Can Vision Language Models Track a Heartbeat? A Benchmark on Frame-Level Echocardiogram Understanding

Dingming Liu[1] DINGMING.LIU@CHARITE.DE
Nabil Jabareen[*1] NABIL.JABAREEN@GMAIL.COM
Soeren Lukassen[*1] SOEREN.LUKASSEN@BIH-CHARITE.DE
[1] *Center of Digital Health, Berlin Institute of Health at Charite – Universitaetsmedizin Berlin, Berlin, Germany*

Editors: Under Review for MIDL 2026

Abstract

Echocardiogram videos are among the most common and clinically vital imaging modalities in cardiovascular medicine. They capture dynamic cardiac motion, and their accurate functional assessment requires frame-level temporal precision. Ejection fraction (EF) is an essential metric for assessing cardiac function and is computed from the left-ventricular volumes at end-diastole (EDV) and end-systole (ESV), making its estimation inherently dependent on accurate frame-wise temporal reasoning. Vision Language Models (VLMs) have recently shown strong performance in general video understanding. However, whether they can reliably reason over the fine-grained temporal dynamics required for echocardiographic interpretation remains unclear.

We benchmarked six state-of-the-art open-source VLMs, Gemma 3n, LLaVA-Interleave, LLaVA-NeXT-Video 7B/34B, and Qwen3-VL 8B/32B, on the clinically motivated task of frame-level EDV/ESV localization in apical four-chamber echocardiograms. All models performed poorly on this localization task, with errors far beyond clinically acceptable tolerances, and in some cases indistinguishably from random Monte Carlo baselines. To further test whether explicit structural guidance could compensate for limited temporal reasoning, we additionally provided left-ventricular segmentation overlays as auxiliary visual input for both tasks. However, even with segmentation cues, performance gains remained negligible in this tasks. Prompting the model to focus on masked areas only, omitting any medical context, did not lead to marked improvements.

To reduce the complexity to pure size comparison, we further evaluated a simplified two-frame binary classification task in which each model must distinguish end-diastole (ED) from end-systole (ES). Despite this simplification, performance remained low for most models on original videos, only Qwen3-VL-32B reaches an accuracy of 0.711. Providing segmentation overlays and ignoring medical background knowledge only helped Qwen3-VL in both sizes reaches accuracy over 0.9, with other models resulting in random level.

This work presents the first systematic evaluation of general-purpose VLMs on echocardiogram video analysis across progressively simplified temporal reasoning tasks. Our results reveal a fundamental limitation of current VLMs in frame-level cardiac ultrasound interpretation. This work highlights the importance of medical benchmarks for VLMs and the need for domain-specific temporal modeling in future medical VLMs. To facilitate benchmarking of VLMs on echocardiogram video analysis, we make the benchmark and all associated code publicly available here.

Keywords: Echocardiography, Vision Language Models, Benchmarking

[*] Contributed equally

1. Introduction

Transthoracic echocardiography is the most widely used imaging modality for the assessment of cardiac disease(Lang et al., 2015) (Gillam and Marcoff, 2024). Because it is non-invasive and does not involve ionizing radiation (Gillam and Marcoff, 2024), it is often the first-line imaging method for patients with suspected cardiovascular disease(Steeds, 2011). To assess global cardiac function, the ejection fraction (EF) is typically derived from left ventricular volumes at end-diastole (EDV) and end-systole (ESV). This process requires accurately identifying the corresponding ED and ES frames in echocardiogram videos, then estimating chamber volumes from these frames. Accurate quantification of cardiac chamber size and function is therefore a cornerstone of cardiac imaging (Lang et al., 2015). In routine clinical practice, however, these measurements still rely heavily on time-consuming manual work(Zolgharni et al., 2017).

Vision-language models (VLMs) have recently demonstrated strong performance on a range of general video understanding tasks, including action recognition, temporal event localization, and video question answering (Chambon et al., 2022) (Li et al., 2024) (Team, 2025b). These advances raise the question of whether such models can be repurposed for medical video interpretation. For echocardiography in particular, a promising use case would be automating frame-level tasks such as locating EDV and ESV frames as an intermediate step towards EF estimation and downstream decision support. Before such applications can be considered, VLMs must be systematically evaluated on whether they can reliably reason over fine-grained cardiac motion at the frame level.

In this work, we focused on the clinically motivated task of frame-level EDV and ESV localization in apical four-chamber echocardiogram videos. Our primary objective is to assess whether state-of-the-art open-source VLMs can identify the frames corresponding to the largest and smallest left ventricular cavity. To this end, we defined a frame localization task (T1) in which each input subsequence was constructed to contain exactly one EDV frame and one ESV frame, and the models were prompted to output the indices of these frames within a specified index range.

To study how different forms of visual and textual guidance affect frame localization, we evaluated T1 under three levels of varying context. In the original setting, models received unmodified echocardiogram videos. In the segmented setting, models received videos with a mask highlighting the left ventricle. In the non-medical segmented setting, the same overlays were used, but models were explicitly instructed to disregard medical context and consider only the size of the masked region. This hierarchy of input conditions allowed us to test whether explicit structural cues or the removal of medical terminology could compensate for limited temporal reasoning.

In addition to this main benchmark, we included a simpler two-frame discrimination task (T2) as an auxiliary analysis. In T2, models were asked to select the EDV or ESV frame from a pair of annotated frames extracted from the same video. This auxiliary task removed long-range temporal context and reduced the problem to direct size comparison between two candidate frames, which helped interpret failures observed in the more challenging localization setting.

Across these tasks and input conditions, we systematically assessed whether current VLMs could achieve clinically meaningful frame-level performance on echocardiogram videos.

2. Methods

2.1. Dataset

We based our experiments on a subset containing 100 videos from the EchoNet-Dynamic dataset (Ouyang et al., 2020), which contains over 10,000 apical four-chamber echocardiography videos from individuals who underwent imaging at Stanford University Hospital. For each video, the left ventricle was traced along the endocardial border at one EDV and one ESV frame within a single cardiac cycle. This provided ground-truth frame indices for a single EDV-ESV pair per video. Our subset was chosen randomly from the validation set.

