EchoNarrator: Generating natural text explanations for ejection fraction predictions

Sarina Thomas[1,2], Qing Cao[3], Anna Novikova[4], Daria Kulikova[4], and Guy Ben-Yosef[5,†]

[1] University of Oslo, Oslo, Norway
[2] Department of Cardiovascular Ultrasound, GE Healthcare, Oslo, Norway
[3] GE Healthcare, Wuxi, China
[4] GE Healthcare, Kharxiv, Ukraine
[5] GE HealthCare Technology & Innovation Center, Niskayuna, New York, USA
[†] corresponding author
guy.ben-yosef@gehealthcare.com

Abstract. Ejection fraction (EF) of the left ventricle (LV) is considered as one of the most important measurements for diagnosing acute heart failure and can be estimated during cardiac ultrasound acquisition. While recent successes in deep learning research successfully estimate EF values, the proposed models often lack an explanation for the prediction. However, providing clear and intuitive explanations for clinical measurement predictions would increase the trust of cardiologists in these models. In this paper, we explore predicting EF measurements with Natural Language Explanation (NLE). We propose a model that in a single forward pass combines estimation of the LV contour over multiple frames, together with a set of modules and routines for computing various motion and shape attributes that are associated with ejection fraction. It then feeds the attributes into a large language model to generate text that helps to explain the network's outcome in a human-like manner. We provide experimental evaluation of our explanatory output, as well as EF prediction, and show that our model can provide EF comparable to state-of-the-art together with meaningful and accurate natural language explanation to the prediction. The project page can be found at https://github.com/guybenyosef/EchoNarrator .

Keywords: Echocardiography · Graph Neural Networks · Explainable-AI. · Natural Language Explanation

1 Introduction

The release of the extensive Dynamic EchoNet echocardiography dataset [16] has accelerated the adoption of deep learning models for ejection fraction (EF) prediction and left ventricle (LV) contour delineation. Several innovative approaches have been introduced, including LV segmentation [16, 2, 12], direct video regression [16, 9, 18, 15], graph- and keypoints-based models [14, 23], and attention-based models [13] - Their potential is somewhat diminished by a common short-

fall: the lack of clinically meaningful explanations for the predicted EF. Explainability of visual deep learning models is often linked with activation maps such as Class Activation Mapping (CAM) [27] and Grad-CAM [21], which associate image regions with their contribution to the prediction. Although useful in certain contexts, these methods often fall short in medical imaging, where they may only highlight obvious regions such as the LV to predict EF, providing clinically correct but not meaningful information (an interesting example of this phenomenon was shown in [13]). To overcome these limitations and improve the explainability with human-like text, a novel approach has been developed within the subfield of *Natural Language Explanations* (NLE). This approach leverages advancements in vision-language and language models to generate text explanations that accompany model outcomes, providing context and clarity that activation maps cannot provide. NLE models include the generation of text explanations based on object attributes [4], models for associating text explanations with image regions [5], and language models such as GPT2 [20, 8] and GPT3 [19]. Considering that abnormal EF values are often linked to visible changes in the LV, providing explanations based on those visual cues shall enhance cardiologists' confidence in the deep learning predictions. Inspired by attribute-based NLE strategies (e.g., [4, 5]), where explanations are generated based on a predefined set of attributes, we create attributes that influence EF predictions, such as wall thickening in the interventricular septum (hereafter referred to as bulge), regional wall motion abnormalities, and foreshortening due to acquisition. Furthermore, we incorporate Large Language Models (LLMs) into our pipeline, utilizing models like LLaMA [24] as the final step to synthesize smooth and coherent explanations. Our innovation harnesses the capabilities of LLMs to assimilate the estimation of relevant LV attributes, generating text explanations that are informative and aligned with clinical practices. This method marks a significant step toward developing an AI assistant capable of engaging with clinicians through human-like language.

