fidelity, we employ an external specialist model (Duffy et al., 2022; Ouyang et al., 2019) to extract clinical metrics (LVID, IVS, LVPW, LV Area) from the predicted frame \tilde{f} and the ground-truth f . Implementation details are provided in Appendix B.2. The reward is penalized by the relative error vector E : $R_{sim} = 1 - \|E(f, \tilde{f})\|_1$.

Consistency Reward (R_{con}). To mitigate internal hallucinations where predicted attributes contradict each other, we evaluate the logical coherence of the output triplet. We assign a bonus of 0.5 for each mutually consistent pair among the predicted duration, timestamp, and frame index.

Total Reward. The final optimization objective is the weighted summation of all components:

$$R_{\text{total}} = R_{\text{fmt}} + R_{\text{IoU}} + R_t + R_i + \alpha R_{\text{sim}} + R_{\text{con}} \quad (5)$$

where α is a balancing coefficient empirically set to 3 to strictly enforce clinical feature alignment.

4 Experiments

4.1 Implementation

We build EchoMLLM upon Qwen2.5-VL-3B (Bai et al., 2025b). All experiments are conducted using 8 NVIDIA A100 GPUs with 80GB of memory. All random seeds are fixed during the training procedures to ensure reproducibility. For the SFT stage, we adopt a learning rate of 5e-6 with a global accumulation batch size of 48. For the CoT-SFT and reinforcement learning stages, we employ a learning rate of 5e-6 and 3e-6 with accumulation batch sizes of 48 and 32, respectively. The GRPO algorithm in the reinforcement post-training stage is configured with a maximum completion length of 1024 tokens and samples 8 responses per input. Across both training stages and task types, the sampling frame rate is set to 150 (exceeding the maximum frame rate of the actual video data), and the maximum resolution parameter is configured at 139,552.

4.2 Evaluation Setup

For Keyframe Grounding, we evaluate on the standard test splits of the respective source datasets, totaling 11,847 samples. For Report Generation, we curate a held-out set of 350 randomly sampled examples. Note that all test instances are processed using the exact same pipeline as EchoMM-120k.

We employ the following metrics with **full definitions** detailed in Appendix B.

Keyframe Grounding. We evaluate this task from two complementary dimensions: index error and clinical similarity. **(1) Index Error (Idx).** We compute the Mean Absolute Error (MAE) between the predicted frame index and the *nearest valid candidate frame* in the ground truth, adopting a minimum distance principle to account for cardiac periodicity. **(2) Clinical Similarity.** Since temporal distance does not always correlate with diagnostic accuracy, we further evaluate the functional similarity between the predicted and ground-truth frames. We adopt the most appropriate metric for different task by computing the Relative Metric Error for geometric indicators like Left Ventricular Area (**Area**), Left Ventricle Internal Diameter (**LVID**), Ejection

Fraction (**EF**), pixel intensity (**PI**) and visual cosine similarity (**Sim**) for texture-based pathologies (e.g., Myocardial Infarction).

Report Generation. Given the limitations of traditional n -gram metrics in medical contexts, we adopt an LLM-as-a-judge paradigm. We employ GPT-4 to evaluate generated reports on a 0–5 scale across four clinical dimensions: (1) *Structure & Terminology*, (2) *Completeness*, (3) *Accuracy*, and (4) *Conclusion Consistency*.

We compare the proposed EchoMLLM against state-of-the-art medical MLLMs including Lingshu-3B/7B (Team et al., 2025) and Hulu-Med-3B/7B (Jiang et al., 2025), advanced general-purpose MLLMs such as Qwen2.5-VL-3B and Qwen3-VL-4B (Bai et al., 2025a), as well as proprietary models GPT-4o and Gemini-2.5-Pro.

4.3 Main Results

Quantitative Results on Keyframe Grounding. As reported in Table 1, EchoMLLM establishes a new state-of-the-art across both cycle- and pathology-conditioned benchmarks. Most notably, compared to its backbone Qwen2.5-VL-3B, our model reduces frame index errors by approximately 76% on the LVH-Cycle task (from 17.35 to 4.18) and 73% on EF estimation (from 20.60 to 5.48), effectively resolving the invalid (“N/A”) predictions caused by cycle inconsistency. EchoMLLM also demonstrates marked superiority over GPT-4o in complex scenarios (LVH: 4.18 vs. 7.41) and achieves the lowest errors on clinical metrics (e.g., Amyloid Area error of 0.11 vs. baselines >0.15 ; PFO Intensity error of 1.96 vs. >2.24). Furthermore, the RL stage proves essential for precision, reducing index errors by 8.9% and clinical metric errors by up to 23.5% compared to the SFT baseline, validating the effectiveness of our reward-driven optimization.

Quantitative Results on Report Generation. As shown in Table 2, EchoMLLM achieves best-in-class performance with an average score of 2.56, ranking first across all four evaluation dimensions. In direct comparison with the generalist Qwen2.5-VL-3B (Avg: 1.55), EchoMLLM delivers a substantial performance leap of +1.01 points representing a 65% relative improvement. Moreover, EchoMLLM outperforms the strongest proprietary baseline, Gemini-2.5 Pro (Avg: 2.00), by +0.56 points ($\approx 28\%$), and surpasses the top open-source medical model, Lingshu-32B (Avg: 1.85), by +0.71

Table 1: **Keyframe grounding results.** We report the Frame Index Error (Idx) and Clinical Similarity metrics across Cycle- and Pathology-conditioned benchmarks. EchoMLLM significantly outperforms both proprietary and open-source models. Note that for all metrics, lower values indicate better N/A: invalid EF calculations caused by predicting paired frames across inconsistent cardiac cycles.

Methods	Idx	Area	Idx	LVID	Idx	Area	EF	Idx	Area	Idx	Area	Idx	Sim	Idx	Area	Int	Idx	LVID
	Cycle-conditioned				Pathology-conditioned													
	Dy-cycle		LVH-cycle		EF			AS		Amyloid		MI		PFO			LVH	
<i>API-based</i>																		
GPT-4o	5.11	0.18	7.41	0.13	5.94	0.15	0.05	5.64	0.16	5.37	0.15	19.21	0.25	67.89	0.36	2.66	4.38	0.11
<i>Open-source</i>																		
Qwen2.5-VL-3B	14.51	0.19	17.35	0.14	20.60	0.07	N/A	14.78	0.19	15.42	0.19	7.38	0.12	125.61	0.54	3.57	19.73	0.13
Qwen2.5-VL-7B	14.96	0.19	17.08	0.13	20.54	0.07	N/A	14.94	0.19	15.41	0.19	7.38	0.12	95.96	0.39	3.57	19.61	0.13
Qwen3-VL-4B	16.26	0.19	18.86	0.13	28.82	0.09	N/A	15.37	0.08	12.12	0.18	25.90	0.82	91.13	0.45	2.24	28.19	0.16
Hulu-Med-4B	13.95	0.19	16.34	0.13	20.59	0.07	N/A	14.91	0.19	15.34	0.19	7.38	0.12	95.85	0.38	3.77	19.55	0.13
Hulu-Med-7B	14.21	0.19	16.60	0.13	20.59	0.07	N/A	14.94	0.19	15.40	0.19	7.38	0.12	95.96	0.39	3.57	19.62	0.13
Hulu-Med-32B	14.51	0.19	16.62	0.13	20.59	0.07	N/A	14.94	0.19	15.41	0.19	6.79	0.11	95.96	0.39	3.57	19.62	0.13
Lingshu-7B	14.84	0.19	17.04	0.13	20.59	0.07	N/A	15.05	0.19	15.44	0.19	7.15	0.12	95.96	0.39	3.31	19.56	0.13
EchoMLLM (SFT)	6.73	0.14	6.52	0.10	4.52	0.14	0.07	3.22	0.15	4.27	0.14	2.38	0.11	83.85	0.31	4.82	3.85	0.08
EchoMLLM (SFT+RL)	5.14	0.12	4.18	0.10	5.48	0.10	0.07	4.67	0.11	4.29	0.11	2.55	0.06	64.14	0.25	1.96	4.63	0.08

Table 2: **Report generation results.** Metrics include **Fmt.** (Structure & Terminology), **Comp.** (Completeness), **Acc.** (Accuracy), **Consis.** (Conclusion Consistency), and **Avg.** (Average score).

