LEARNING WHERE TO LOOK: ANATOMICAL GROUNDING AND GUIDED ATTENTION FOR CARDIAC MRI VISION–LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Cardiac magnetic resonance imaging (CMR) enables assessment of cardiac anatomy, ventricular function, and myocardial tissue characteristics. Clinicians interpret these images by identifying cardiac structures and focusing on the regions relevant to each clinical question, motivating anatomically guided vision–language models (VLMs). Yet CMR-specific supervision for anatomical localisation and clinical question answering remains limited. To address this gap, we investigate fine-grained CMR visual question answering through anatomical grounding and guided attention. We construct 128,915 anatomical-grounding and 42,799 clinical QA pairs across short-axis cine, late gadolinium enhancement, and long-axis cine. These datasets support anatomical recognition, localisation, and clinical assessment without requiring paired reports for individual training images. To help the model *learn where to look*, we introduce Cardiac Anatomy-Routed Attention (CARA), which selects predicted anatomical priors according to the question and guides decoder attention with learned task-specific strengths. Combining anatomical grounding pretraining with CARA yields our model, CARA-VL. Experiments demonstrate CARA-VL's strengths in clinical assessment and regional localisation across CMR imaging settings, with promising generalization to an external clinical cohort. Together, our data and method provide a practical framework for studying and advancing cardiac visual understanding in VLMs. We will release the QA data derived from public datasets upon publication.

1 INTRODUCTION

Cardiac magnetic resonance imaging (CMR) enables comprehensive assessment of cardiac anatomy, ventricular function, and myocardial tissue characteristics (Schulz-Menger et al., 2020). Clinicians interpret CMR by first identifying the relevant cardiac structures and then assessing their morphology, motion, and tissue characteristics for the diagnosis. For example, hypertrophy is assessed from myocardial wall thickness, ventricular function from changes across cardiac phases, and myocardial injury from the presence, extent, and transmurality of scar on late gadolinium enhancement (LGE) (Kim et al., 2000; Schulz-Menger et al., 2020). However, reliable CMR interpretation requires specialist expertise and often time-consuming quantitative analysis. This motivates cardiac visual question answering (VQA), in which AI models answer clinically relevant questions about cardiac anatomy, function, and tissue characteristics from CMR images.

Recent medical vision–language models (VLMs) have shown strong capability in medical image understanding and VQA through large-scale multimodal pretraining and medical instruction tuning (Li et al., 2023; Moor et al., 2023; Chen et al., 2024; Sellergren et al., 2026). Yet their visual training is largely based on broad online biomedical image–text corpora and commonly used medical imaging datasets, with comparatively limited supervision on cardiac MRI. This specialization gap can lead to poor generalization on fine-grained CMR tasks (O'Sullivan et al., 2026). Motivated by this gap, cardiac-specific vision–language models have recently begun to emerge, using CMR–report pairs, cardiac-specific pretraining, or modality-specialized experts to improve domain adaptation (Nakashima et al., 2026; Shad et al., 2026; O'Sullivan et al., 2026; Qu et al., 2026). Despite these advances, accessible training data and explicit anatomical grounding remain important challenges. Many cardiac VLMs rely on institutional or non-public CMR cohorts, limiting access to their train-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

ing data and reproducibility. Moreover, clinical answer supervision alone does not explicitly teach models to localise cardiac structures or guide attention towards the anatomy relevant to each question.

To address these limitations, we introduce an anatomically grounded vision–language model for fine-grained CMR VQA, connecting anatomical recognition and localisation with clinical assessment. We construct complementary anatomical-grounding and clinical VQA datasets from public and in-house CMR resources. Expert segmentation annotations provide 128,915 grounding QA pairs, while quality-controlled image-derived measurements yield 42,799 clinical QA pairs across 14 tasks informed by clinical report analysis. These datasets support learning about cardiac structure, function, and myocardial scar without requiring a paired report for each training image. Clinicians direct their attention to different cardiac structures depending on the clinical question, such as the myocardium for hypertrophy and the ventricular cavities for size and function (Schulz-Menger et al., 2020). Inspired by this practice, we introduce **Cardiac Anatomy-Routed Attention (CARA)** to help the model *learn where to look*. CARA predicts soft anatomical occupancy maps from visual features and selects the relevant map for each clinical question. The resulting prior guides decoder attention over visual tokens with a learned task-specific strength (Figure 2). We refer to the resulting model, trained with anatomical grounding and CARA, as CARA-VL. Our main contributions are summarised as follows:

- We systematically investigate cardiac visual understanding in vision–language models through fine-grained CMR VQA, examining how anatomical recognition and localisation support clinical assessment of cardiac structure, function, and myocardial scar.
- We construct 128,915 anatomical-grounding and 42,799 clinical QA pairs across SAX cine, LGE, and LAX cine. Report-guided task selection, segmentation annotations, and quantitative measurements provide complementary supervision without requiring paired reports for individual training images.
- We develop CARA-VL, combining anatomical grounding pretraining with CARA to learn where to look for clinical question answering. CARA routes predicted anatomical priors according to the question and guides decoder attention with learned task-specific strengths.
- Experiments across 14 clinical tasks show that CARA-VL outperforms the evaluated baselines on all internal tasks and demonstrates promising external generalization. Ablations support the complementary benefits of anatomical grounding pretraining and CARA.

2 Clinically grounded QA dataset construction

We construct complementary anatomical grounding and clinical QA datasets (Figure 1). Below, we describe report-guided task selection, image collection and standardisation, and the construction of anatomical grounding and clinical QA pairs.

2.1 Report-guided clinical QA task specification

We define clinical QA tasks through report standardisation, finding-level frequency analysis, and clinically informed selection. The corpus comprises 3,215 de-identified CMR reports from three sources. Preprocessing standardises the findings and impression sections and groups synonymous clinical expressions. We quantify the report-level frequency of each clinical finding, distinguishing normal from abnormal mentions to identify candidate tasks. The final selection considers clinical relevance, the availability of explicit measurement or grading criteria, and whether the available data and analysis tools support reliable reference answers. The resulting 14 clinical QA tasks SAX cine, LGE, and LAX cine. Eight SAX cine tasks assess hypertrophy, its segmental distribution, left ventricular (LV) size, right ventricular (RV) size, ejection fraction (EF), regional wall motion, global radial strain (GRS), and global circumferential strain (GCS). Three LGE tasks assess scar presence, transmurality, and extent, while three LAX cine tasks assess EF, end-diastolic volume (EDV), and global longitudinal strain (GLS). Each task has a predefined answer set and explicit assessment criteria, as summarised in Section 2.4 and more details are provided in Appendix C.