Our paper presents three major contributions to the field of cardiovascular ultrasound analysis and interpretation:

(1) *Novel NLE Model for EF Prediction:* We introduce the first NLE model tailored for EF prediction in cardiovascular ultrasound. This model synergizes the analytical depth of modern LLMs with spatiotemporal analysis of geometric features, setting a new benchmark for accuracy and explainability.
(2) *Self-Instruction Training Method:* We develop a novel training approach for the LLaMA model, utilizing GPT-4 to augment explanation examples. By releasing a dataset of approx. 800 self-instructions, we lay the groundwork for future advancements in training LLMs for echocardiography-related tasks.
(3) *Evaluation Metrics for Explanation Output:* Our research extends into the development and application of evaluation metrics specifically designed for assessing the quality of natural language explanations. We show that our model not only achieves precise EF predictions, but also generates explanations that are clinically relevant in a human-like language.

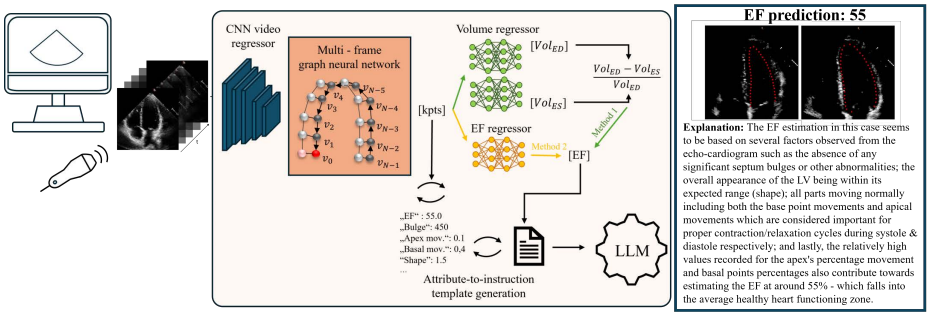

Fig. 1. Overview of the proposed pipeline A US video is fed into a CNN video encoder that outputs a feature representation. The features are passed to a spatio-temporal GCN that returns keypoints for ED and ES. The keypoints serve as 1) input for MLPs that regress the LV volumes (in green) or the EF directly (in orange) and 2) computation of geometrical attributes that are converted into text snippets that can be parsed into an LLM. The LLM provides a human-like explanation for the EF.

2 Methods

Our approach introduces a streamlined pipeline that enriches a GCN-based EF prediction with clinically meaningful explanations. With echocardiography videos as input, a video encoder extracts feature representations, then fed to a spatio-temporal Graph Convolutional Network (GCN) which identifies anatomical keypoints. From these keypoints, the EF is predicted along with geometrical attributes essential for our text generator model, which produces natural language explanations of the EF predictions. An overview is shown in Fig. 1.

2.1 Multi-frame GCN Model for EF Prediction in a Single Pass

Central to our pipeline is the multi-frame Graph Convolutional Network (GCN) model, engineered to perform EF prediction and contour detection in a single integrated operation. Previous work [23] designed a multi-task multi-frame model that derived the EF value directly from the input encoder. That approach resulted in EF and keypoints prediction being disentangled and less interpretable. Since manual EF computation relies solely on the contours, we modified the architecture to ensure that EF prediction follows the mathematical concept of the ratio between both keypoints volumes. By adding two volume regressors and then fusing the results, we base the model on prior knowledge about the dynamics rather than relying purely on black box predictions. We explored different levels of adding prior knowledge, either by directly regressing the EF from keypoints versus having two separate volume regression branches.

2.2 Attribute generation

Beyond automatically predicting LV contours, our model also derives attributes that reflect structural changes and temporal deviations due to pathology or

acquisition which affect the EF values. Based on clinical insights, we developed geometrical processing routines to compute attributes from LV contour points, capturing the intuition of cardiologists in EF assessment. This section outlines the attributes, grounded in clinical feedback, along with the computation.

Septal bulge: A septal bulge can be a morphological sign for early hypertensive heart disease [3] and is an asymmetric, localized thickening of the basal-to-mid part of the inter-ventricular septum. It could be detected by calculating a wall thickness ratio over 1.4 [11]. We compute a bulge using the LV convexity which is the distance between a convex hull and the true contour. Ground truth contours were visually inspected while three manual thresholds were set to distinguish prominent bulges from undetected convexity calculated using OpenCV[6].

Segment motion: The 17-segment model[1] is widely used for regional wall motion analysis in multiple views. To simplify the process for 4CH-view, we divide the contour into 7 distinctive segments and calculate the segment movement direction relative to the overall motion as well as the vertical basal movement.