In addition to the original videos, we used frame-by-frame semantic segmentations (Ouyang et al., 2020) of the left ventricle on our chosen subset videos of EchoNet-Dynamic. The segmented version of each video contained a red mask indicating the left ventricular region in every frame. These segmentation overlays allowed us to study whether explicit structural cues helped VLMs focus on the relevant anatomy.

2.2. Frame localization task (T1)

Our primary task is frame-level localization of EDV and ESV. For each video, we extracted a contiguous subsequence corresponding to approximately one cardiac cycle. This subsequence was chosen such that the annotated EDV and ESV frames both lay strictly within the subsequence and were neither the first nor the last frame. This design ensured that each input contained exactly one EDV–ESV pair while reducing the temporal length of the sequence to make the task more tractable.

For each subsequence, we re-centered the frame indices to a local range $[0, \ell_i + 1)$ for subsequence length ℓ_i. The models were informed about the number of frames and the valid index range and were required to select indices only from this range. Prompts also instructed the models to output frame indices in a predefined format. Structured prompts such as "*I sampled exactly ℓ_i frames from the video. Only use these ℓ_i frames. Find the frame indices of ESV (smallest Left Ventricle cavity) and EDV (largest Left Ventricle cavity) from input frames indexing ranging in [0, $\ell_i + 1$). Reply exactly as: Frame: EDV = <int>, ESV = <int>*". To probe instruction sensitivity, we considered three prompting variants, asking for EDV first and ESV second, reversing this order, and requesting only a single phase (EDV or ESV) at a time.

T1 was evaluated under three different context conditions:

- **Original**. Models received the original echocardiogram frames.
- **Segmented**. Models received videos with a red mask overlay highlighting the left ventricle. Prompts explained that "the left ventricle is marked as the red region in all frames" and that EDV and ESV correspond to the frames with the largest and smallest red region area.
- **Non-medical**. Models received the same segmented videos but were explicitly instructed to ignore medical context and to focus solely on the size of the red region, for example by asking for "the frame with the largest red region".

Together, these settings defined the main benchmark conditions of T1 on original, segmented, and non-medical inputs. For the full prompts see here.

2.3. Auxiliary two-frame discrimination task (T2)

To complement the localization benchmark, we defined an auxiliary two-frame discrimination task. For each video, we extracted the annotated EDV and ESV frames and formed a short two-frame sequence. In separate runs, each model was instructed to either identify the EDV frame or to identify the ESV frame. The same three context conditions as in T1 were considered.

T2 removed the need for long-range temporal integration and transferred the problem to direct comparison of cavity size between two candidate frames. We used this task primarily as a diagnostic tool to assess whether failures on T1 were due to temporal reasoning, difficulty in attending to the relevant regions, or more basic limitations in comparing differences in cavity size.

2.4. Models and evaluation pipeline

We benchmarked six state-of-the-art open-source VLMs that support video or multi-image input, namely LLaVA-Interleave-7B(Li et al., 2024), LLaVA-NeXT-Video-7B, LLaVA-NeXT-Video-34B (Zhang et al., 2024)(Liu et al., 2024), Qwen3-VL-8B-Instruct, Qwen3-VL-32B-Instruct(Team, 2025b)(Bai et al., 2023), Gemma 3n E4B IT (Instruct)(Team, 2025a).

All models were accessed via Hugging Face. For each task and input modality, all models received semantically equivalent text instructions, adapted only as needed to match model-specific formatting requirements.

Because of architectural differences, models accepted video input in different forms. Gemma 3n and LLaVA-Interleave processed multiple images, which were provided as individual frames sampled from the same video subsequence. LLaVA-NeXT-Video and Qwen3-VL processed a single video tensor directly, which were fed with temporally ordered frames.

For each model and experimental condition, we ran ten different random seeds. For each seed, three independent passes were performed per video and the predictions were aggregated to reduce stochastic variability. For T1, we reported frame-level mean absolute error (MAE) and correlation-based R^2 of frame indices between predictions and ground truth. For a better understanding of the model performance, we further calculated the EF from the volumes of the predicted EDV and ESV frames by $EF = \frac{EDV - ESV}{EDV}$ for each video. MAE was also calculated from calculated EF and ground truth EF. For T2, we reported classification accuracy.

2.5. Monte Carlo random baselines

To quantify improvements over chance, we estimated random baselines using Monte Carlo simulation for both localization and binary discrimination tasks.

For T1, each video and each condition specified a searchable frame index range $[0, i_{\max})$ and a ground-truth EDV or ESV frame index i^*. In each Monte Carlo trial, we sampled a single integer frame index uniformly at random from this range for each video and computed the MAE of the index error averaged over all videos. Repeating this procedure $T = 10{,}000$ times yielded an empirical distribution of random MAE values from which we derived the mean random MAE and its 95% confidence interval.

For T2, each evaluation run consisted of 100 two-frame samples with a binary ground-truth label vector $y \in \{0,1\}^{100}$ that indicated which frame was EDV or ESV. Repeating this process $T = 10{,}000$ times per run yielded an empirical distribution of random accuracies. We reported the mean and 95% confidence intervals of these random baselines alongside model accuracies.

3. Experiments and Results

In this section, we present a comprehensive evaluation of six open-source VLMs on the frame localization task across different context conditions, followed by a brief analysis of the auxiliary two-frame discrimination task. Unless otherwise specified, we focus on frame-level mean absolute error (MAE) for T1 and classification accuracy for T2.

Across all T1 settings, we found that current VLMs struggled substantially, with localization errors far beyond clinically acceptable tolerances and often close to random performance.