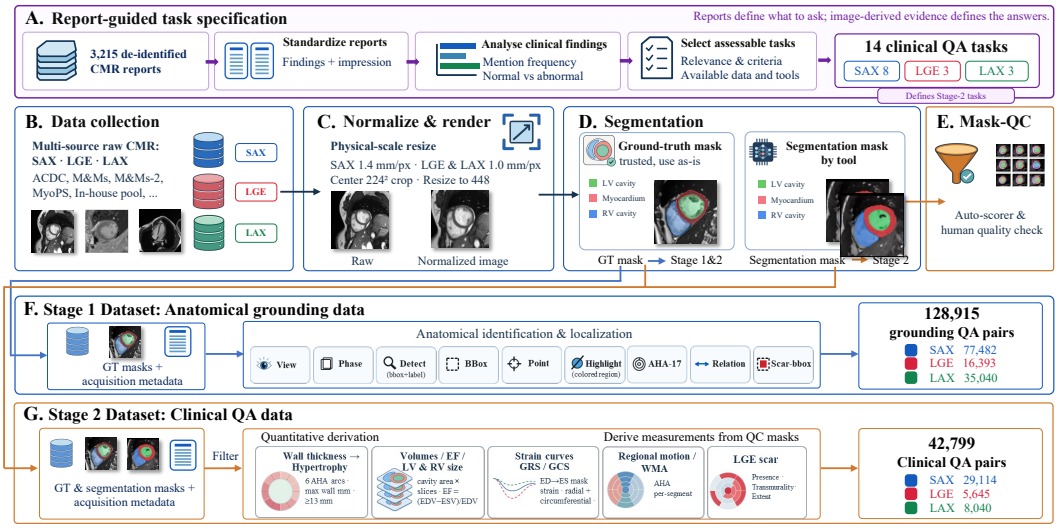


Figure 1: CMR VQA dataset construction pipeline. (A) Clinical report analysis guides the selection of 14 assessable clinical QA tasks. (B–E) Multi-source CMR images are standardised and paired with expert or quality-controlled automatic masks. (F) Expert masks and acquisition metadata generate nine types of anatomical-grounding QA for identifying and localising cardiac structures. (G) Image-derived measurements and task-specific decision rules generate clinical QA on cardiac structure, function, and myocardial scar.

2.2 Image Data Collection and Standardization

We combine public datasets and in-house CMR cohorts to construct large-scale anatomical grounding and clinical QA datasets for pretraining, clinical adaptation, and evaluation. The collection covers SAX cine, LGE, and LAX cine. Public sources include ACDC (Bernard et al., 2018), M&Ms, M&Ms-2 (Campello et al., 2021; Martín-Isla et al., 2023), MyoPS-related cohorts (Zhuang, 2018; Qiu et al., 2023; Ding et al., 2023; 2025), EMIDEC (Lalande et al., 2020), CMR-MULTI (Qu et al., 2026), and the Kaggle Data Science Bowl Challenge (Data Science Bowl, 2015). The public datasets used here do not provide paired clinical reports. Our in-house data comprise an image-only collection (In-house A) used for training, and an image–report cohort (In-house B) reserved exclusively for external evaluation. To standardize the physical scale of images within each imaging modality, we resample SAX cine to 1.4 mm and LGE and LAX cine to 1.0 mm in-plane resolution using acquisition-specific pixel spacing. Images and available masks are center-cropped to 224×224 pixels. The images are then upsampled to 448×448 for model input, with grounding coordinates scaled accordingly. Available ground-truth segmentation masks are used directly for data generation. For images without reference masks, we obtain segmentations from a deployed internal CMR segmentation pipeline and manually review their quality before clinical QA construction. Details are provided in Appendix C.

2.3 Stage-1: Anatomical Grounding Dataset

Before clinical QA adaptation, we first pretrain the VLM to recognize and localize cardiac anatomy as a structural foundation for subsequent reasoning, consistent with recent evidence on anatomical grounding in medical VLMs (Gu et al., 2026; Zhang et al., 2026). Accordingly, from the multi-source CMR pool described in Section 2.2, we select expert-annotated cases for anatomical grounding QA, covering 878 SAX patients, 440 LAX series, and 2,241 LGE slices. We restrict Stage-1 to ground-truth segmentation masks, allowing all grounding targets to be derived deterministically from the annotations. The resulting nine QA task types cover imaging context, anatomical identification and localisation, and cardiac-specific spatial reasoning (Table 6). Specifically, view and phase establish imaging context; detection, bounding-box, point, highlighted-region, and scar bounding-box tasks provide structure- and region-level grounding; and segment and relation tasks supervise identification of American Heart Association (AHA) myocardial segments (e.g., anteroseptal and

inferolateral segments) and LV–RV spatial relationships, respectively. Together, these tasks teach the model what structures are present, where they are located, and how they relate spatially. The final Stage-1 dataset contains 128,915 QA pairs over 45,773 images, including 77,482 SAX cine, 35,040 LAX cine, and 16,393 LGE pairs. Further details are provided in Appendix C.

2.4 STAGE-2: CLINICAL QA DATASET

Stage-2 extends anatomical grounding to clinical assessment through the 14 tasks defined in Section 2.1. We construct reference answers from anatomical measurements, available reference volumes, and feature-tracking outputs, enabling clinical QA generation from images without paired reports. These quantities capture wall thickness, chamber size, ventricular function and deformation, and myocardial scar. Task-specific decision rules map them to binary findings, severity grades, or regional labels. All masks used for answer generation undergo manual quality control, with low-quality or unreliable masks excluded before deriving clinical measurements and reference answers. Each accepted label is paired with a question template specifying the imaging context and a canonical answer. Depending on the task, the input contains a single image or an ordered end-diastolic (ED) and end-systolic (ES) pair. The model receives the images and question, without mask overlays or reference measurements. The resulting dataset contains 42,799 QA pairs: 29,114 from SAX cine, 5,645 from LGE, and 8,040 from LAX cine. Patient-level exclusions are shared across tasks, views, and both training stages, with the image–report cohort reserved for external evaluation. Details of measurement construction, grading, review, and sampling are provided in Appendix C.5; split definitions and evaluation references are given in Appendix C.6.

3 CARA: CARDIAC ANATOMY-ROUTED ATTENTION

We present Cardiac Anatomy-Routed Attention (CARA), which guides decoder attention towards question-relevant cardiac structures (Figure 2). Below, we detail its anatomical occupancy prediction, question-guided region routing, and attention injection.