Apex movement: Foreshortening is a common problem in 2D echocardiography which results in underestimating the LV volume and inaccurate EF estimation. A foreshortened apex translates throughout the cycle, whereas a true apex almost remains at the same point. Following the approach of [22], we compute the apex movement in the direction of the LV long axis. Based on the distribution in the dataset we set a threshold to indicate suspicious apex movement.

Length-width ratio: A normal LV has a bullet-like shape. Cardiovascular diseases such as hypertension or heart failure may change the LV shape despite age and gender being also effecting factors. We decided to use the length-width ratio as a shape measure which is computed by dividing the apex-basal distance by the horizontal mid-septal distance. The length-width ratio is typically around **2**, as observed during our experiments, while in cases with dilated LV, resulting in a reduced length-width ratio.

Sector intersection: One important requirement for manual and automatic EF computation is to ensure full visibility of the LV within the ultrasound sector. Therefore, we calculate the ratio of the intersection with the detected LV contours as a quality metric for the EF computation.

Image quality: Image quality will affect the visibility of the LV contours and the wall movement which influences the EF estimation. We calculate the intensity difference between the LV cavity and the myocardial wall as a measure of their contrast. A higher contrast indicates an improved visibility. However, this metric, though practical, does not cover all dimensions of image quality.

All of the aforementioned attributes are clustered based on their distribution in the annotated dataset. Thresholds were defined for each attribute to create text templates that could be fed into a language model.

[6] www.opencv.org

2.3 Generation of text explanations

The ability to translate complex geometrical and spatiotemporal attributes into understandable natural language explanations is critical to bridge the gap between advanced echocardiographic analysis and clinical practice. Our approach encompasses two primary phases: (1) converting the attributes into basic text sentences that describe the underlying clinical observations, and (2) refining these basic sentences into coherent natural language explanations suitable for clinical use, leveraging the capabilities of a Large Language Model (LLM).

From attribute values to LLM inputs. Computed attributes are numerical values that need to be converted into text tokens digestable for the LLM. The first phase involves translating the list of visual and geometric attributes into **basic sentences** by using thresholds. For instance, the numerical value $bulge = 500$ is translated to "A bulge value of 500 means that there is no bulge".

Natural Language Refinement with LLM. In the second phase, we employ the LLaMA model [24], a LLM variant, and train it for generating medical text specifically. We fine-tune LlaMA on a dataset with clinical explanations to ensure that the generated text aligns with clinical terminology and reasoning. To solve the limited availability of expert-generated explanations, we further implement some data augmentation strategies to enrich the training dataset.

Synthetic Explanations. By adding prior clinical knowledge, we build more elaborated sentences from the basic sentences as synthetic expert explanations. These sentences articulate the clinical significance of each attribute in a structured format. For instance, an attribute indicating a significant septal bulge would be converted into a more elaborated sentence like "There is a large septal bulge, which may adversely affect the EF."

Data Augmentation through Self-instruction. We adopted a self-instruction method [25] using the GPT-4 model to augment a small initial dataset containing experts explanations. By feeding 5 expert explanations into GPT-4, along with a chain-of-thought prompt [26] that includes examples of the input (basic sentences) and the desired output (expert explanation), we instruct GPT-4 to simulate medical expert explanations for novel sets of basic text. This use of chain-of-thought processing with GPT-4 effectively enlarges our dataset towards a ten times expansion of the initial set.

This dual process ensures that our model not only accurately identifies the visual and geometrical attributes indicative of specific cardiac conditions, but also communicates findings in a way that clinicians can immediately interpret.

2.4 A novel metric to evaluate the EF explanation via LLMs

Evaluating unstructured text is crucial to identify errors in clinical LLM, yet human evaluation is time consuming and potentially subjective, highlighting the need for automated metrics. However, initial experiments indicated that even simple adversarial examples could deceive most of the existing metrics for sentence similarity. Given that the output of the proposed LLMs is unstructured text focused on key attributes, we designed a metric specifically aimed at

assessing factual correctness. To accomplish this, we use the recently released Mistral model [7], another and faster LLM variant. By creating nine targeted prompts with instructions and one-shot context for Mistral, we evaluate whether attributes appear in the text as positive (pathological), negative (normal), or unspecified. This allows a comparison between ground truth and prediction beyond mere textual similarity. In cases where an attribute is unspecified, its status is considered normal. We quantify the performance by reporting the accuracy, the count of true contradictions, hallucinations, and of missing attributes.