3.1. T1 on different settings

We first evaluated all models on T1 using original echocardiogram videos as defined in Section 2.2. The results are summarized in Table 1. In the setting where EDV was requested first and ESV second, which reflects the typical order in clinical practice, Gemma-3n achieved the best in EDV localization and LLaVA-NeXT-Video-34B had the best score in ESV localization. Qwen3-VL-8B outperformed in the ESV first and EDV second setting with the best scores in both EDV and ESV predictions. These models had also the best performances in ESV-only (Gemma-3n: 4.336) and EDV-only (Qwen3-VL-8B: 6.545) tasks (Figure 1, Figure 2). However, these errors were still large compared to typical human variability, which is often within 2 to 3 frames. The calculated MAE of EF further proved this, even the model with the best EF prediction failed. The remaining models exhibited even larger errors and, in some conditions, performed only marginally better than the Monte Carlo random baseline (Figure 1, Figure 2). Overall, we did not observe a consistent pattern of model performance across prompting variants. Instead, each model's performance remained unpredictable across settings, supporting the suspicion that the models did not truly understand the videos but were effectively outputting frame indices close to chance.

We next repeated T1 on segmented videos where the left ventricle was highlighted by a red mask, as described in Section 2.2. Full results shown in Table 2. Contrary to our expectation, providing segmentation information did not improve performance for most models (Figure 1, Figure 2). In many cases, localization errors became even larger than on original videos.

LLaVA-NeXT-Video-34B achieved the best performance among segmented-input models when prompted to report EDV first and ESV second, while Gemma 3n performed best on the ESV-first variant and ESV only settings. However, only the ESV predictions of these settings clearly outperformed the random baseline. For the single-phase localization variants, none of the models consistently outperformed the Monte Carlo baseline, suggesting that apparent successes on joint EDV and ESV prediction may have been driven by chance.

To test whether medical terminology and anatomical priors distracted VLMs from the visual evidence, we performed an additional variant of T1 on segmented videos with non-

		EDV (frame)		ESV (frame)		EF (%)
Model	**Output format**	**MAE↓**	R^2 ↑	**MAE↓**	R^2 ↑	**MAE↓**
Random	-	11.874	-	11.616	-	-
LLaVA-Interleave	edv_esv	100.678	0.023	85.281	0.082	51.489
	esv_edv	46.044	0.046	38.630	0.038	65.320
	edv	7538.389	0.032	-	-	-
	esv	-	-	15.199	0.081	-
LLaVA-NeXT-Video-7B	edv_esv	25.718	0.007	20.747	0.032	34.364
	esv_edv	28.144	0.033	22.583	0.025	52.662
	edv	33.137	0.039	-	-	-
	esv	-	-	28.298	0.030	-
LLaVA-NeXT-Video-34B	edv_esv	10.087	0.170	**6.292**	0.372	37.045
	esv_edv	16.491	0.333	10.756	0.145	60.292
	edv	8.517	0.053	-	-	-
	esv	-	-	10.956	0.168	-
Gemma-3n	edv_esv	**7.653**	**0.294**	10.311	0.295	41.308
	esv_edv	18.845	**0.591**	11.294	0.138	75.020
	edv	9.683	**0.530**	-	-	-
	esv	-	-	**4.336**	**0.575**	-
Qwen3-VL-8B	edv_esv	8.024	0.191	7.058	**0.762**	28.151
	esv_edv	**7.125**	0.011	**10.320**	**0.238**	**36.567**
	edv	**6.545**	0.064	-	-	-
	esv	-	-	11.341	0.091	-
Qwen3-VL-32B	edv_esv	10.049	0.029	7.828	0.544	**25.734**
	esv_edv	12.131	0.045	18.206	0.011	59.155
	edv	6.641	0.103	-	-	-
	esv	-	-	11.222	0.251	-

Table 1: Metrics of model performance in T1 original. We highlighted the best model performance in different context conditions.

medical instructions, again following the setup in Section 2.2. As results listing in Table 3, most of the models surpassed the random baselines. Gemma-3n, however, achieved the best MAE of 2.875 in ESV-only prediction setting, but a lot worse in EDV-only and other tasks. Overall, removing medical context and relying purely on geometric size cues did not help current VLMs perform meaningful frame localization, even when the relevant region was explicitly highlighted.

For some seeds and prompting variants, we observed degenerate behavior. The model repeatedly output the same frame index for many videos, which led to artificially low variability in predictions without reflecting meaningful localization ability. When we reversed the order in which the model was instructed to report ESV and EDV, we observed a consistent bias, with models tending to assign earlier frame indices to whichever term was requested first and later indices to the term requested second. Asking for only a single phase at a time reduced this order bias and led to predictions that were more evenly dis-