3.1 PREDICTING ANATOMICAL OCCUPANCY

We build CARA on Qwen2.5-VL-3B-Instruct (Bai et al., 2025). For a QA instance with M input CMR images, $\mathcal{I} = \{I^{(m)}\}_{m=1}^{M}$, each $I^{(m)} \in \mathbb{R}^{448\times448\times3}$ is processed by the shared vision encoder. Qwen2.5-VL uses 14×14 image patches and spatially merges each 2×2 group of neighboring patch tokens, yielding a 16×16 grid of post-merger visual representations for each image:

$$F^{(m)} \in \mathbb{R}^{16\times16\times2048}, \qquad m = 1, \ldots, M. \tag{1}$$

which are the same visual features passed to the language model. A convolutional head g_ϕ, shared across images, views, and sequences, predicts question-independent soft occupancy maps for the LV cavity, myocardium, and RV cavity:

$$P^{(m)} = \sigma\Big(g_\phi\Big(F^{(m)}\Big)\Big) \in [0,1]^{16\times16\times3}, \tag{2}$$

where σ denotes the element-wise sigmoid function. We denote the occupancy map for structure $r \in \{\mathrm{LV}, \mathrm{MYO}, \mathrm{RV}\}$ by $P_r^{(m)} \in [0,1]^{16\times16}$. The predictor is supervised with soft occupancy targets, where each target measures the fraction of the corresponding visual-token cell occupied by the anatomical structure. Thus, $P_r^{(m)}$ provides an image-specific coarse anatomical prior that is spatially aligned with the VLM visual-token grid. Segmentation masks are used only to supervise g_ϕ during training; at inference, all occupancy maps are predicted directly from the input CMR images. Architectural details are provided in Appendix D.1.

3.2 ANATOMY-ROUTED ATTENTION INJECTION

Each clinical question is associated with one of 14 predefined clinical tasks, indexed by k. For each input image m, a fixed routing function ρ_k selects the anatomical evidence relevant to the task:

$$a_k^{(m)} = \rho_k\Big(P_{\mathrm{LV}}^{(m)}, P_{\mathrm{MYO}}^{(m)}, P_{\mathrm{RV}}^{(m)}\Big) \in [0,1]^{16\times16}. \tag{3}$$

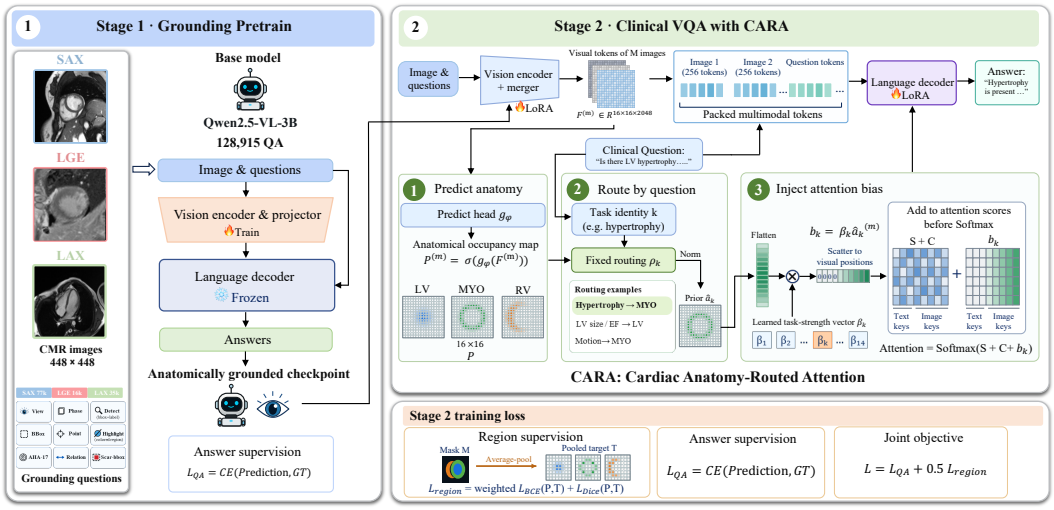


Figure 2: Two-stage training with CARA. Stage 1 learns anatomical grounding while keeping the language model frozen. Stage 2 jointly trains clinical answer generation and anatomical occupancy prediction. CARA selects the question-relevant occupancy map and injects the resulting anatomical prior into decoder attention with a learned task-specific strength.

For example, hypertrophy-related tasks are routed to the myocardium, whereas LV-size and EF tasks are routed to the LV cavity. Because occupancy confidence may vary across images, We peak-normalise each routed map independently to make its scale comparable across images:

$$\widehat{a}_k^{(m)} = \frac{a_k^{(m)}}{\max\left(\epsilon, \left\| a_k^{(m)} \right\|_\infty\right)}, \qquad \epsilon = 10^{-6}. \tag{4}$$

Each task has a learned scalar injection strength β_k. After flattening the 16×16 prior, let $\widehat{a}_{k,j}^{(m)}$ denote the value associated with the j-th visual token of image m, and let $\pi_m(j)$ map this token to its position in the language-model sequence. We construct a sequence-aligned bias $b_k \in \mathbb{R}^L$ as

$$b_{k,p} = \begin{cases} \beta_k\, \widehat{a}_{k,j}^{(m)}, & p = \pi_m(j), \\ 0, & p \notin \mathcal{I}_{\text{visual}}, \end{cases} \tag{5}$$

where $\mathcal{I}_{\text{visual}}$ denotes the visual-token positions. This directly aligns each image-specific anatomical prior with its corresponding visual tokens, including multi-image inputs. CARA then injects this bias into the pre-softmax decoder attention. Let $S_{i,p}^{(\ell,h)}$ denote the scaled query–key score from query position i to key position p in decoder layer ℓ and head h, and let $C_{i,p}$ denote the original causal and padding mask:

$$\widetilde{S}_{i,p}^{(\ell,h)} = S_{i,p}^{(\ell,h)} + C_{i,p} + b_{k,p}. \tag{6}$$

Because $b_{k,p}$ depends only on the key position, the same anatomical bias is shared across decoder layers, heads, and query positions. Positive β_k promotes attention to the routed anatomy, zero recovers the original attention pattern, and negative β_k suppresses it. During autoregressive generation, the bias remains attached to the original visual tokens and is zero for newly generated positions, allowing CARA to remain fully mask-free at inference.