Models	Fmt.	Comp.	Acc.	Consis.	Avg.
<i>API-based</i>					
GPT-4o	1.98	0.39	0.22	0.18	0.69
GPT-4-turbo	2.34	0.57	0.26	0.21	0.85
Gemini2.5-pro	3.55	1.53	1.58	1.42	2.02
<i>Open-source</i>					
Qwen2.5-VL-3B	2.60	1.27	1.17	1.14	1.55
Qwen3-VL-4B	1.92	0.94	0.89	0.80	1.14
Lingshu-7B	3.07	1.32	1.21	1.19	1.70
Lingshu-32B	3.29	1.54	1.32	1.27	1.86
Hulu-Med-7B	2.23	0.79	0.97	0.85	1.21
Hulu-Med-32B	2.82	1.20	1.09	1.02	1.53
EchoMLLM	4.07	2.50	1.82	1.84	2.56

points ($\approx 38\%$). These consistent gains verify that our cycle-aware training paradigm enables the generation of professionally structured, accurate, and logically consistent diagnostic reports. We further strengthen the reliability analysis of the report generation by computing fine-grained max / mean / standard deviation statistics across repeated 3 runs, thereby providing a more precise estimate of evaluation stability and reducing the possibility that the observed performance differences arise from

incidental variation or scoring randomness. The resulting detailed results in Table 6 of Appendix B, where the consistently small deviations further support the robustness and repeatability of our conclusions.

Human Evaluation. To further validate the clinical reliability of EchoMLLM beyond automatic metrics, we conduct a double-blinded evaluation with senior cardiologists. As detailed in Table 9 and Table 10, EchoMLLM achieves consistently higher clinician preference in keyframe grounding, reaching a win rate (WR) of $\sim 65\%$ and an adjusted win rate (AWR) of up to 85.12%.

For report generation, EchoMLLM substantially outperforms all baselines across all evaluation dimensions, achieving an average score of 4.31/5, significantly higher than both open-source and proprietary models. These results further confirm that our method not only improves quantitative performance but also yields clinically preferred and higher-quality outputs in real-world settings.

4.4 Ablation

We analyze the impact of different training stages in Table 3. While SFT and CoT-SFT establish a solid foundation for perceptual alignment and explicit reasoning, the RL stage proves critical for fine-grained refinement. By leveraging coarse-to-fine reasoning patterns, the RL stage consistently reduces errors across both cycle-conditioned and

Table 3: Ablation Studies of the proposed EchoMLLM.

Methods	Idx	Area	Idx	LVID	Idx	Area	EF	Idx	Area	Idx	Area	Idx	Sim	Idx	Area	Int	Idx	LVID
	Cycle-conditioned				Pathology-conditioned													
	Dy Cycle		LVH Cycle		EF			AS		Amyloid		MI		PFO			LVH	
<i>Specific Training Stages</i>																		
EchoMLLM (SFT)	6.73	0.14	6.52	0.10	4.52	0.14	0.07	3.22	0.15	4.27	0.14	2.38	0.11	83.85	0.31	4.82	3.85	0.08
EchoMLLM (CoT-SFT)	5.11	0.18	7.41	0.13	5.94	0.15	0.05	5.64	0.16	5.37	0.15	19.21	0.25	67.89	0.36	2.66	4.38	0.11
EchoMLLM (SFT+RL)	5.14	0.12	4.18	0.10	5.48	0.10	0.07	4.67	0.11	4.29	0.11	2.55	0.06	64.14	0.25	1.96	4.63	0.08
<i>Specific RL Configs</i>																		
EchoMLLM-e2	5.82	0.11	5.47	0.08	7.72	0.10	0.10	5.33	0.11	6.43	0.10	3.55	0.06	108.53	0.25	1.55	6.29	0.08
EchoMLLM-derup	5.88	0.11	5.57	0.08	7.90	0.10	0.10	5.38	0.11	6.49	0.10	14.11	0.20	110.80	0.31	2.23	6.36	0.08
EchoMLLM-term-only	20.17	0.15	10.08	0.11	6.35	0.10	N/A	15.37	0.08	12.12	0.18	25.90	0.82	91.13	0.45	2.24	5.57	0.09
EchoMLLM-eq	12.43	0.15	11.09	0.10	17.39	0.09	N/A	13.49	0.15	15.90	0.15	9.05	0.53	95.85	0.38	3.77	12.18	0.11
EchoMLLM-consist	6.89	0.11	5.86	0.08	8.59	0.10	0.11	6.95	0.11	6.71	0.11	5.09	0.30	126.55	0.31	3.19	7.20	0.08
EchoMLLM-th18	6.08	0.12	6.69	0.09	7.55	0.10	0.11	6.23	0.11	7.47	0.11	10.63	0.11	94.93	0.34	4.61	6.69	0.09
EchoMLLM-th14	7.31	0.11	5.77	0.08	7.25	0.10	N/A	7.08	0.11	6.89	0.11	7.71	0.08	123.42	0.30	4.76	5.77	0.08
EchoMLLM	5.14	0.12	4.18	0.10	5.48	0.10	0.07	4.67	0.11	4.29	0.11	2.55	0.06	64.14	0.25	1.96	4.63	0.08

pathology-conditioned tasks. For example, on the LVH Cycle task, the frame index error decreases from 6.52 (SFT) and 7.41 (CoT-SFT) to 4.18 with RL. Similarly, clinical metrics are improved, with Area error reduced to 0.11 and PFO Intensity significantly decreasing from 4.82 to 1.96. These improvements validate the effectiveness of our reward-driven optimization strategy in enhancing physiological fidelity alongside temporal precision.

Furthermore, we investigate specific RL configurations to address the challenge of "reward hacking" induced by cardiac periodicity. We observe that over-optimizing for temporal precision without adequate constraints leads to model collapse. Specifically, using a purely temporal reward ("-term_only") causes the frame index error to spike dramatically (e.g., Dy Cycle: 20.17 vs. 5.14 in the full model), as the model exploits cyclic ambiguity. Similarly, naively extending training epochs ("-e2") or reducing dataset size ("-derup") leads to degraded performance across multiple tasks, indicating overfitting or insufficient regularization. Our final configuration, which balances cycle-aware temporal rewards with clinical similarity constraints, successfully mitigates this hacking phenomenon and achieves the best overall performance.

4.5 Case Study

We qualitatively compare temporal grounding results on EchoNet-LVH `lvh_cycle` in Fig. 4, where the goal is to identify end-diastole (ED) and

end-systole (ES) keyframes. We report predictions from Qwen2.5-VL-3B, our SFT model, and the RL-enhanced model against the Ground Truth.

Cardiac periodicity: SFT vs. Baseline. The Qwen2.5-VL-3B baseline predicts both ED and ES near the beginning of the sequence. While the ED prediction coincides with the GT in this case, it reflects a bias toward early frames rather than modeling cardiac periodicity. In contrast, the SFT model selects ED/ES frames from later cycles that exhibit phase-consistent morphology, indicating an improved understanding of the cardiac motion.