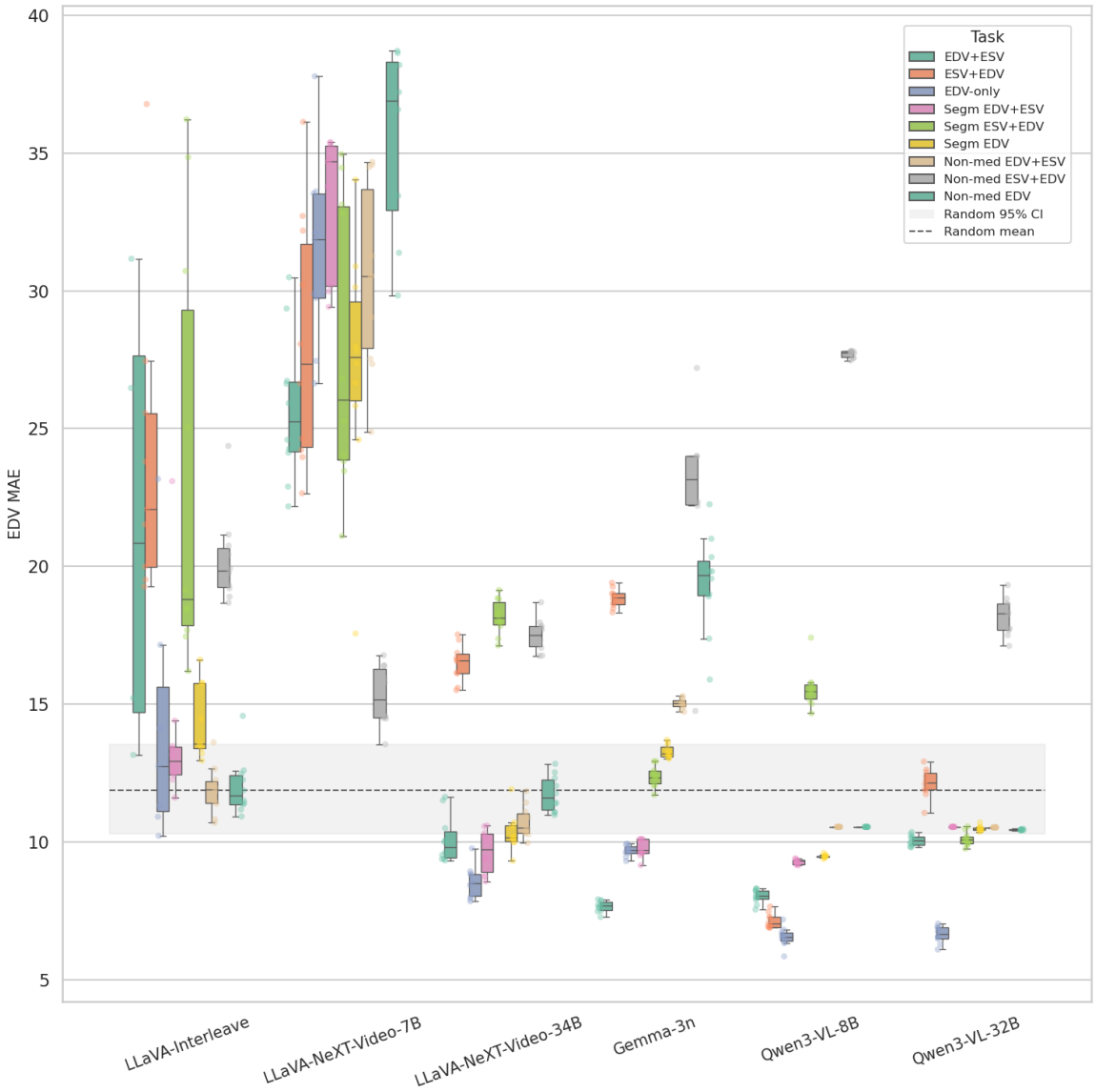


Figure 1: Frame MAE of frame localization task T1 - EDV

tributed along the diagonal when plotting ground-truth versus predicted indices. However, the overall MAEs remained high and far from clinical requirements.

We also observed systematic differences in how strictly models followed the specified index range. Smaller models often produced indices outside the admissible range, and in some cases even returned clearly nonsensical values, including negative indices or variants such as “minus zero”. In contrast, the larger models, in particular LLaVA-NeXT-Video-34B and Qwen3-VL-32B, were more likely to respect the output constraints and restrict their predictions to the instructed index range.