3.3 Two-stage training with anatomical supervision

Training proceeds from anatomical grounding to clinical VQA. In Stage 1, we fully fine-tune the vision encoder and multimodal projector on 128,915 anatomical-grounding QA pairs while keeping the language model frozen. This checkpoint initialises Stage 2, where we apply LoRA to adapt the VLM on 42,799 clinical QA pairs across 14 tasks, jointly optimising CARA's occupancy head g_ϕ and attention strengths $\{\beta_k\}$.

We supervise answer generation with next-token cross-entropy over answer tokens. For anatomical supervision, reference masks are average-pooled onto the visual-token grid to obtain soft occupancy targets $R_r^{(m)} \in [0,1]^{16\times16}$ for image m and structure r. The predicted occupancy maps are supervised using positive-class-weighted binary cross-entropy and soft Dice:

$$\mathcal{L}_{\text{region}} = \mathcal{L}_{\text{WBCE}} + \mathcal{L}_{\text{Dice}}, \qquad \mathcal{L} = \mathcal{L}_{\text{QA}} + \lambda\mathcal{L}_{\text{region}}, \quad \lambda = 0.5. \tag{7}$$

Only valid anatomical targets contribute to the region loss; images without valid masks contribute only to the QA loss.

Visual features are detached before entering g_ϕ, so the region loss does not update the vision encoder. The QA loss also updates g_ϕ and β_k through the differentiable attention bias. At inference, CARA predicts anatomical priors directly from visual features without requiring masks as model inputs. Detailed loss definitions and annotation-validity rules are provided in Appendix D.

4 Experiments

4.1 Implementation details

Implementation and evaluation. We initialise Qwen2.5-VL-3B-Instruct and train the CARA-VL model on four NVIDIA B200 GPUs using AdamW, bfloat16 precision, and cosine decay with 5% warm-up. The base learning rate is 10^{-4}, with 10^{-2} for Stage 2 CARA injection strengths. Each stage runs for three epochs, with effective batch sizes of 64 and 16 for Stage 1 and Stage 2, respectively. Stage 2 uses LoRA with rank 16, scaling factor 32, and dropout 0.05. All evaluations use the final Stage 2 checkpoint. Additional training settings are provided in Appendix D.

We evaluate all 14 clinical tasks using the ACDC test dataset for SAX cine, the M&Ms-2 test dataset for LAX cine, and a held-out 50-patient LGE set. The eight SAX tasks are additionally evaluated on external In-house B. Patients are held out across tasks, views, and both training stages; performance is scored per QA instance. For In-house B, we evaluate the same predictions separately against two reference standards. Rule-based labels are generated using the same measurement-based labelling framework as the training data. Report-based labels are extracted directly from clinical reports, with echocardiographic grading thresholds used for selected findings. Cohort-specific criteria and differences between the two standards are detailed in Appendix C.6. For classification, we report balanced accuracy (BA; mean sensitivity and specificity) and macro-F_1. Regional hypertrophy and wall motion use micro-F_1 over question–segment pairs, with each anatomical segment labelled affected or unaffected. Scoring details are provided in Appendix D.4.

Baselines We compare our CARA-VL model with general medical VLMs and cardiac-specific systems. The general medical baselines comprise four open models: Lingshu-7B (Xu et al., 2026), HuatuoGPT-Vision-7B (Chen et al., 2024) (Huatuo-Vision), MedGemma-4B-it (Sellergren et al., 2026), and LLaVA-Med-7B (Li et al., 2023). All four are evaluated zero-shot, without adaptation to our CMR training data. The cardiac-specific baselines include CMR-CLIP (Nakashima et al., 2026), the contrastive CMR model of Shad et al. (2026), and BAAI Cardiac Agent (Qu et al., 2026). For BAAI Cardiac Agent, we evaluate its VLM component (BAAI-VLM) for direct image-based question answering without external tool calls.

4.2 Main results

Limited zero-shot generalization of general medical VLMs. General medical VLMs struggle with fine-grained CMR clinical QA (Tables 1 and 2). Their mean SAX BA ranges from 46.7% to 51.9%, close to the binary chance level. Performance on LGE and LAX is also generally limited. These results indicate that broad medical pretraining alone does not reliably support the cardiac visual recognition and localisation required by our tasks.

Internal evaluation: clinical assessment and localisation. CARA-VL achieves the highest scores across all 14 internal tasks, leading in both BA and macro-F_1 for classification and micro-F_1 for regional localisation (Table 1). On SAX, it reaches 90.7% mean BA and 90.0% mean macro-F_1, exceeding CMR-CLIP by 22.6 and 30.3 percentage points, respectively. Its advantage extends to

Table 1: QA-level clinical VQA results on the ACDC test set (SAX), the 50-patient LGE test set, and the M&Ms-2 test set (LAX). Classification cells show binary BA / macro-F_1; LV size, EF, and EDV are evaluated as normal versus abnormal. Regional tasks (†) report micro-F_1. All scores are percentages(%).

(a) SAX cine

Model	Hyper.	LV size	RV size	EF	GRS	GCS	**Avg.**	Reg. hyper.†	Reg. motion†
General medical VLMs									
LLaVA-Med-7B	50.0 / 9.9	50.0 / 24.3	50.0 / 42.4	47.5 / 39.9	50.0 / 41.0	50.0 / 38.8	49.6 / 32.7	9.2	31.2
MedGemma-4B-it	56.6 / 55.7	60.7 / 46.0	36.7 / 35.5	50.0 / 27.8	50.0 / 23.4	51.1 / 39.4	50.9 / 38.0	9.2	31.9
Huatuo-Vision	50.0 / 47.1	50.0 / 24.3	50.0 / 20.9	50.0 / 27.8	52.1 / 35.5	59.5 / 59.5	51.9 / 35.9	0.9	24.1
Lingshu-7B	50.0 / 47.1	44.4 / 27.6	50.8 / 22.7	50.0 / 27.8	39.2 / 39.5	45.7 / 43.5	46.7 / 34.7	3.2	16.9
CMR-specific models									
CMR-CLIP	80.0 / 60.6	62.5 / 47.9	39.4 / 39.7	81.4 / 77.2	65.4 / 57.4	80.1 / 75.6	68.1 / 59.7	20.3	52.0
Shad et al. (2026)	42.9 / 43.6	50.0 / 41.9	51.1 / 44.7	50.0 / 27.8	47.7 / 39.8	41.1 / 37.1	47.1 / 39.2	14.6	21.9
BAAI-VLM	49.1 / 46.7	50.0 / 40.4	50.0 / 42.4	64.4 / 63.2	59.3 / 58.5	76.7 / 77.8	58.3 / 54.8	11.3	26.6
CARA-VL (Ours)	**86.7 / 84.3**	**89.9 / 88.4**	**86.6 / 88.8**	**92.1 / 90.4**	**92.6 / 92.6**	**96.4 / 95.7**	**90.7 / 90.0**	**61.8**	**55.8**