Reasoning-aware grounding: RL vs. SFT. The RL-enhanced model further refines temporal grounding by reasoning over the cardiac cycle. It first identifies a plausible temporal segment and then selects ED and ES at frames where ventricular dimensions reach their extrema. Compared to SFT, the predicted ED frame shows a larger left ventricular internal diameter (LVID), and the predicted ES frame corresponds to a more contracted ventricle. This leads to better alignment between temporal localization and clinically meaningful phase definitions, illustrating how RL reduces phase ambiguity beyond supervised training.

5 Conclusion

In this work, we present **EchoMLLM**, a unified framework designed to automate the clinical echocardiography workflow by integrating fine-grained keyframe grounding with professional re-

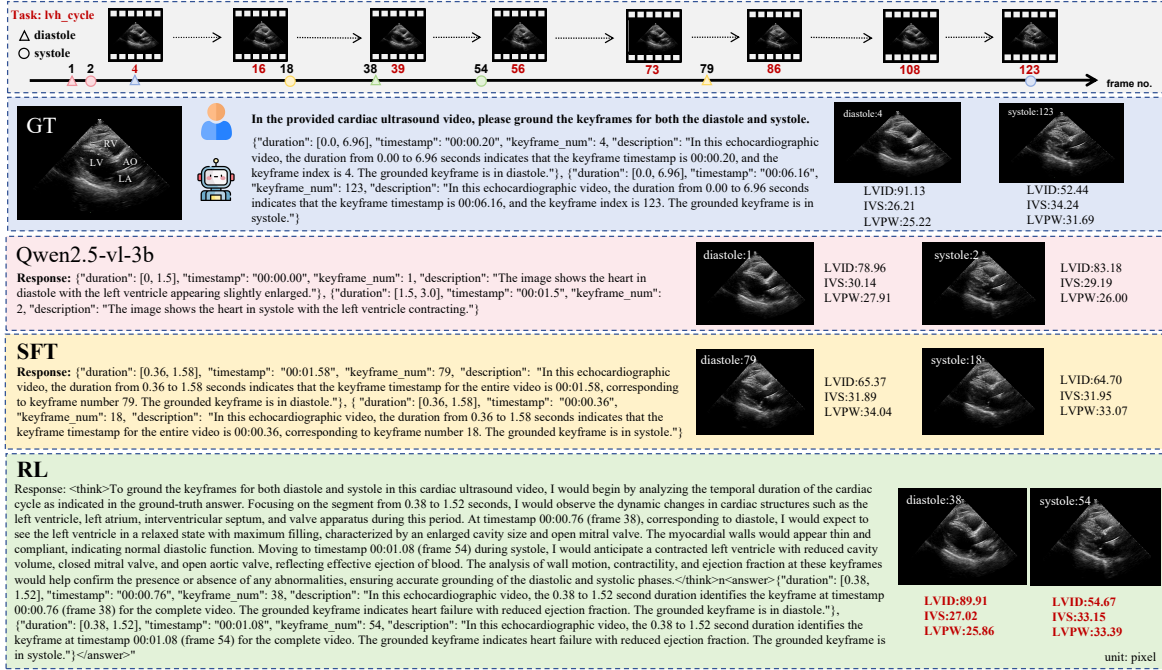


Figure 4: Qualitative comparison on EchoNet-LVH. We compare the Ground Truth (GT) with predictions from baseline, our SFT and RL-enhanced model. EchoMLLM leverages cycle-aware reasoning to identify a temporally consistent keyframe pair, resulting in physiologically accurate clinical measurements (LVID, IVS, LVPW).

port generation. To address the data scarcity in this domain, we curate **EchoMM-120k**, a large-scale, cycle-annotated instruction dataset. Technically, we propose a novel multi-stage training paradigm that incorporates a cycle-aware reinforcement learning strategy. This approach effectively resolves the temporal ambiguity inherent in periodic cardiac monitoring and enforces physiological fidelity through clinical similarity rewards. Extensive experiments demonstrate that EchoMLLM significantly outperforms state-of-the-art proprietary and open-source models, establishing a new benchmark for interpretable and accurate cardiac video understanding. We believe this work serves as a foundational step toward reliable, automated diagnostic assistants in cardiology.

Acknowledgement

The work is initiated and supported by AI9Stars Team. We thank Xiazhen Qin for his valuable discussions and suggestions.

Limitation

Despite the strong empirical performance of EchoMLLM, a main limitation arises from the scarcity of triplet-aligned echocardiography data (video, keyframes, and reports) due to strict clinical privacy regulations. Consequently, EchoMLLM re-

lies on a task-decomposed formulation, utilizing multi-task learning rather than fully unified end-to-end supervision. While this approach yields strong performance by mimicking clinical workflows, it remains a multi-stage pipeline potentially subject to inter-task error propagation. Future work aims to curate more comprehensive datasets where keyframes are explicitly linked to report sections. This would enable unified supervision, facilitating a truly end-to-end system and further reducing cascading errors for robust clinical deployment.

Ethics Statement

This work involves the development of a Multi-Modal Large Language Model (MLLM) for echocardiography analysis. We strictly adhere to the ACL Ethics Policy and address the potential ethical implications as follows:

Data Privacy and Usage. Our model is trained and evaluated on publicly available, de-identified datasets, including MIMIC-IV-ECHO, EchoNet-Dynamic, EchoNet-LVH, HMC-QU, and EchoCP. Access to the MIMIC-IV suite was obtained via PhysioNet under a signed Data Use Agreement (DUA) after completing the required CITI human subjects research training. We have verified that all utilized data sources strictly follow regulations

or their respective ethical guidelines regarding patient anonymity. No private or identifiable patient information was collected or processed during this research.

Clinical Safety and Deployment. While EchoM-LLM demonstrates state-of-the-art performance in keyframe grounding and report generation, it is intended solely as a research prototype and an assistive tool for clinicians, not as a standalone diagnostic decision-maker. Echocardiography interpretation requires expert judgment, and AI-generated reports may contain hallucinations or factual errors, although we have mitigated this via our cycle-aware reward mechanism. We explicitly warn against the deployment of this system in real-world clinical settings without rigorous, prospective clinical validation and human-in-the-loop oversight.