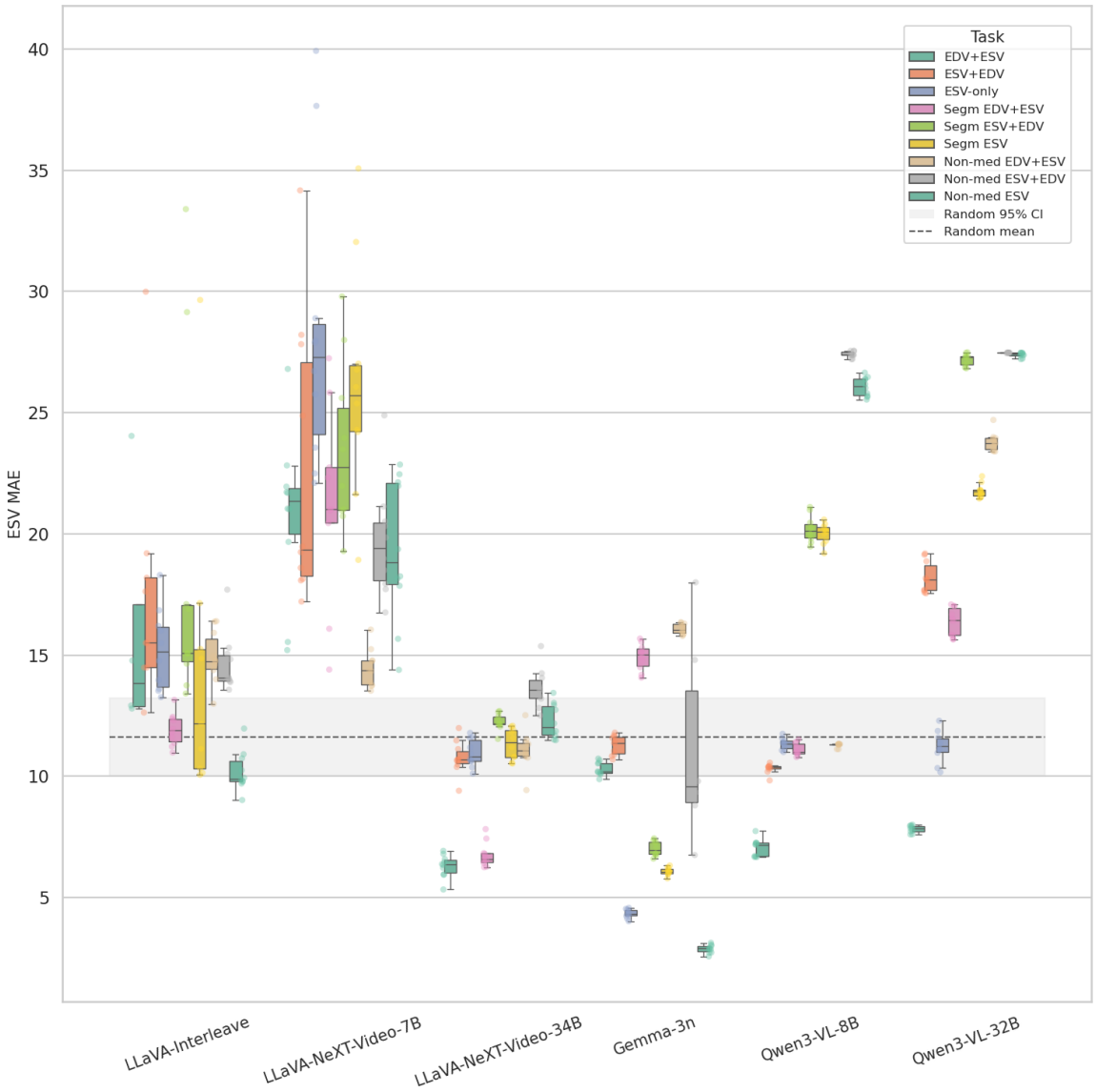


Figure 2: Frame MAE of frame localization task T1 - ESV

Among all evaluated models, Gemma-3n achieved a surprisingly low MAE on the ESV-only task across different context conditions (Figure 2), which at first glance might suggest that it understood the ESV-only prediction task better than other models. However, this performance was largely driven by a simple heuristic behavior, namely a strong tendency to output frame indices near the middle of the provided index range (Seen in Appendix A Figure 5, Figure 4). In our setup, the ESV frame is typically interpreted second in the sequence, so a mid-range guess can accidentally align with the true ESV position more often than chance. As a result, the apparently favorable MAE for Gemma-3n in the ESV-only

		EDV (frame)		ESV (frame)		EF (%)
Model	**Output format**	**MAE↓**	R^2 ↑	**MAE↓**	R^2 ↑	**MAE↓**
Random	-	11.874	-	11.616	-	-
LLaVA-Interleave	edv_esv	13.902	0.148	11.927	0.222	49.422
	esv_edv	23.420	0.197	18.358	0.077	62.332
	edv	18.053	0.114	-	-	-
	esv	-	-	22.016	0.111	-
LLaVA-NeXT-Video-7B	edv_esv	33.941	0.015	21.006	0.014	44.493
	esv_edv	27.987	0.019	23.483	0.022	46.534
	edv	27.289	0.011	-	-	-
	esv	-	-	26.143	0.011	-
LLaVA-NeXT-Video-34B	edv_esv	9.659	0.074	**6.751**	0.436	38.858
	esv_edv	18.169	**0.435**	12.241	0.084	67.417
	edv	10.354	0.136	-	-	-
	esv	-	-	11.333	0.120	-
Gemma-3n	edv_esv	9.758	**0.392**	14.920	0.036	58.559
	esv_edv	12.374	0.265	**7.015**	0.103	**35.394**
	edv	13.280	**0.432**	-	-	-
	esv	-	-	**6.078**	**0.646**	-
Qwen3-VL-8B	edv_esv	**9.272**	0.051	11.104	**0.563**	**31.633**
	esv_edv	15.568	0.008	20.194	**0.157**	48.883
	edv	**9.478**	0.002	-	-	-
	esv	-	-	20.017	0.044	-
Qwen3-VL-32B	edv_esv	10.540	-	16.364	0.067	31.809
	esv_edv	**10.106**	0.089	27.192	0.013	55.900
	edv	10.480	0.005	-	-	-
	esv	-	-	21.777	0.004	-

Table 2: Metrics of model performance in T1 segmented. The best model performance among each ouput format was highlighted.

setting does not reflect genuine frame-level understanding, but rather an artifact of this mid-range bias.

3.2. Auxiliary results on T2

Finally, we summarize the results of the auxiliary two-frame discrimination task. In T2, each model was asked to select the EDV or ESV frame from a pair of annotated frames extracted from the same video. This removed long-range temporal context and reduced the problem to comparing cavity size between two candidate frames. All results shown in Table 4.

On original frames (Figure 3), several models such as LLaVA-NeXT-Video in both sizes and Gemma-3n failed to achieve accuracies substantially above the random baseline of 0.5 and sometimes tended to output the same index across many videos. Qwen3-VL-32B was the only model that consistently distinguished EDV and ESV on original frames with

Model	Output format	EDV (frame) MAE↓	EDV (frame) R^2 ↑	ESV (frame) MAE↓	ESV (frame) R^2 ↑	EF (%) MAE↓
Random	-	11.874	-	11.616	-	-
LLaVA-Interleave	edv_esv	11.872	0.213	14.913	0.125	53.305
	esv_edv	20.245	0.299	14.637	0.177	59.582
	edv	12.009	0.218	-	-	-
	esv	-	-	10.197	0.283	-
LLaVA-NeXT-Video-7B	edv_esv	30.490	0.034	14.444	0.052	53.601
	esv_edv	**15.291**	0.038	19.660	0.009	59.490
	edv	36.687	0.056	-	-	-
	esv	-	-	19.312	0.048	-
LLaVA-NeXT-Video-34B	edv_esv	10.717	0.161	**11.079**	0.152	52.074
	esv_edv	17.504	0.301	13.635	0.042	66.697
	edv	11.736	0.126	-	-	-
	esv	-	-	12.268	0.060	-
Gemma-3n	edv_esv	15.015	**0.438**	16.073	0.382	91.337
	esv_edv	22.414	0.516	**11.247**	**0.443**	62.770
	edv	19.395	**0.453**	-	-	-
	esv	-	-	**2.875**	**0.670**	-
Qwen3-VL-8B	edv_esv	10.540	-	11.282	**0.663**	**31.386**
	esv_edv	27.705	**0.593**	27.397	0.000	55.712
	edv	10.540	-	-	-	-
	esv	-	-	26.077	0.010	-
Qwen3-VL-32B	edv_esv	**10.521**	0.000	23.788	0.017	36.995
	esv_edv	18.204	0.070	27.471	0.003	**54.360**
	edv	**10.435**	0.047	-	-	-
	esv	-	-	27.388	0.011	-

Table 3: Metrics of model performance in T1 non-medical. The best model performance among each output format was highlighted.

accuracies above 0.58. Segmentation overlays improved the performance of Qwen3-VL-8B and Qwen3-VL-32B, and in the non-medical setting Qwen3-VL-32B reached accuracies above 0.9 on some sub-tasks.