3 Experiments

3.1 Data

Dataset: We use the EchoNet-Dynamic dataset [16], which contains 10,030 echocardiography videos from healthy and pathological patients. Each video is annotated with 40 LV contour points, one basal and apex point at the end-diastolic (ED) and end-systolic (ES) frame, along with the EF. The training, validation and test splits provided by EchoNet are used for benchmarking.

Annotations: We trained the GCNs on the annotated keypoints from the EchoNet dataset, employing a multi-frame strategy in [23]. To simplify the processing, we selected ED and ES frames and sampled 14 evenly spaced intermediate frames from each video. For our experiments, ED and ES frames were assumed to be known as they can be computed from the ECG or automatically. A lack of Electronic Health Records (EHR) prompted us to enlist two cardiology experts who annotate a subset of the EchoNet data with video-text pairs. The experts watched the videos and provided text descriptions including an EF assessment and reasoning, focusing on attributes like LV shape, wall movement, and bulge presence. They were allowed to use different structures or description formats to ensure a diverse text reflection of real-world scenarios. 89 image-text pairs were generated for training, with additional 48 pairs designated for testing.

3.2 NLE evaluation metrics

We incorporated several different evaluation metrics to evaluate our LLM outputs from different aspects utilizing the advantages of each. We exploit the **ClinicalBERT** [6] and Sentence-based BERT models (**sBERT**) [17] that generate single embeddings either per word or per sentence followed by cosine similarity and are pre-trained on clinical texts. Additional text similarity models are provided in the suppl. material. The Mistral score was introduced to evaluate explanations against specific clinical attributes, leveraging a recently developed **Mistral LLM** [7] tailored for this purpose. This custom metric (sec. 2.4), aims to provide a more nuanced assessment of the clinical relevance and accuracy of the explanations generated. In addition to accuracy metrics, we also used the Flesch Reading Ease score to measure contextual richness of the explanation (a lower Flesch score means contextually richer text).

3.3 EF predictions with generated explanations

Implementation: Our GCN model uses a ResNet-3D-18 as video encoder backend, optimizing for the efficient processing of echocardiography videos. GCN model training focuses on accurately predicting LV keypoints, which are crucial for the subsequent estimation of EF and the generation of explanations. For the NLE component, we train the LLaMA model to generate clinically relevant explanations based on attributes derived from the GCN output. LLM training included a low-rank adaptation on the LLaMA-1 model through 8-bit quantization, with a learning rate of 0.0003 and a batch size of 32. Training was performed for 5 hours on two A6000 GPUs, each equipped with 48GB memory. The end-to-end inference pipeline ensures a seamless transition from raw video data to EF predictions accompanied by understandable explanations. For this experiment, the GCN with the lowest mean absolute error (MAE) was used. Detailed implementation specifics are available on GitHub for reproducibility[7].

Results: In evaluating our end-to-end inference system, we focus on both the EF prediction accuracy and the quality of the generated explanations. While competitors exist for EF prediction, our approach is unique in integrating NLE, setting a benchmark in the field. Tab. 1 lists the details of our comparison with previous models using the Dice score, the mean keypoint error (MKE) and the accuracy of the prediction of EF. We show that our single task NLE GCN can reach a lower MKE than the EchoGraph while maintaining the same EF accuracy. Predicting from volumes (Vol) versus predicting directly from keypoints (EF) performed similarly well. For NLE prediction, we evaluate the coherence and clinical relevance of the generated explanations quantitatively (Tab. 2) and qualitatively (Fig. 2). In Tab. 2, we added a prediction of the LLaVA-Med model [10] when we input an image showing ED and ES frames, followed by the instruction to explain the EF (see suppl. material for details). Although different from our model in the design and goal, it seems to be the closest in providing baseline text explanations.

Analysis: Our results demonstrate notable accuracy in EF prediction (Tab. 1) combined with the generation of clinically meaningful explanations. Compared to competitors in EF predictions, the scores in Tab. 2 further show that our system not only achieves good EF estimation, but also introduces the capability of generating accurate, complete, and human-like explanations. Tab. 2 also shows that predicted explanations are better from simple baselines. The enhancement of NLE prediction in our model can be attributed to the use of synthetic data and data augmentation techniques (Sec. 2.3). The utilization of the self-instruction by using Chain-of-Thought in GPT4 further refines the model's capability to generate plausible and contextually more rich explanations that meet clinical expectations, as shown by its lower scores in the Reading Ease metric.