References

- Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibao Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. 2025a. [Qwen3-vl technical report](#).
- Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. 2025b. [Qwen2.5-vl technical report](#). *arXiv preprint arXiv:2502.13923*.
- Junying Chen, Chi Gui, Ruyi Ouyang, Anningzhe Gao, Shunian Chen, Guiming Hardy Chen, Xidong Wang, Ruifei Zhang, Zhenyang Cai, Ke Ji, Guangjun Yu, Xiang Wan, and Benyou Wang. 2024. [Huatuogpt-vision, towards injecting medical visual knowledge into multimodal llms at scale](#).
- Aysen Degerli, Serkan Kiranyaz, Tahir Hamid, Rashid Mazhar, and Moncef Gabbouj. 2024. [Early myocardial infarction detection over multi-view echocardiography](#). *Biomedical Signal Processing and Control*, 87:105448.
- Garrett Duffy, Patrick P. Cheng, Nan Yuan, et al. 2022. [High-throughput precision phenotyping of left ventricular hypertrophy with cardiovascular deep learning](#). *JAMA Cardiology*, 7(4):386–395.
- Brian Gow, Tom Pollard, Nathaniel Greenbaum, Benjamin Moody, Alistair Johnson, Elizabeth Herbst, Jonathan W. Waks, Parastou Eslami, Ashish Chaudhari, Tanner Carbonati, Seth Berkowitz, Roger Mark, and Steven Horng. 2023. [MIMIC-IV-ECHO: Echocardiogram matched subset](#).
- Songtao Jiang, Yuan Wang, Sibao Song, Tianxiang Hu, Chenyi Zhou, Bin Pu, Yan Zhang, Zhibo Yang, Yang Feng, Joey Tianyi Zhou, Jin Hao, Zijian Chen, Ruijia Wu, Tao Tang, Junhui Lv, Hongxia Xu, Hongwei Wang, Jun Xiao, Bin Feng, Fudong Zhu, Kenli Li, Weidi Xie, Jimeng Sun, Jian Wu, and Zuozhu Liu. 2025. [Hulu-med: A transparent generalist model towards holistic medical vision-language understanding](#).
- Alistair Johnson, Tom Pollard, Steven Horng, Leo Anthony Celi, and Roger Mark. 2023. [MIMIC-IV-Note: De-identified free-text clinical notes](#).
- Yuxiang Lai, Jike Zhong, Ming Li, Shitian Zhao, Yuheng Li, Konstantinos Psounis, and Xiaofeng Yang. 2025. [Med-r1: Reinforcement learning for generalizable medical reasoning in vision-language models](#).
- Chunyu Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. 2023a. [Llava-med: Training a large language-and-vision assistant for biomedicine in one day](#). *Advances in Neural Information Processing Systems*, 36:28541–28564.
- Chunyu Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. 2023b. [Llava-med: Training a large language-and-vision assistant for biomedicine in one day](#).
- KunChang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. 2025a. [Videochat: Chat-centric video understanding](#). *Science China Information Sciences*, 68(10):200102.
- Yifan Li, Fenghe Tang, Yingtai Li, and Shaohua Kevin Zhou. 2025b. [Medreason-r1: Learning to reason for ct diagnosis with reinforcement learning and local zoom](#).
- Fuwen Luo, Shengfeng Lou, Chi Chen, Ziyue Wang, Chenliang Li, Weizhou Shen, Jiyue Guo, Peng Li, Ming Yan, Ji Zhang, Fei Huang, and Yang Liu. 2025. [Museg: Reinforcing video temporal understanding via timestamp-aware multi-segment grounding](#).
- David Ouyang, Bryan He, Amirata Ghorbani, Matthew P. Lungren, Euan A. Ashley, David H. Liang, and James Y. Zou. 2019. [Echonet-dynamic: A large new cardiac motion video data resource for medical machine learning](#). In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 1–10.

Jiazhen Pan, Che Liu, Junde Wu, Fenglin Liu, Jiayuan Zhu, Hongwei Bran Li, Chen Chen, Cheng Ouyang, and Daniel Rueckert. 2025. [Medvlm-r1: Incentivizing medical reasoning capability of vision-language models \(vlms\) via reinforcement learning.](#)

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. 2024. [Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300.*](#)

LASA Team, Weiwen Xu, Hou Pong Chan, Long Li, Mahani Aljunied, Ruifeng Yuan, Jianyu Wang, Chenghao Xiao, Guizhen Chen, Chaoqun Liu, Zhaodonghui Li, Yu Sun, Junao Shen, Chaojun Wang, Jie Tan, Deli Zhao, Tingyang Xu, Hao Zhang, and Yu Rong. 2025. [Lingshu: A generalist foundation model for unified multimodal medical understanding and reasoning.](#)

Rahul Thapa, Andrew Li, Qingyang Wu, Bryan He, Yuki Sahashi, Christina Binder-Rodriguez, Angela Zhang, David Ouyang, and James Zou. 2025. [Mimiciv-echo-ext-mimicechoqa: A benchmark dataset for echocardiogram-based visual question answering.](#)

Milos Vukadinovic, Xiu Tang, Neal Yuan, Paul Cheng, Debiao Li, Susan Cheng, Bryan He, and David Ouyang. 2024. [Echoprime: A multi-video view-informed vision-language model for comprehensive echocardiography interpretation.](#)

Tianchen Wang, Zhihe Li, Meiping Huang, Jian Zhuang, Shanshan Bi, Jiawei Zhang, Yiyu Shi, Hongwen Fei, and Xiaowei Xu. 2021. [Echocp: An echocardiography dataset in contrast transthoracic echocardiography for patent foramen ovale diagnosis.](#)

Huihui Xu, Yuanpeng Nie, Hualiang Wang, Ying Chen, Wei Li, Junzhi Ning, Lihao Liu, Hongqiu Wang, Lei Zhu, Jiyao Liu, Xiaomeng Li, and Junjun He. 2025. [Medground-r1: Advancing medical image grounding via spatial-semantic rewarded group relative policy optimization.](#)

Tianyu Yu, Zefan Wang, Chongyi Wang, Fuwei Huang, Wenshuo Ma, Zhihui He, Tianchi Cai, Weize Chen, Yuxiang Huang, Yuanqian Zhao, et al. 2025. [Mimicpm-v 4.5: Cooking efficient mllms via architecture, data, and training recipe. *arXiv preprint arXiv:2509.18154.*](#)

A Detailed Data Construction Pipeline

The construction of the **EchoMM-120K** dataset follows a dual-stream pipeline designed to bridge raw multimodal medical resources with clinical expert-level understanding. The process is bifurcated into (1) Fine-grained Keyframe Grounding (KG) and (2) Temporal-Matched Report Generation (RG).

A.1 Fine-grained Keyframe Grounding (KG) Data Construction

The KG stream transforms video-level labels into frame-level reasoning instructions by leveraging physiological priors and specialist models across four primary sources.

- **EchoNet-LVH (PLAX View):** We merged the training and validation sets (11,657/343 samples). Since official labels only cover partial frames, we utilized open-source regression weights to predict structural indicators—LVID, IVS, and LVPW—frame-by-frame. End-Diastole (ED) and End-Systole (ES) ground truths were defined by global pixel area extrema and further refined using a *Rank-Sum filtering algorithm* (sorting ES by minimal LVID and ED by maximal LVID) to select the most representative cardiac cycle. Relative Wall Thickness (RWT) was calculated as $RWT = \frac{2 \times PWT_d}{LVID_d}$ to categorize four LV geometric patterns.
- **EchoNet-Dynamic (A4C View):** Left ventricular pixel area (size) serves as the core temporal indicator. We implemented beat-level Ejection Fraction (EF) estimation by aggregating results from multiple heartbeats to improve stability. For Aortic Stenosis (AS) and Amyloidosis, we utilized pre-trained models to obtain video-level confidence scores, aligning these with the best cardiac cycle identified via the Rank-Sum algorithm.
- **EchoCP (PFO Diagnosis):** Original image sequences were synthesized into AVI format at 35 FPS and converted to grayscale. We computed mean grayscale intensity $S(c, t)$ for cardiac chambers to characterize contrast microbubble dynamics. Keyframes were selected based on the Top-2 highest combined intensity values. Microbubble distribution was categorized into discrete labels (Max, Min, Mean) to model temporal differences in bubble shunts.
- **HMC-QU (MI Diagnosis):** We leveraged expert-annotated ED/ES frames and video-level Myocardial Infarction (MI) labels to build question-answering pairs. These instructions guide the model to locate structural and functional abnormalities associated with infarction at critical timestamps.