These auxiliary results indicated that even when the task was simplified to a two-frame comparison, most general-purpose VLMs did not reliably distinguish EDV from ESV. Stronger models benefited from segmentation cues in this simplified setting, but this did not translate into robust frame localization performance in T1.

4. Discussion

Our benchmark focused on the clinically motivated task of frame-level EDV and ESV localization in apical four-chamber echocardiogram videos across multiple context conditions. The core result is that current open-source VLMs were unable to achieve clinically meaning-

		Accuracy	
		EDV	ESV
LLaVA-NeXT-Video-34B	original video	0.488	0.528
	segmented	0.467	0.532
	non_medical	0.480	0.510
LLaVA-NeXT-Video-7B	original video	0.512	0.502
	segmented	0.510	0.500
	non_medical	0.523	0.499
LLaVA-Interleave	original video	0.486	0.497
	segmented	0.489	0.511
	non_medical	0.507	0.476
Qwen3-VL-32B	original video	**0.711**	**0.626**
	segmented	0.942	**0.888**
	non_medical	**0.961**	0.932
Qwen3-VL-8B	original video	0.510	0.513
	segmented	**0.956**	0.685
	non_medical	0.906	**0.962**
Gemma-3n	original video	0.514	0.514
	segmented	0.514	0.514
	non_medical	0.486	0.514

Table 4: Accuracy of model performance in T2. The best accuracy in three context conditions are highlighted.

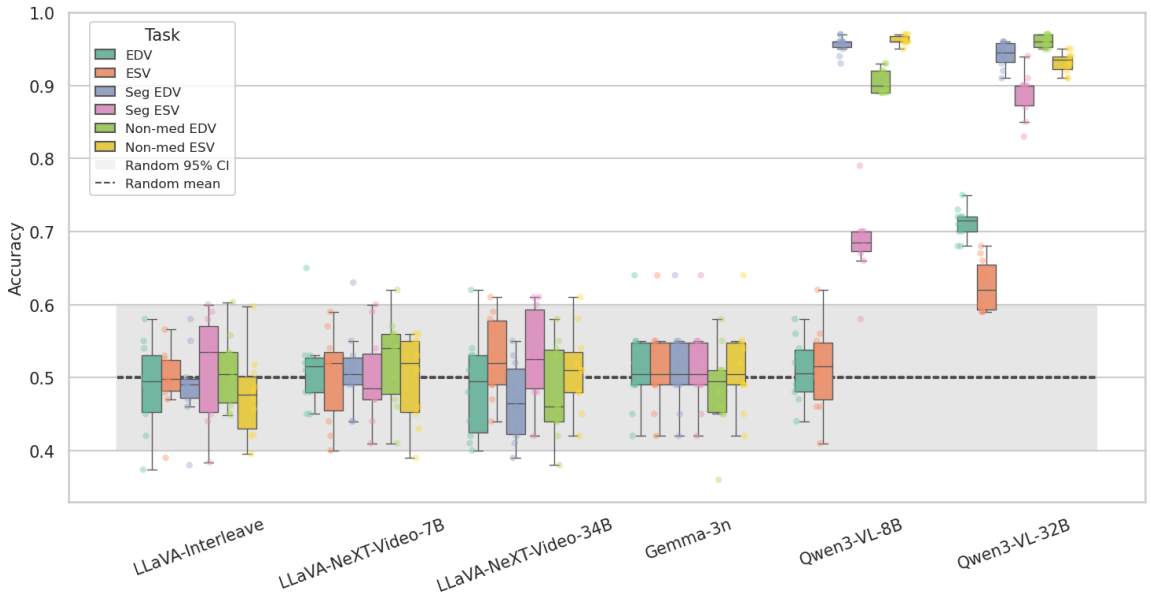


Figure 3: Accuracy of binary task T2

ful accuracy on this task, regardless of whether they were given original videos, segmentation overlays, or non-medical instructions.

Echocardiogram interpretation could become substantially more time-efficient if VLMs were able to assist with frame-level tasks such as EDV and ESV localization. For an experienced clinician, accurately identifying ED and ES and quantifying volumes typically takes up to a few minutes per study, with a frame-level variability on the order of 2 to 4 frames (Zolgharni et al., 2017). In contrast, all evaluated models exhibited localization errors exceeding 6 frames on average in our single-phase localization tasks. When we translated these frame errors into EF estimation, the best-case mean absolute error remained high and far from clinically acceptable thresholds. Prior work has shown that an error of just two to three frames in detecting ES can already elicit an approximate 10 percent difference in segmental ES strain (Mada et al., 2015), which underlines how demanding frame-level accuracy is for downstream functional assessment.

Our experiments further showed that naive structural cues did not solve the problem. Adding segmentation masks that highlighted the left ventricle did not systematically improve performance on frame localization and often worsened it. Removing medical context and asking models to focus solely on the red region likewise failed to yield robust improvements in T1. These findings suggest that current VLMs did not automatically exploit segmentation overlays as structured guidance at the level of precision required for cardiac ultrasound.

The auxiliary two-frame discrimination task provided additional insight into these limitations. Even when temporal context was removed, but to a direct comparison between two candidate frames, many models still performed close to random. Only the strongest models benefited consistently from segmentation cues in this simplified setting, achieving high accuracies when asked to focus on the masked region. This pattern indicated that failures on T1 were not solely due to long-range temporal reasoning but also reflected a difficulty in reliably interpreting pixel-level differences in medical video frames.

Taken together, these results point to a fundamental gap between general-purpose VLM capabilities and the requirements of frame-level echocardiogram understanding. General VLMs excel at semantic reasoning and coarse video understanding, but they are not trained to perform precise temporal localization of clinically relevant phases based on subtle changes in grayscale intensity. Bridging this gap will likely require models that incorporate explicit temporal modeling of cardiac cycles, tighter integration with segmentation and motion cues, and training on large-scale echocardiography data.