Model	**(b) Late gadolinium enhancement**			**(c) Long-axis cine**		
	Scar presence	Transmurality	Extent	EF	EDV	GLS
General medical VLMs						
LLaVA-Med-7B	50.0 / 21.0	50.0 / 12.2	50.0 / 40.6	50.0 / 36.8	50.0 / 22.0	50.0 / 41.8
MedGemma-4B-it	56.7 / 57.0	53.9 / 20.2	50.0 / 40.6	48.9 / 36.3	48.3 / 30.9	49.0 / 47.1
Huatuo-Vision	57.5 / 48.3	50.0 / 12.2	48.3 / 26.0	50.0 / 29.5	50.0 / 22.0	47.0 / 38.0
Lingshu-7B	50.0 / 21.0	50.0 / 12.2	50.0 / 40.6	50.0 / 29.5	50.0 / 22.0	52.8 / 51.4
CMR-specific models						
CMR-CLIP	63.1 / 63.1	56.7 / 53.8	51.1 / 26.5	70.0 / 64.8	54.8 / 31.9	68.5 / 67.6
Shad et al. (2026)	50.3 / 44.3	50.0 / 12.2	52.4 / 31.8	53.0 / 41.9	48.2 / 46.2	30.5 / 31.4
BAAI-VLM	50.1 / 49.6	43.0 / 39.5	33.0 / 27.4	69.3 / 68.9	57.5 / 58.1	55.9 / 53.2
CARA-VL (Ours)	**82.9 / 82.0**	**68.5 / 71.1**	**82.5 / 82.3**	**78.1 / 74.9**	**84.9 / 81.0**	**77.1 / 79.6**

structural, functional, and scar assessment in LAX and LGE. Regional hypertrophy shows a particularly pronounced gain: 61.8% micro-F_1 versus 20.3% for the strongest baseline; regional wall motion reaches 55.8% versus 52.0%. These gains extend beyond global abnormality classification to identifying affected myocardial regions, supporting fine-grained clinical assessment across CMR imaging settings.

External evaluation: generalization across cohorts and clinical standards. On In-house B, CARA-VL leads under Rule-based labels on five of six classification tasks and both regional tasks (Table 2). Its mean BA / macro-F_1 reaches 79.5% / 72.4%, compared with 67.5% / 49.1% for CMR-CLIP, with EF remaining the exception. Results are more mixed under Report-based labels: CMR-CLIP performs better on hypertrophy and EF, and achieves regional hypertrophy and wall-motion micro-F_1 of 57.3% and 54.4%, compared with 30.9% and 34.5% for CARA-VL. CMR-CLIP's report-supervised pretraining may contribute to its stronger agreement with these labels. These results support generalization to an unseen cohort, while highlighting a remaining gap in agreement with clinical reporting. The contrast between the two reference standards also shows that external performance depends on how clinical findings are defined and recorded. Differences in reference criteria and anatomical coverage are detailed in Appendix C.6 (Table 9).

4.3 Ablation

Table 2: External VQA results on In-house B. Classification cells show BA / macro-F_1; regional cells show micro-F_1 (%). Avg. covers six classification tasks. Report-based GRS/GCS use wall-motion-score proxies; reference definitions, regional coverage, and report-field availability are detailed in Appendix C.6.

Model	Reference	Hyper.	LV size	RV size	EF	GRS	GCS	**Avg.**	Reg. hyper.	Reg. motion
General medical VLMs										
LLaVA-Med-7B	Rule-based	50.0 / 13.3	50.6 / 16.2	50.0 / 49.4	49.0 / 47.3	50.0 / 47.6	50.0 / 48.5	49.9 / 37.0	7.2	10.5
	Report-based	50.0 / 23.0	50.4 / 41.2	50.0 / 40.4	49.6 / 49.4	50.0 / 42.8	50.0 / 43.0	50.0 / 40.0	32.2	28.9
MedGemma-4B-it	Rule-based	58.3 / 48.6	56.1 / 43.0	52.2 / 48.9	49.8 / 27.6	48.9 / 13.3	46.5 / 36.1	51.9 / 36.2	7.2	12.7
	Report-based	52.9 / 50.7	55.9 / 55.7	51.3 / 48.8	50.1 / 21.8	48.0 / 23.5	47.4 / 44.0	50.9 / 40.7	32.2	39.0
Huatuo-Vision	Rule-based	49.5 / 48.9	50.0 / 11.1	50.0 / 2.3	50.0 / 24.1	47.3 / 19.3	55.8 / 46.7	50.4 / 25.4	3.3	9.2
	Report-based	49.5 / 45.8	50.0 / 36.9	50.0 / 24.4	50.0 / 17.9	48.3 / 29.7	49.5 / 49.2	49.6 / 34.0	5.3	26.1
Lingshu-7B	Rule-based	50.0 / 45.8	53.5 / 51.3	50.8 / 4.0	50.0 / 24.1	49.2 / 42.5	49.5 / 49.3	50.5 / 36.2	1.9	8.8
	Report-based	50.0 / 41.2	51.5 / 45.6	49.8 / 25.7	50.0 / 17.9	52.1 / 50.1	50.2 / 46.0	50.6 / 37.8	22.1	25.6
CMR-specific models										
CMR-CLIP	Rule-based	85.8 / 74.2	62.5 / 34.6	51.1 / 46.6	**76.8 / 71.7**	59.8 / 28.2	68.7 / 39.2	67.5 / 49.1	18.9	21.2
	Report-based	**78.8 / 78.3**	63.4 / **62.9**	50.9 / **49.9**	**75.9 / 64.0**	60.9 / 43.3	69.4 / 59.1	**66.5** / 59.6	**57.3**	**54.4**
Shad et al. (2026)	Rule-based	52.9 / 41.7	47.1 / 47.3	50.0 / 49.4	49.9 / 24.0	49.5 / 49.4	39.5 / 18.5	48.2 / 38.4	10.8	8.7
	Report-based	47.3 / 43.6	45.1 / 36.0	50.0 / 40.4	49.9 / 17.8	49.2 / 46.3	46.0 / 30.6	47.9 / 35.8	29.3	18.2
BAAI-VLM	Rule-based	50.0 / 46.1	50.0 / 46.7	50.0 / 49.4	51.9 / 45.1	54.3 / 55.4	58.2 / 53.0	52.4 / 49.3	5.9	12.5
	Report-based	49.9 / 41.4	50.0 / 29.4	50.0 / 40.4	52.7 / 49.7	52.7 / 49.5	57.5 / 58.1	52.1 / 44.7	28.9	34.4
CARA-VL (Ours)	Rule-based	**87.1 / 90.7**	**82.5 / 66.6**	**62.5 / 59.1**	73.5 / 66.3	**87.3 / 86.0**	**84.1 / 65.8**	**79.5 / 72.4**	**55.1**	**36.7**
	Report-based	69.9 / 72.7	**64.0** / 60.7	**52.4** / 46.8	72.6 / 58.2	**62.7 / 64.5**	**75.0 / 77.2**	66.1 / **63.4**	30.9	34.5