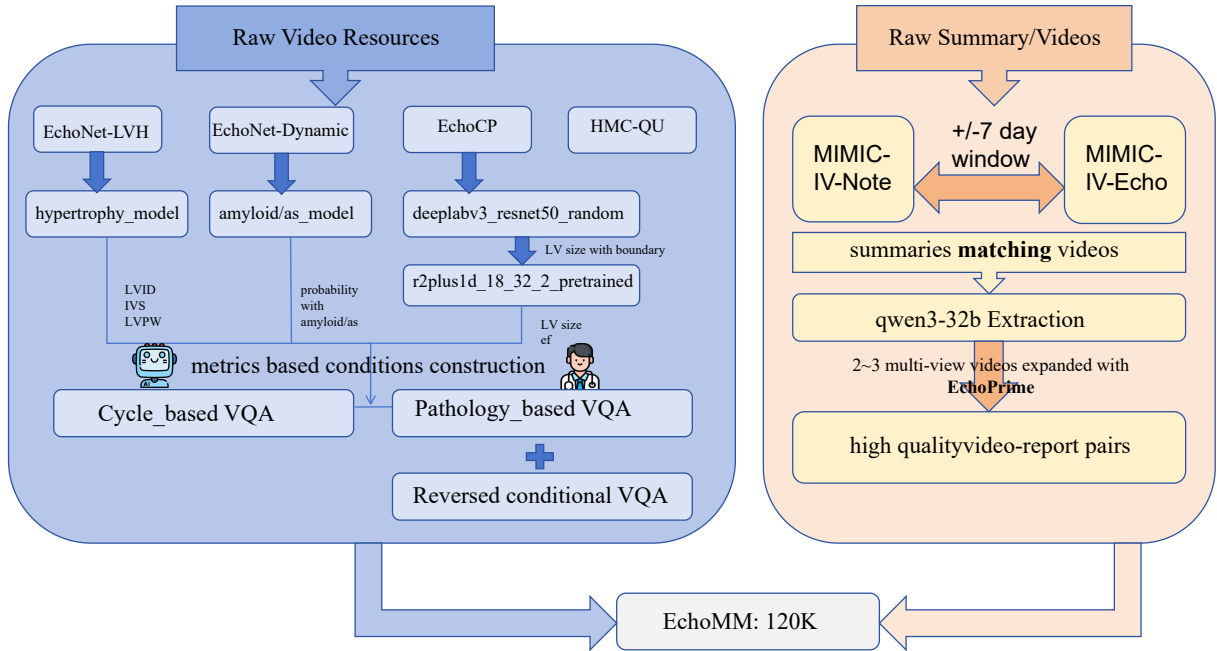


Figure 5: The overall construction framework of the keyframe grounding and echo-report datasets construction

A.2 Multimodal Report Generation (RG) Data Construction

The RG stream curates high-fidelity clinical summaries through a rigorous matching and extraction pipeline based on the MIMIC-IV database.

- **Spatio-Temporal Matching.** We use `subject_id` and `study_id` to align **MIMIC-IV-Echo** videos with **MIMIC-IV-Note** discharge summaries, following $Admission\ Time \leq study_datetime \leq Discharge\ Time$. The closest note within a 7-day window is selected.
- **Expert Content Extraction.** To handle lengthy discharge notes, we employ expert models (e.g., **Lingshu**, **Qwen3-32B**) to extract echocardiography-relevant diagnostic findings while preserving original clinical wording.
- **Backtracking & Integration.** For **MIMICE-choQA**, note consistency is verified using a word inclusion rate:

$$\frac{|W_{EchoQA} \cap W_{Discharge}|}{|W_{EchoQA}|}$$

Extracted contents are merged using the priority: **Lingshu** > **EchoQA** > **Qwen3-32B** > **Qwen3-4B**.

- **Multi-View Expansion.** To simulate real-world examinations, each report is paired with multiple video views. For missing standard views

(A4C, A2C, PLAX), we employ **EchoPrime** to synthesize additional views, enabling robust multi-view learning.

A.3 Splitting for the test

We follow the official dataset splits when available and construct processed splits after preprocessing and filtering. Specifically, for EchoNet-LVH and EchoNet-Dynamic, frame-level annotations are first generated, and samples with incomplete measurements are removed to obtain clean training and test sets. For datasets without official splits (e.g., EchoCP and HMCQU), we manually define train/test partitions.

Table 4 summarizes the resulting data distribution. *Raw* denotes the original dataset splits, while *Proc.* refers to the processed splits used in our experiments. The *Test* column indicates the number of evaluation queries. For cycle-level tasks (e.g., *lvh_cycle* and *dynamic_cycle*), each video typically corresponds to one test sample. In contrast, for disease-level tasks (e.g., EF, AS, and amyloidosis), multiple queries can be generated from a single video based on different clinical attributes, resulting in larger test sets.

Overall, the evaluation covers multiple datasets, views (PLAX and A4C), and task types, providing diverse and clinically grounded test scenarios.

Table 4: **Datasets for keyframe grounding.** Raw videos follow official splits when available; processed videos denote clips after preprocessing.

Dataset	Task	View	Size	Raw (Tr/Val/Te)	Proc. (Tr/Te)	Test
EchoNet-LVH	lvh_cycle	PLAX	12k	10490/1167/343	10134/299	299
	lvh_disease	PLAX			–	897
EchoNet-Dynamic	dynamic_cycle	A4C	10k	7465/1288/1277	8748/1277	1277
	ef_disease	A4C			–	3122
	as_disease	A4C			–	3444
	amyloid_disease	A4C			–	2628
EchoCP	pfo_disease	A4C	60	N/A	48/12	138
HMCQU	MRI_disease	A4C	163	N/A	156/7	42

A.4 The supplement for the construction of the queries

This section provides a supplementary description of the query construction used in our evaluation. Table 5 summarizes the query templates associated with each evaluation task. For each task, we define a set of sub-query categories that capture different semantic aspects of the target clinical concept, and these categories are used to systematically construct the final prompts.

B Details of the Evaluation Metrics

To comprehensively evaluate the performance of multi-modal models in cardiac ultrasound video analysis tasks, we propose a unified evaluation framework covering **keyframe and timestamp grounding, frame similarity, and consistency**. For each input video V , let \mathcal{P} denote the model’s prediction set and \mathcal{G} denote the ground-truth set. We conduct quantitative evaluation through the following three dimensions.

B.1 Keyframe Grounding

This metric aims to measure the ability of the model in grounding key cardiac phases (e.g., End-Diastole [ED], End-Systole [ES]) using the indicator of relative error. Given the periodic nature of ultrasound videos, we adopt the *minimum distance principle*, allowing the model’s predictions to align with any valid candidate cycle except the ground-truth. Let the predicted index of the keyframe be \hat{f} . Let the set of candidate keyframes provided by the ground-truth be $\mathcal{C}_f = \{c_1, c_2, \dots\}$

Minimum Frame Error This calculates the L_1 distance between the predicted frame and the near-

est candidate ground-truth frame:

$$E_{frame} = \min_{c \in \mathcal{C}_f} |\hat{f} - c| \quad (6)$$

For dual-phase tasks (e.g., calculating Ejection Fraction requires locating both ED and ES), we calculate the error for both phases and report the arithmetic mean:

$$E_{avg_frame} = \frac{1}{2}(E_{frame}^{ED} + E_{frame}^{ES}) \quad (7)$$

B.2 Clinical Similarity

Beyond keyframe grounding, the similarity of the identified frames with the ground truth in terms of clinical measurement is crucial. We design task-specific metrics to evaluate this consistency across different datasets.

(1) Relative Metric Error We compute the relative error between the measurement on the predicted frame \hat{f} and the ground-truth frame f^* . Let $M(\cdot)$ be the function that extracts specific clinical measurements. The error is defined as:

$$E_{rel} = \frac{1}{|\mathcal{M}|} \sum_{m \in \mathcal{M}} \frac{|m(\hat{f}) - m(f^*)|}{|m(f^*)| + \epsilon} \quad (8)$$

where \mathcal{M} is the set of task-specific metrics and ϵ is a small constant (e.g., 10^{-6}) for numerical stability. We instantiate \mathcal{M} for each dataset as follows:

- **EchoNet-LVH (lvh_disease, lvh_cycle):** The metric set \mathcal{M} includes three morphological indicators: Left Ventricle Internal Diameter (LVID), Interventricular Septum (IVS), and LV Posterior Wall (LVPW). We report the mean relative error of these three metrics at both End-Diastole (ED) and End-Systole (ES) phases.