This work has several limitations. We focused on a single public dataset and a single view, so performance may differ on other acquisitions, institutions, or pathologies. We also evaluated only a fixed set of prompts and instruction formats and restricted our analysis to open-source models due to data usage constraints. Future work could explore prompt optimization, instruction tuning, and architectures tailored to periodic motion, as well as evaluations on curated datasets that can be shared with both open and closed source models.

Despite these limitations, our benchmark provides a concrete and clinically relevant stress test for VLMs in medical video analysis. It highlights that strong performance on general video understanding does not automatically translate into clinically meaningful frame-level accuracy in echocardiography. Our non-medical experiments further suggest that these failures are not only due to temporal reasoning or the specifics of medical im-

agery, but also reflect a broader limitation in how current VLMs handle fine-grained spatial information. We hope this work will motivate the development of medical VLMs with explicit temporal reasoning capabilities and encourage the community to design more targeted benchmarks for high-precision tasks in cardiology and beyond.

Acknowledgments

The authors would like to thank the Charité Scientific Computing group for infrastructure support. This work was supported by the German Ministry for Research, Technology and Space (BMFTR, junior research group "Medical Omics", 01ZZ2001).

References

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023.

Pierre Chambon, Christian Bluethgen, Curtis P. Langlotz, and Akshay Chaudhari. Adapting pretrained vision-language foundational models to medical imaging domains, 2022. URL https://arxiv.org/abs/2210.04133.

Linda D. Gillam and Leo Marcoff. Echocardiography: Past, present, and future. *Circulation: Cardiovascular Imaging*, 17(4):e016517, 2024. doi: 10.1161/CIRCIMAGING.124.016517. URL https://www.ahajournals.org/doi/abs/10.1161/CIRCIMAGING.124.016517.

Roberto M. Lang, Luigi P. Badano, Victor Mor-Avi, Jonathan Afilalo, Anderson Armstrong, Laura Ernande, Frank A. Flachskampf, Elyse Foster, Steven A. Goldstein, Tatiana Kuznetsova, Patrizio Lancellotti, Denisa Muraru, Michael H. Picard, Ernst R. Rietzschel, Lawrence Rudski, Kirk T. Spencer, Wendy Tsang, and Jens-Uwe Voigt. Recommendations for cardiac chamber quantification by echocardiography in adults: An update from the american society of echocardiography and the european association of cardiovascular imaging. *Journal of the American Society of Echocardiography*, 28(1):1–39.e14, 2015. ISSN 0894-7317. doi: https://doi.org/10.1016/j.echo.2014.10.003. URL https://www.sciencedirect.com/science/article/pii/S0894731714007457.

Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models, 2024. URL https://arxiv.org/abs/2407.07895.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, January 2024. URL https://llava-vl.github.io/blog/2024-01-30-llava-next/.

Razvan Mada, Peter Lysyansky, Ana Daraban, Jürgen Duchenne, and Jens-Uwe Voigt. How to define end-diastole and end-systole? *JACC Cardiovascular Imaging*, 01 2015. doi: 10.1016/j.jcmg.2014.10.010.

David Ouyang, Bryan He, Amirata Ghorbani, Neal Yuan, Joseph Ebinger, Curtis P. Langlotz, Paul A. Heidenreich, Robert A. Harrington, David H. Liang, Euan A. Ashley, and James Y. Zou. Video-based AI for beat-to-beat assessment of cardiac function. *Nature*, 580(7802):252–256, 2020. doi: 10.1038/s41586-020-2145-8.

R. P. Steeds. Echocardiography: frontier imaging in cardiology. *The British Journal of Radiology*, 84(Spec No 3):S237–S245, December 2011. ISSN 0007-1285. doi: 10.1259/bjr/77730594.

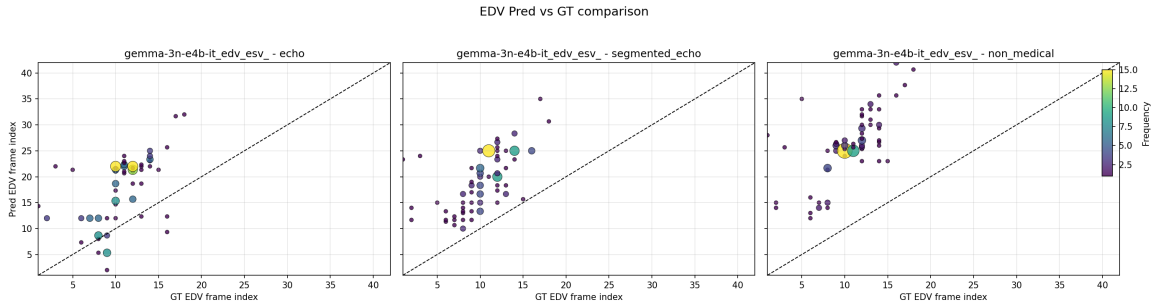
Gemma Team. Gemma 3n. 2025a. URL https://ai.google.dev/gemma/docs/gemma-3n.

Qwen Team. Qwen3 technical report, 2025b. URL https://arxiv.org/abs/2505.09388.

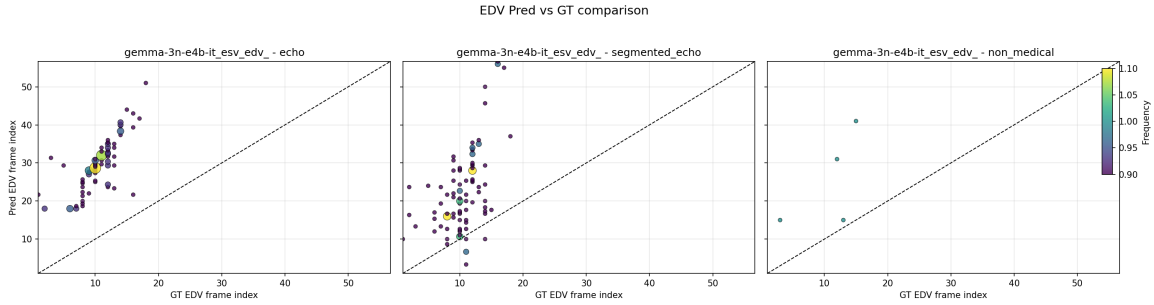
Yuanhan Zhang, Bo Li, haotian Liu, Yong jae Lee, Liangke Gui, Di Fu, Jiashi Feng, Ziwei Liu, and Chunyuan Li. Llava-next: A strong zero-shot video understanding model, April 2024. URL https://llava-vl.github.io/blog/2024-04-30-llava-next-video/.