Table 3: Ablation on ACDC SAX clinical QA after Stage-2 training. BA and macro-F_1 are averaged across the six global classification tasks (%).

Model	Stage-1	CARA	BA	Macro-F_1
CARA-VL (Ours)	✓	✓	90.7	90.0
w/o Stage-1	×	✓	88.4	87.9
w/o CARA	✓	×	86.8	86.5
w/o Stage-1 and CARA	×	×	86.2	86.0

Table 3 evaluates the contributions of anatomical grounding pretraining and CARA. Removing Stage-1 reduces mean BA and macro-F_1 by 2.3 and 2.1 percentage points, respectively, while removing CARA produces larger drops of 3.9 and 3.5 points. The full model outperforms the variant without either component by 4.5 points in BA and 4.0 points in macro-F_1. Stage-1 also provides a larger gain when CARA is enabled, suggesting that anatomical grounding and guided attention offer complementary benefits. Figure 3 compares attention maps from models trained without and with CARA. For the same SAX image, the model trained with CARA shows more concentrated attention on the myocardium for hypertrophy and the LV cavity for LV size. The alignment between the routed priors and decoder attention provides a qualitative view of CARA's anatomical guidance. These examples illustrate how anatomical routing guides attention towards different cardiac structures according to the clinical question.

5 Discussion

Our study demonstrates the value of anatomical supervision for developing and evaluating cardiac VLMs. Our dataset covers 14 clinical QA tasks across three CMR imaging modalities, deriving reference answers from segmentation masks, quantitative measurements, and task-specific criteria. These tasks assess cardiac structure, function, tissue characteristics, and regional localisation. CARA connects anatomical grounding with anatomy-routed attention, using learned anatomical priors to guide image interpretation. Its leading performance across all 14 internal tasks, together with its performance on an external clinical cohort, demonstrates capabilities in global abnormality recognition and regional assessment. Together, the data and method provide a practical framework for studying cardiac visual understanding and highlight anatomical learning as a promising direction for specialised medical VLMs.

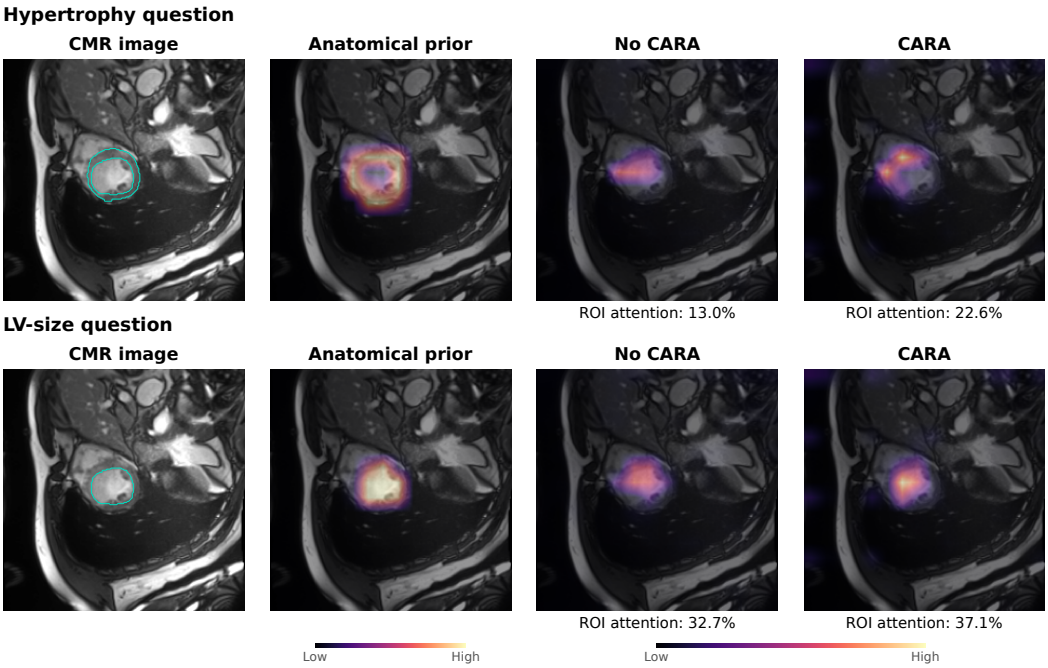


Figure 3: Attention maps from models trained without and with CARA for hypertrophy and LV-size questions on the same SAX image. Attention maps average all heads in the final eight decoder layers at the last prompt token. Attention maps share a colour scale within each row; anatomical priors are separately normalised to unit peak. ROI attention quantifies the fraction of visual attention allocated to the target anatomy.

Limitations. Our study has three main limitations. First, training answers are generated from annotation masks and image-derived measurements using predefined rules, inheriting errors and assumptions from these procedures. Although useful for evaluating visual recognition and localisation, these labels require independent clinical validation, as differences between measurement-based and report-based evaluation further highlight (Appendix C.6). Second, the underlying cohorts remain modest: generating multiple questions per examination increases QA volume without proportionally increasing patient diversity. Access to CMR studies, usable annotations, and quality review constrains expansion, while external evaluation currently covers SAX cine only. Finally, current performance remains insufficient for clinical use, and answering predefined questions from selected images captures only part of comprehensive CMR interpretation. If larger multicentre CMR datasets with paired clinical reports become available, future work could incorporate report supervision and assess clinical utility through independent expert review.