[7] https://github.com/guybenyosef/EchoNarrator
[8] Means random answers to the attributes from sec.2.2

Model	Frames	Dice (%)	MKE (px)	EF MAE	Cycle	Explainable
EchoNet [16]	32	91.7 ± 4.2	2.5 ± 1.2	4.22	Single	No
EchoGraphs [23]	16	90.3 ± 4.3	2.7 ± 1.5	4.01	Single	No
EchoCoTr-S [15]	36	N/A	N/A	3.95	Multi	No
GEMTrans [13]	16	N/A	N/A	4.15	Multi	Heatmaps
NLE EF GCN (Ours)	16	91.5 ± 4.3	2.4 ± 1.1	4.00	Single	Text
NLE Vol GCN (Ours)	16	91.4 ± 4.4	2.4 ± 1.1	4.05	Single	Text

Table 1. Segmentation and EF accuracy for different methods evaluated on EchoNet testset (n=1264) and annotated ED and ES frames (MKE = mean L1 keypoint pixel error). *Multi* refers to whole videos whereas *single* refers to one cycle or ED to ES.

Model	mistral acc↑	halluc.↓	contradict.↓	missing↓	cbert↑	msbert↑	Flesch↓
random[8]	0.45						
LLaVA-Med	0.67	1.49	0.87	1.49	0.92	0.95	54.92
NLE-EF-13B	0.77	0.54	1.02	1.50	0.93	0.96	58.59
NLE-EF-7B	0.77	0.65	0.67	1.44	0.94	0.95	57.20
NLE-EF-13B self-i	0.77	0.44	1.13	1.65	0.91	0.95	30.91
NLE-EF-7B self-i	0.80	0.50	0.90	1.33	0.92	0.95	27.42

Table 2. Evaluation of the NLE-EF versions on 48 samples from the EchoNet test set using different averaged metrics such as mistral score accuracy, average number of mistral contradictions, hallucinations and missing attributes, cbert score (clinical BERT), msbert score (sentence medsBERT), and Flesch reading ease.

4 Discussion and conclusion

Our method introduces an innovative approach by leveraging a GCN and a LLM (LLaMA) to provide LV contours and EF along with geometrical features and a text explanation. The main contribution is the integration of cardiac features (potentially less intuitive for humans) derived from a vision model with an LLM that translates these features into explanatory text. Considering that the primary focus was on the effective combination of these components to enhance interpretability, architectural choices were based on experiments. Leveraging synthetic and augmented data can improve interpretability without compromising prediction accuracy. This balance is vital for wider clinical adoption, where the clarity of the explanations is as important as the accuracy. Our evaluation with another LLM aims to increase sensitivity to contradictions while configurations and prompt design need to be considered carefully. Despite notable successes, we acknowledge limitations such as a relatively small dataset, noisy labels and prompts, which could affect our findings' robustness and generalizability. We incorporated six widely used LV attributes, but clinical feedback suggested extending this to include the right side of the heart. Although GCN and LLM pre-training add more data implicitly, an extension of the dataset, including more attributes, and a clinical evaluation will be future work. The proposed method facilitates AI-assisted diagnosis, reporting, and education by providing cardiologists an accurate visual output with human-readable explanations.