M. Zolgharni, M. Negoita, N. M. Dhutia, M. Mielewczik, K. Manoharan, S. M. A. Sohaib, J. A. Finegold, S. Sacchi, G. D. Cole, and D. P. Francis. Automatic detection of end-diastolic and end-systolic frames in 2d echocardiography. *Echocardiography*, 34(7):956–967, July 2017. doi: 10.1111/echo.13587.

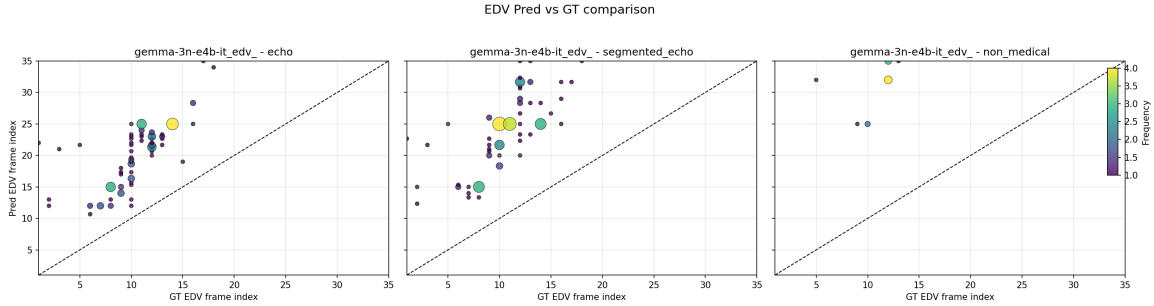
Appendix A. Plots of predicted vs ground truth index from Gemma-3n outputs



(*a*) EDV–ESV input

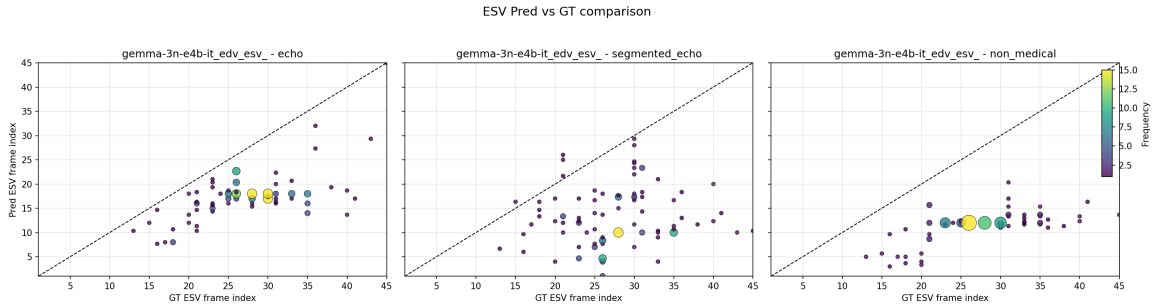


(*b*) ESV–EDV input

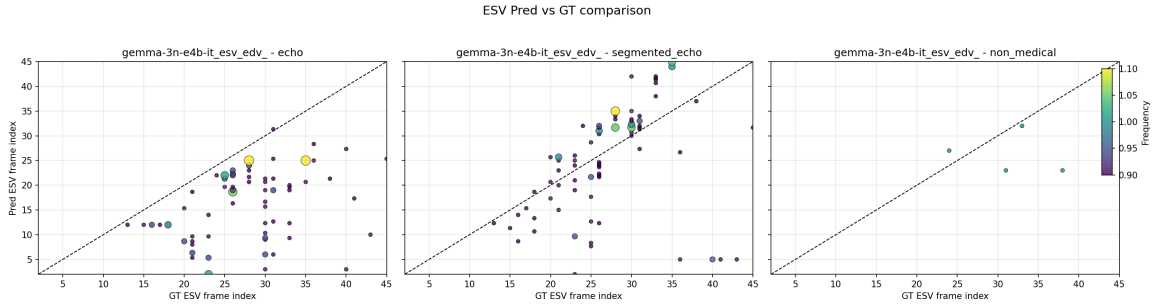


(*c*) ESV only

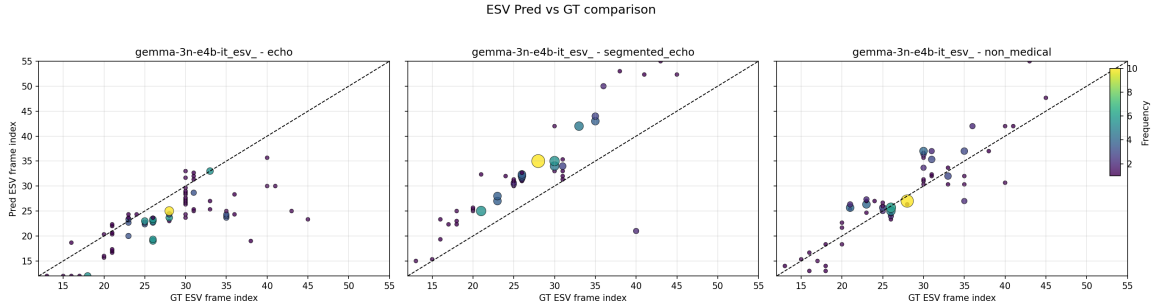
Figure 4: Additional EDV examples for Gemma-3n



(*a*) EDV–ESV input



(*b*) ESV–EDV input



(*c*) ESV only

Figure 5: Additional ESV examples for Gemma-3n