6 Conclusion

We presented CARA-VL, which combines anatomical grounding and anatomy-routed attention for fine-grained CMR question answering. Our data pipeline connects clinical reporting needs with anatomical annotations and quantitative measurements, providing grounding and clinical QA supervision across SAX cine, LGE, and LAX cine without paired reports for individual training images. CARA-VL leads the evaluated models across all 14 internal tasks, with ablations supporting the complementary benefits of anatomical learning and guided attention. External evaluation shows generalization to an unseen clinical cohort alongside remaining gaps in agreement with clinical reporting. Future work could extend anatomical guidance to full cine sequences, linking cardiac structure with motion over time, and explore more open-ended clinical questions. These directions provide opportunities to advance generative cardiac VQA and study how anatomical knowledge supports visual understanding in medical VLMs.

REPRODUCIBILITY STATEMENT

We provide detailed descriptions of data construction, model implementation, training, and evaluation in Appendices C and D to facilitate reproducibility. We will release a subset of the QA data constructed from public CMR datasets, in accordance with their licences and redistribution conditions, to support further research on cardiac visual question answering.

AI USE STATEMENT

Generative AI tools assisted with manuscript drafting, language editing and translation, literature search, code development and review, figure preparation, and interpretation and consistency checking of reported results. The authors take responsibility for the final manuscript and associated research artifacts, including the scientific claims, references, implementation, and reported results.

REFERENCES

Leon Axel, Mikael Kanski, Amit Jhaveri, Meng Ye, Bangwei Guo, Xiaoxiao He, and Dimitris Metaxas. Analysis of cardiac dynamic global function. *JRSM Cardiovascular Disease*, 15: 20480040261439970, 2026.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report, 2025. URL https://arxiv.org/abs/2502.13923.

Olivier Bernard, Alain Lalande, Clement Zotti, Frederick Cervenansky, Xin Yang, Pheng-Ann Heng, Irem Cetin, Karim Lekadir, Oscar Camara, Miguel Angel Gonzalez Ballester, et al. Deep learning techniques for automatic mri cardiac multi-structures segmentation and diagnosis: is the problem solved? *IEEE transactions on medical imaging*, 37(11):2514–2525, 2018.

Victor M Campello, Polyxeni Gkontra, Cristian Izquierdo, Carlos Martin-Isla, Alireza Sojoudi, Peter M Full, Klaus Maier-Hein, Yao Zhang, Zhiqiang He, Jun Ma, et al. Multi-centre, multi-vendor and multi-disease cardiac segmentation: the m&ms challenge. *IEEE Transactions on Medical Imaging*, 40(12):3543–3554, 2021.

Junying Chen, Chi Gui, Ruyi Ouyang, Anningzhe Gao, Shunian Chen, Guiming Hardy Chen, Xidong Wang, Zhenyang Cai, Ke Ji, Xiang Wan, et al. Towards injecting medical visual knowledge into multimodal llms at scale, 2024.

Data Science Bowl. Data Science Bowl Cardiac Challenge Data. https://www.kaggle.com/c/second-annual-data-science-bowl/data, 2015. Accessed: 2026-09-17.

Wangbin Ding, Lei Li, Junyi Qiu, Sihan Wang, Liqin Huang, Yinyin Chen, Shan Yang, and Xiahai Zhuang. Aligning multi-sequence cmr towards fully automated myocardial pathology segmentation. *IEEE Transactions on Medical Imaging*, 42(12):3474–3486, 2023.

Wangbin Ding, Lei Li, Junyi Qiu, Bogen Lin, Mingjing Yang, Liqin Huang, Lian-Ming Wu, Sihan Wang, and Xiahai Zhuang. Cinemyops: Segmenting myocardial pathologies from cine cardiac mr. *IEEE Transactions on Medical Imaging*, 44(12):5027–5037, 2025.

Yunguan Fu, Wenjia Bai, Weixi Yi, Charlotte Manisty, Anish N Bhuva, Thomas A Treibel, James C Moon, Matthew J Clarkson, Rhodri Huw Davies, and Yipeng Hu. Development and validation of a versatile foundation model for cine cardiac magnetic resonance image analysis. *Communications Medicine*, 2026.

Yunhe Gao, Mu Zhou, and Dimitris N Metaxas. Utnet: a hybrid transformer architecture for medical image segmentation. In *International conference on medical image computing and computer-assisted intervention*, pp. 61–71. Springer, 2021.

Difei Gu, Yunhe Gao, Mu Zhou, and Dimitris Metaxas. Anatomy-vlm: a fine-grained vision-language model for medical interpretation. In *2026 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 2838–2847. IEEE, 2026.

Bangwei Guo, Meng Ye, Yunhe Gao, Bingyu Xin, Leon Axel, and Dimitris Metaxas. Verse: Integrating multiple queries as prompts for versatile cardiac mri segmentation. In *International Conference on Information Processing in Medical Imaging*, pp. 373–387. Springer, 2025.

Bangwei Guo, Yunhe Gao, Meng Ye, Difei Gu, Yang Zhou, Leon Axel, and Dimitris Metaxas. K-prism: A knowledge-guided and prompt integrated universal medical image segmentation model. In *International Conference on Learning Representations*, volume 2026, pp. 146455–146480, 2026.

Nadine Kawel-Boehm, Scott J Hetzel, Bharath Ambale-Venkatesh, Gabriella Captur, Christopher J Francois, Michael Jerosch-Herold, Michael Salerno, Shawn D Teague, Emanuela Valsangiacomo-Buechel, Rob J Van der Geest, et al. Reference ranges ("normal values") for cardiovascular magnetic resonance (cmr) in adults and children: 2020 update. *Journal of cardiovascular magnetic resonance*, 22(1):87, 2020.

Raymond J Kim, Edwin Wu, Allen Rafael, Enn-Ling Chen, Michele A Parker, Orlando Simonetti, Francis J Klocke, Robert O Bonow, and Robert M Judd. The use of contrast-enhanced magnetic resonance imaging to identify reversible myocardial dysfunction. *New England Journal of Medicine*, 343(20):1445–1453, 2000.

Christopher M Kramer, Jörg Barkhausen, Chiara Bucciarelli-Ducci, Scott D Flamm, Raymond J Kim, and Eike Nagel. Standardized cardiovascular magnetic resonance imaging (cmr) protocols: 2020 update. *Journal of cardiovascular magnetic resonance*, 22(1):17, 2020.

Alain Lalande, Zhihao Chen, Thomas Decourselle, Abdul Qayyum, Thibaut Pommier, Luc Lorgis, Ezequiel de La Rosa, Alexandre Cochet, Yves Cottin, Dominique Ginhac, et al. Emidec: a database usable for the automatic evaluation of myocardial infarction from delayed-enhancement cardiac mri. *Data*, 5(4):89, 2020.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in neural information processing systems*, 36:28541–28564, 2023.