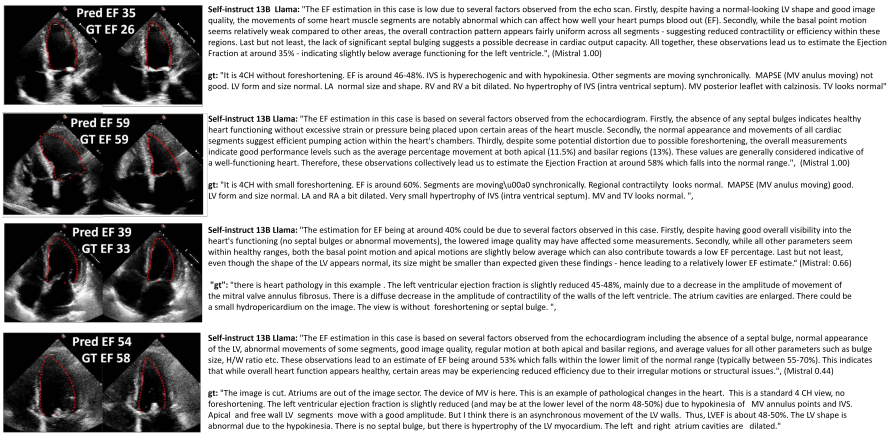

Fig. 2. (Zoom in for optimal view) LV contour estimation, EF prediction and its text explanation as provided by the NLE-EF-13B self-instruct on EchoNet test examples.

Acknowledgments. This work was financially supported by the Research Council of Norway (RCN), through an innovation project (EchoAI, 313756) and its Centre for Research-based Innovation (Visual Intelligence, 309439). We thank E. Steen, J. Sprem and S. A. Aase for their guidance and feedback.

Disclosure of Interests. The authors declare no further competing interests. A patent has been filed for the methods and technologies described in this document.

References

1. Cerqueira, M., Weissman, N., Dilsizian, V., Jacobs, A., Kaul, S., Laskey, W., Pennell, D., Rumberger, J., Ryan, T., et al.: Standardized myocardial segmentation and nomenclature for tomographic imaging of the heart: a statement for healthcare professionals from the cardiac imaging committee of the council on clinical cardiology of the american heart association. Circulation **105**(4), 539–42 (2002)
2. Dai, W., Li, X., Ding, X., Cheng, K.T.: Cyclical self-supervision for semi-supervised ejection fraction prediction from echocardiogram videos. IEEE Transactions on Medical Imaging **42**(5), 1446–61 (2023)
3. Gaudron, P.D., Liu, D., Scholz, F., Hu, K., Florescu, C., Herrmann, S., Bijnens, B., Ertl, G., Störk, S., Weidemann, F.: The septal bulge - an early echocardiographic sign in hypertensive heart disease. Journal of the Am. Society of Hypertension **10**, 70–80 (2016)
4. Hendricks, L.A., Akata, Z., Rohrbach, M., Donahue, J., Schiele, B., Darrell, T.: Generating visual explanations. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14. pp. 3–19. Springer (2016)
5. Hendricks, L.A., Hu, R., Darrell, T., Akata, Z.: Grounding visual explanations. In: Proc. of the European conference on computer vision (ECCV). pp. 264–79 (2018)