- **EchoNet-Dynamic (All tasks) & EchoCP (pfo_disease)**: For these tasks, the fundamental component of \mathcal{M} is the Left Ventricular (LV) area size (measured in pixels) at the ED and ES phases. We further extend \mathcal{M} for specific tasks to align with their clinical definitions:

1. For **ef_disease**, we additionally incorporate the **Ejection Fraction (EF)** relative error derived from the predicted cardiac cycle.
2. For **pfo_disease**, given that the diagnosis relies on a saline contrast "bubble study" where microbubbles passing through the shunt manifest as high-intensity signals, we introduce a **pixel intensity** relative error. This metric captures the brightness shifts within the LV cavity, which geometric area alone cannot reflect.

(2) **Visual Cosine Similarity** For the **MI_disease** task in the **HMC-QU** dataset, where pathological patterns are texture-dependent, we compute the cosine similarity between the *grayscale pixel vectors* of the predicted frame \hat{f} and the ground-truth frame f^* . Let $\mathbf{v}_{\hat{f}} \in \mathbb{R}^{HW}$ and $\mathbf{v}_{f^*} \in \mathbb{R}^{HW}$ denote the vectorized (flattened) grayscale images after spatial alignment (e.g., resizing to the same resolution). The visual similarity is:

$$S_{\text{vis}}(\hat{f}, f^*) = \frac{\mathbf{v}_{\hat{f}}^\top \mathbf{v}_{f^*}}{\|\mathbf{v}_{\hat{f}}\|_2 \|\mathbf{v}_{f^*}\|_2}. \quad (9)$$

A higher S_{vis} indicates that the predicted frame is more visually consistent with the ground-truth frame. We further convert similarity to an error term:

$$E_{\text{vis}} = 1 - \frac{S_{\text{vis}}(\hat{f}_d, f_d^*) + S_{\text{vis}}(\hat{f}_s, f_s^*)}{2}. \quad (10)$$

B.3 Report Generation

For report generation task, we employ a comprehensive multi-dimensional evaluation to assess the quality of generated echocardiography reports against ground-truth references. Given the structured clinical nature of medical report generation, traditional NLP metrics (e.g., BLEU, ROUGE) are insufficient to capture clinical accuracy and safety. Therefore, we leverage GPT-4.1 as an expert evaluator to conduct fine-grained assessments across four

critical dimensions: (1) Structure & Terminology, evaluating the logical organization, appropriate use of medical terminology, and professional formatting; (2) Completeness, measuring the coverage of essential findings and the presence of critical clinical information; (3) Accuracy, assessing factual correctness, precise numerical values, and accurate descriptions of pathological severity; and (4) Conclusion Consistency, examining the alignment between diagnostic conclusions and supporting findings. Each dimension is scored on a 0-5 scale with detailed clinical reasoning, where higher scores indicate better alignment with clinical standards. This LLM-as-a-judge approach enables nuanced evaluation that reflects real-world clinical requirements beyond surface-level textual similarity. The detailed evaluation prompt for report generation is shown in Table 7, and the quantitative results are summarized in Table 6, where EchoMLLM consistently outperforms both open-source and proprietary baselines..

C More Examples of EchoMLLM

To further substantiate the robustness of EchoMLLM, we present expanded qualitative comparisons across diverse tasks. As shown in Fig. 6 (Dynamic Cycle), EchoMLLM successfully identifies a functional cardiac cycle with significant LV area reduction, whereas the baseline predicts static frames with negligible contraction. For pathology-conditioned grounding (Fig. 7), our model explicitly reasons about wall thickness and cavity size to correctly identify "Concentric Remodeling," demonstrating superior diagnostic logic over baselines. Finally, in report generation (Fig. 8), EchoMLLM generates professional-grade documentation with rigorous structure and fine-grained clinical details (e.g., wall motion scores), significantly outperforming the superficial outputs of generalist MLLMs.

D Manual scoring

To strengthen clinical rigor beyond LLM-based judging, we conducted an independent double-blinded clinician evaluation with senior cardiologists from Class-A tertiary hospitals via a Gradio-based platform. For keyframe grounding in Table 9, clinicians performed side-by-side comparisons between EchoMLLM and baselines and reported Win/Loss/Tie/Skipped; we summarize results using WR (Win Rate = Win/Total) and AWR

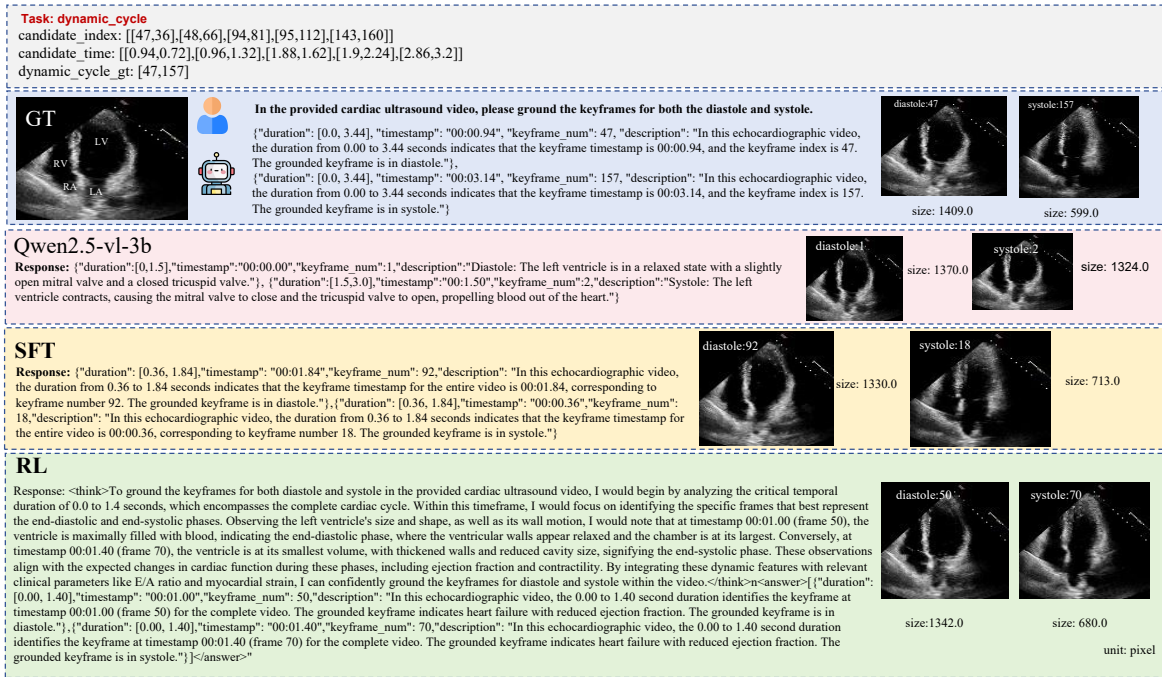


Figure 6: A qualitative case study on the EchoNet-Dynamic `dynamic_cycle` task. The figure compares the ground truth (GT) annotations with predictions from Qwen2.5-VL-3B, the SFT model, and the proposed RL-enhanced model.

(Absolute Win Rate = $\text{Win}/(\text{Win}+\text{Loss})$). For report generation in Table 10, clinicians rated randomized, anonymized reports on a 0–5 Likert scale over ST (Structure & Terminology), CP (Completeness), AC (Accuracy), and CC (Conclusion Consistency), and we report the Mean as the average across these dimensions. As shown in the table below, EchoMLLM achieves higher clinician preference in grounding ($\text{WR} \approx 65\%$, AWR up to 85.12%) and substantially stronger report quality (Mean 4.31/5).

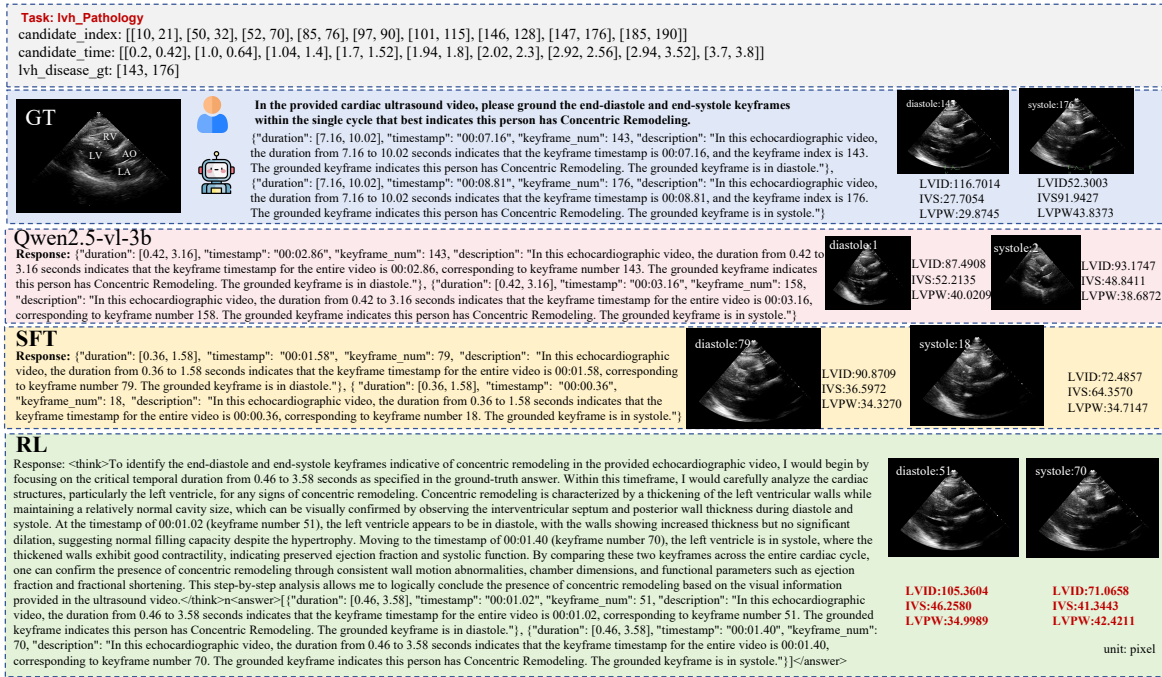


Figure 7: A qualitative case study on the EchoNet-LVH lvh_disease task. The figure compares the ground truth (GT) annotations with predictions from Qwen2.5-VL-3B, the SFT model, and the proposed RL-enhanced model.

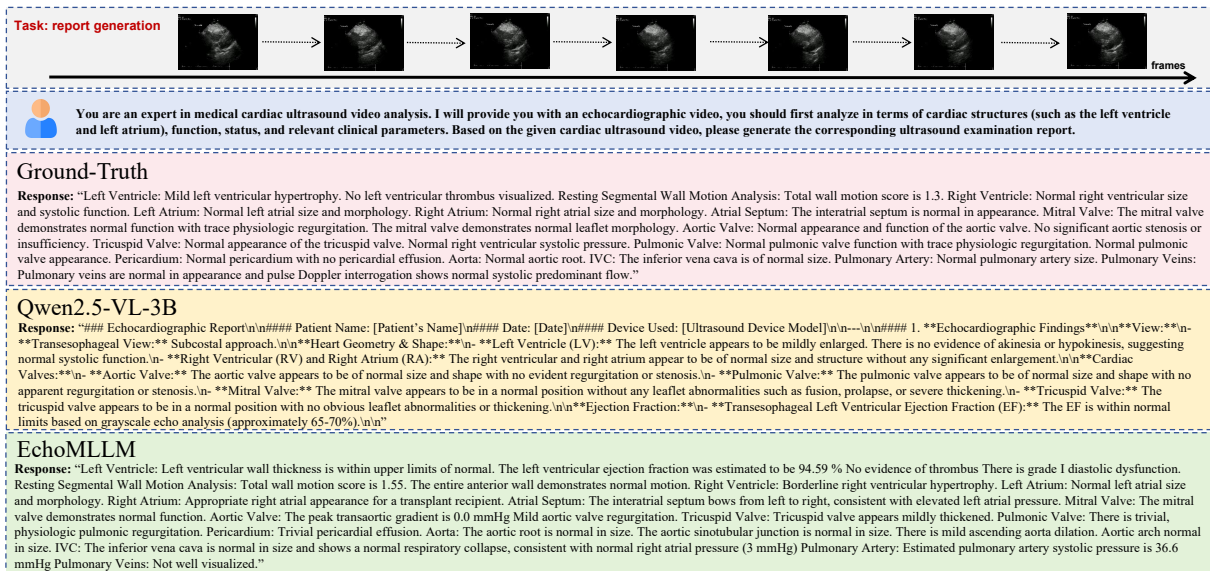


Figure 8: A qualitative case on the MIMIC-IC EchoNote video_to_report task. The figure compares the ground truth (GT) annotations with predictions from Qwen2.5-VL-3B, the SFT model, and the proposed RL-enhanced model.

Table 5: **Query templates for each evaluation task.** Each task contains multiple sub-query categories used to construct prompts.

Task (eval_template)	Query Type (Sub_Query_Category)
dynamic_cycle	Standard_DualPhase
periodic_keyframe_grounding	Standard_DualPhase
disease_lvh	<ul style="list-style-type: none"> • Normal_Geometry • Concentric_Remodeling • Concentric_Hypertrophy • Eccentric_Hypertrophy
disease_ef	<ul style="list-style-type: none"> • Normal_EF_Value • Normal_Function_General • Mild_HFmrEF • Mild_Impaired_Pumping • Mild_Damage_Cause • Reduced_HFrEF • Reduced_Systolic_Failure • Reduced_Significantly_Impaired • Reduced_Severe_Damage_General • Reduced_Severe_Damage_Etiology
disease_as	<ul style="list-style-type: none"> • High_Prob_Value • High_Prob_Anatomy_LVH • High_Prob_Physiology_Afterload • Medium_Prob_Value • Medium_Prob_Anatomy_Suspicious • Medium_Prob_Physiology_EarlyStage • Low_Prob_Value_No • Low_Prob_Anatomy_DoesNot • Low_Prob_Physiology_NotHigh
disease_amyloid	<ul style="list-style-type: none"> • High_Prob_Value • High_Prob_Anatomy • High_Prob_Physiology • Medium_Prob_Value • Medium_Prob_Generic • Medium_Prob_Anatomy • Medium_Prob_Physiology • Medium_Prob_Symptoms • Low_Prob_Value_No • Low_Prob_Anatomy_DoesNot • Low_Prob_Physiology_Normal
disease_pfo	<ul style="list-style-type: none"> • Shunt_Fraction_High • Shunt_Fraction_Low • Bubble_Count_High • Bubble_Count_Low • Septum_Mobility • Septal_Aneurysm • Risk_Embolism • Risk_Stroke • Diagnosis_General • Visual_Bubble_Distribution
disease_MI	<ul style="list-style-type: none"> • RWMA • Remodeling • Thinning_Fibrosis • Ischemic_Cardiomyopathy • Mitral_Regurgitation • General_Lesion

Table 6: Report generation performance with mean \pm standard deviation. Fmt., Comp., Acc., and Consis. denote Format, Completeness, Accuracy, and Consistency, respectively. Best results are highlighted in bold.