6. Huang, K., Altosaar, J., Ranganath, R.: Clinicalbert: Modeling clinical notes and predicting hospital readmission. arXiv preprint arXiv:1904.05342 (2019)
7. Jiang, A.Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D.S., Casas, D.d.l., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., et al.: Mistral 7b. arXiv preprint arXiv:2310.06825 (2023)
8. Kayser, M., Emde, C., Camburu, O.M., Parsons, G., Papiez, B., Lukasiewicz, T.: Explaining chest x-ray pathologies in natural language. In: Medical Image Comp. and Computer Assisted Intervention–MICCAI 2022: 25th Int. Conference, Singapore, September 18–22, 2022, Proceedings, Part V. pp. 701–13. Springer (2022)
9. Kazemi Esfeh, M.M., Luong, C., Behnami, D., Tsang, T., Abolmaesumi, P.: A deep bayesian video analysis framework: towards a more robust estimation of ejection fraction. In: Int. Conf. on Medical Image Comp. and Computer-Assisted Intervention. pp. 582–90. Springer (2020)
10. Li, C., Wong, C., Zhang, S., Usuyama, N., Liu, H., Yang, J., Naumann, T., Poon, H., Gao, J.: Llava-med: Training a large language-and-vision assistant for biomedicine in one day. Adv. in Neural Information Processing Systems **36** (2024)
11. Marciniak, M., Gilbert, A., Loncaric, F., Fernandes, J.F., Bijnens, B., Sitges, M., King, A., Crispi, F., Lamata, P.: Septal curvature as a robust and reproducible marker for basal septal hypertrophy. Journal of hypertension **39**(7), 1421 (2021)
12. Meng, Y., Zhang, Y., Xie, J., Duan, J., Zhao, Y., Zheng, Y.: Weakly/semi-supervised left ventricle segmentation in 2d echocardiography with uncertain region-aware contrastive learning. In: Liu, Q., Wang, H., Ma, Z., Zheng, W., Zha, H., Chen, X., Wang, L., Ji, R. (eds.) Pattern Recognition and Computer Vision. pp. 98–109. Springer Nature Singapore, Singapore (2024)
13. Mokhtari, M., Ahmadi, N., Tsang, T.S.M., Abolmaesumi, P., Liao, R.: Gemtrans: A general, echocardiography-based, multi-level transformer framework for cardiovascular diagnosis. In: Machine Learning in Medical Imaging. pp. 1–10. Springer Nature Switzerland, Cham (2023)
14. Mokhtari, M., Tsang, T., Abolmaesumi, P., Liao, R.: Echognn: Explainable ejection fraction estimation with graph neural networks. In: Int. Conf. on Medical Image Comp. and Computer-Assisted Intervention. pp. 360–69. Springer (2022)
15. Muhtaseb, R., Yaqub, M.: Echocotr: Estimation of the left ventricular ejection fraction from spatiotemporal echocardiography. In: Int. Conf. on Medical Image Comp. and Computer-Assisted Intervention. pp. 370–79. Springer (2022)
16. Ouyang, D., He, B., Ghorbani, A., Yuan, N., Ebinger, J., Langlotz, C.P., Heidenreich, P.A., Harrington, R.A., Liang, D.H., Ashley, E.A., et al.: Video-based ai for beat-to-beat assessment of cardiac function. Nature **580**(7802), 252–56 (2020)
17. Rasmy, L., Xiang, Y., Xie, Z., Tao, C., Zhi, D.: Med-bert: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. NPJ digital medicine **4**(1), 86 (2021)
18. Reynaud, H., Vlontzos, A., Hou, B., Beqiri, A., Leeson, P., Kainz, B.: Ultrasound video transformers for cardiac ejection fraction estimation. In: Int. Conf. on Medical Image Comp. and Computer-Assisted Intervention. pp. 495–505. Springer (2021)
19. Sammani, F., Deligiannis, N.: Uni-nlx: Unifying textual explanations for vision and vision-language tasks. In: VLAR, International Conference on Computer Vision Workshops (ICCVW) 2023. vol. Workshop, pp. 1–4. IEEE (2023)
20. Sammani, F., Mukherjee, T., Deligiannis, N.: Nlx-gpt: A model for natural language explanations in vision and vision-language tasks. In: Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition. pp. 8322–32 (2022)

21. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proc. of the IEEE Int. conference on computer vision. pp. 618–26 (2017)
22. Smistad, E., Østvik, A., Salte, I.M.M., Melichova, D., Nguyen, T.M., Haugaa, K., Brunvand, H., Edvardsen, T., Leclerc, S., Bernard, O., Grenne, B.B., Løvstakken, L., Ostvik, A., Salte, I.M.M., Melichova, D., Nguyen, T.M., Haugaa, K., Vrunvand, H., Edvardsen, T., Leclerc, S., Bernard, O., Grenne, B.B., Lovstakken, L.: Real-time automatic ejection fraction and foreshortening detection using deep learning. IEEE Trans. Ultrason. Ferroelectr. Freq. Control **67**(12), 2595–2604 (2020)
23. Thomas, S., Gilbert, A., Ben-Yosef, G.: Light-weight spatio-temporal graphs for segmentation and ejection fraction prediction in cardiac ultrasound. In: Int. Conf. on Medical Image Comp. and Computer-Assisted Intervention. pp. 380–90. Springer (2022)
24. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al.: Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971 (2023)
25. Wang, Y., Kordi, Y., Mishra, S., Liu, A., Smith, N.A., Khashabi, D., Hajishirzi, H.: Self-instruct: Aligning language model with self generated instructions. arXiv preprint arXiv:2212.10560 (2022)
26. Wei, J., Wang, X., Schuurmans, D., Bosma, M., hsin Chi, E.H., Xia, F., Le, Q., Zhou, D.: Chain of thought prompting elicits reasoning in large language models. ArXiv **abs/2201.11903** (2022)
27. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. In: Proc. of the IEEE conference on computer vision and pattern recognition. pp. 2921–29 (2016)