Category	Model	Fmt.	Comp.	Acc.	Consis.	Avg.
API-based	GPT-4o	1.8895 \pm 0.0703	0.3876 \pm 0.0076	0.2114 \pm 0.0118	0.1724 \pm 0.0067	0.6652 \pm 0.0227
	GPT-4-turbo	2.2105 \pm 0.0943	0.5714 \pm 0.0094	0.2514 \pm 0.0070	0.2124 \pm 0.0047	0.8115 \pm 0.0256
	Gemini-2.5-pro	3.5152 \pm 0.0218	1.5390 \pm 0.0050	1.6171 \pm 0.0262	1.4276 \pm 0.0088	2.0248 \pm 0.0035
Open-source	Qwen2.5-VL-3B	2.5929 \pm 0.0100	1.2829 \pm 0.0086	1.1800 \pm 0.0057	1.1300 \pm 0.0129	1.5464 \pm 0.0021
	Qwen3-VL-4B	1.8819 \pm 0.0295	0.9276 \pm 0.0047	0.8819 \pm 0.0082	0.7781 \pm 0.0082	1.1174 \pm 0.0153
	Lingshu-7B	3.0300 \pm 0.0443	1.3114 \pm 0.0057	1.2186 \pm 0.0100	1.1729 \pm 0.0129	1.6832 \pm 0.0132
	Lingshu-32B	3.3171 \pm 0.0314	1.5171 \pm 0.0200	1.3157 \pm 0.0071	1.2629 \pm 0.0086	1.8532 \pm 0.0011
	Hulu-Med-7B	2.2300 \pm 0.0043	0.8100 \pm 0.0186	0.9514 \pm 0.0229	0.8343 \pm 0.0171	1.2064 \pm 0.0064
	Hulu-Med-32B	2.8329 \pm 0.0157	1.1871 \pm 0.0100	1.0714 \pm 0.0171	1.0157 \pm 0.0071	1.5268 \pm 0.0046
	EchoMLLM	4.0886 \pm 0.0230	2.4781 \pm 0.0082	1.7924 \pm 0.0153	1.8152 \pm 0.0153	2.5436 \pm 0.0105

Table 7: **Evaluation Prompt for Report Generation.** This prompt is used by the LLM judge (e.g., GPT-4) to assess the clinical quality of generated reports against ground truth.

System Instruction:

You are an experienced **Senior Cardiac Ultrasound Expert** with over 10 years of clinical experience in cardiovascular diagnostics. You are reviewing an AI-generated echocardiography report. Your primary goal is to critically assess the AI report’s **clinical safety, accuracy, and professionalism** from a clinical perspective.

Task:

Please strictly compare the [Ground Truth Report] and the [Generated Report]. Assign a **detailed sub-score** (0-5) for each evaluation criterion, accompanied by professional, in-depth reasoning.

Evaluation Dimensions:

- **1. Structure & Terminology (0-5):** Proper organization and logical flow; Correct use of medical terminology; Professional formatting.
- **2. Completeness (0-5):** Coverage of all important findings; No missing critical information; Appropriate level of detail.
- **3. Accuracy (0-5):** Factual correctness; Correct negations and affirmations; Accurate severity descriptions; Precise numerical values.
- **4. Conclusion Consistency (0-5):** Agreement with reference report’s diagnosis; Logical consistency between findings and conclusions; Appropriate clinical impressions.

Scoring Guidelines:

- **5: Excellent** - Perfect or near-perfect match with ground truth.
- **4: Good** - Minor differences that don’t affect clinical interpretation.
- **3: Acceptable** - Some inaccuracies but main points are correct.
- **2: Poor** - Significant errors or missing information.
- **1: Very Poor** - Major errors that could affect diagnosis.
- **0: Unacceptable** - Completely incorrect or irrelevant.

Output Format:

Please output the evaluation result in JSON format:

```
{
  "structure_terminology": <0-5>,
  "completeness": <0-5>,
  "accuracy": <0-5>,
  "conclusion_consistency": <0-5>,
  "reasoning": {
    "structure_terminology": "<Brief explanation>",
    "completeness": "<Brief explanation>",
    "accuracy": "<Brief explanation>",
    "conclusion_consistency": "<Brief explanation>"
  }
}
```

Table 8: A Case of Pathology-Aware Keyframe Grounding Reasoning Prompt (Aortic Stenosis).

Prompt:
 You are a medical specialist in cardiac ultrasound, with expertise in interpreting and analysing echocardiography videos. I will provide the following inputs:

- A cardiac ultrasound video of a healthy individual or a patient with aortic stenosis
- A question about grounding keyframes of normal or aortic stenosis conditions within the video
- The groundtruth answer to the keyframe information, where `duration` indicates the time period that best demonstrates the disease progression and keyframes in the video, `timestamp` represents the temporal position of the keyframe within the video, `keyframe_num` denotes the sequential frame number of the keyframe in the video, and `description` provides a characterization of the corresponding keyframe information

Your task is to simulate a step-by-step reasoning process that determines how to ground keyframes in the ground-truth answer. However, you must assume you do not know the keyframes' information in advance.

- Carefully analyze both the video and question, while referring to the symptoms/conditions/states provided in the question to identify normal or diseased cases, which helps ground the keyframes.
- Refer to the content in the ground-truth answer (Do NOT directly copy or repeat any content from the ground-truth answer) to form a reasoning flow: First, in the cardiac ultrasound video, focus on identifying the critical temporal duration corresponding to the cardiac motion timeline/trajectory/process as specified in `duration`, then describe and analyze the characteristics and manifestations observed during this temporal segment as outlined in `description`. Additionally, incorporate reasoning and analysis across the following dimensions: cardiac structures (such as the left ventricle, left atrium, interventricular septum, and valve apparatus), function (including systolic and diastolic function, ejection fraction, and contractility), status (such as wall motion abnormalities, chamber dilation, hypertrophy, and perfusion), and relevant clinical parameters (including ejection fraction, fractional shortening, E/A ratio, and myocardial strain).
- Reason step-by-step as an ultrasound medical expert would, leveraging the cardiac beating/dynamic features and visual information in ultrasound video, along with textual cues in question-answer pair, to reach a logical reasoning and synthesis.
- The reasoning should proceed from observation and analysis to conclusions.
- Output the reasoning in a one paragraph and the reasoning/thinking should not be repetitive, overly redundant and too long.

Question: {question}
 Answer: {answer}

Table 9: Comparison results on the Keyframe Grounding task. WR and AWR denote Win Rate and Adjusted Win Rate (excluding ties and skipped samples), respectively. Best results are highlighted in bold.

Task	Model Pair	Win	Loss	Tie	Skip	Total	WR (%)	AWR (%)
Keyframe Grounding	EchoMLLM vs. Lingshu	44	10	14	17	68	64.71	81.48
Keyframe Grounding	EchoMLLM vs. Qwen2.5-VL	143	25	49	21	217	65.90	85.12

Table 10: Results on the report generation task (0–5 Likert scale). ST, CP, AC, and CC denote Structure, Clinical Precision, Accuracy, and Coherence, respectively. Best results are highlighted in bold.

Task	Model	ST	CP	AC	CC	Mean
Report Generation	Qwen3-VL	3.54	2.67	2.29	1.88	2.59
Report Generation	Gemini 2.5 Pro	2.40	1.50	1.50	1.80	1.80
Report Generation	GPT-4o	2.00	1.50	1.25	1.00	1.44
Report Generation	EchoMLLM	4.76	4.35	4.24	3.87	4.31