Xiang Li, Like Li, Yuchen Jiang, Hao Wang, Xinyu Qiao, Ting Feng, Hao Luo, and Yong Zhao. Vision-language models in medical image analysis: From simple fusion to general large models. *Information Fusion*, 118:102995, 2025.

Carolin Lim, Edyta Blaszczyk, Leili Riazy, Stephanie Wiesemann, Johannes Schüler, Florian von Knobelsdorff-Brenkenhoff, and Jeanette Schulz-Menger. Quantification of myocardial strain assessed by cardiovascular magnetic resonance feature tracking in healthy subjects—influence of segmentation and analysis software. *European radiology*, 31(6):3962–3972, 2021.

Yinbin Lu and Alan Wang. Integrating language into medical visual recognition and reasoning: A survey. *Medical image analysis*, 102:103514, 2025.

Carlos Martín-Isla, Víctor M Campello, Cristian Izquierdo, Kaisar Kushibar, Carla Sendra-Balcells, Polyxeni Gkontra, Alireza Sojoudi, Mitchell J Fulton, Tewodros Weldebirhan Arega, Kumaradevan Punithakumar, et al. Deep learning segmentation of the right ventricle in cardiac mri: the m&ms challenge. *IEEE Journal of Biomedical and Health Informatics*, 27(7):3302–3313, 2023.

Michael Moor, Qian Huang, Shirley Wu, Michihiro Yasunaga, Yash Dalmia, Jure Leskovec, Cyril Zakka, Eduardo Pontes Reis, and Pranav Rajpurkar. Med-flamingo: a multimodal medical few-shot learner. In *Machine learning for health (ML4H)*, pp. 353–367. PMLR, 2023.

Makiya Nakashima, Jielin Qiu, Peide Huang, Jihye Lee, Po-hao Chen, Richard Grimm, Christopher Nguyen, Byung-Hak Kim, Ding Zhao, Deborah Kwon, et al. Contrastive language image pre-training for a cardiac magnetic resonance image embedding with zero-shot capabilities. *Nature Communications*, 17(1):4416, 2026.

American Heart Association Writing Group on Myocardial Segmentation, Registration for Cardiac Imaging:, Manuel D. Cerqueira, Neil J. Weissman, Vasken Dilsizian, Alice K. Jacobs, Sanjiv Kaul, Warren K. Laskey, Dudley J. Pennell, John A. Rumberger, Thomas Ryan, and

Mario S. Verani. Standardized myocardial segmentation and nomenclature for tomographic imaging of the heart. *Circulation*, 105(4):539–542, 2002. doi: 10.1161/hc0402.102975. URL https://www.ahajournals.org/doi/abs/10.1161/hc0402.102975.

Jack W O'Sullivan, Mohammad Asadi, Lennart Elbe, Akshay Chaudhari, Tahoura Nedaee, Francois Haddad, Michael Salerno, Li Fe-Fei, Ehsan Adeli, Rima Arnaout, et al. Marcus: An agentic, multimodal vision-language model for cardiac diagnosis and management. *arXiv preprint arXiv:2603.22179*, 2026.

Daniele M Papetti, Kirsten Van Abeelen, Rhodri Davies, Roberto Menè, Francesca Heilbron, Francesco P Perelli, Jessica Artico, Andreas Seraphim, James C Moon, Gianfranco Parati, et al. An accurate and time-efficient deep learning-based system for automated segmentation and reporting of cardiac magnetic resonance-detected ischemic scar. *Computer methods and programs in biomedicine*, 229:107321, 2023.

Gianni Pedrizzetti, Piet Claus, Philip J Kilner, and Eike Nagel. Principles of cardiovascular magnetic resonance feature tracking and echocardiographic speckle tracking for informed clinical use. *Journal of cardiovascular magnetic resonance*, 18(1):51, 2016.

Junyi Qiu, Lei Li, Sihan Wang, Ke Zhang, Yinyin Chen, Shan Yang, and Xiahai Zhuang. Myops-net: Myocardial pathology segmentation with flexible combination of multi-sequence cmr images. *Medical image analysis*, 84:102694, 2023.

Taiping Qu, Hongkai Zhang, Lantian Zhang, Can Zhao, Nan Zhang, Hui Wang, Zhen Zhou, Mingye Zou, Kairui Bo, Pengfei Zhao, et al. Baai cardiac agent: An intelligent multimodal agent for automated reasoning and diagnosis of cardiovascular diseases from cardiac magnetic resonance imaging. *arXiv preprint arXiv:2604.04078*, 2026.

Prabhakar Shantha Rajiah, Christopher J François, and Tim Leiner. Cardiac mri: state of the art. *Radiology*, 307(3):e223008, 2023.

Jeanette Schulz-Menger, David A Bluemke, Jens Bremerich, Scott D Flamm, Mark A Fogel, Matthias G Friedrich, Raymond J Kim, Florian von Knobelsdorff-Brenkenhoff, Christopher M Kramer, Dudley J Pennell, et al. Standardized image interpretation and post-processing in cardiovascular magnetic resonance-2020 update: Society for cardiovascular magnetic resonance (scmr): Board of trustees task force on standardized post-processing. *Journal of cardiovascular magnetic resonance*, 22(1):19, 2020.

Andrew Sellergren, Chufan Gao, Fereshteh Mahvar, Timo Kohlberger, Fayaz Jamil, Madeleine Traverse, Alberto Tono, Bashir Sadjad, Lin Yang, Charles Lau, et al. Medgemma 1.5 technical report. *arXiv preprint arXiv:2604.05081*, 2026.

Rohan Shad, Cyril Zakka, Dhamanpreet Kaur, Mrudang Mathur, Robyn Fong, Joseph Cho, Ross Warren Filice, John Mongan, Kimberly Kallianos, Nishith Khandwala, et al. A generalizable deep learning system for cardiac mri. *Nature Biomedical Engineering*, pp. 1–16, 2026.

Weiwen Xu, Hou Pong Chan, Long Li, Mahani Aljunied, Ruifeng Yuan, Jianyu Wang, Chenghao Xiao, Guizhen Chen, Chaoqun Liu, Zhaodonghui Li, et al. Lingshu: Generalist foundation model for unified multimodal medical understanding and reasoning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026.

Meng Ye, Bingyu Xin, Bangwei Guo, Leon Axel, and Dimitris Metaxas. Learning volumetric neural deformable models to recover 3d regional heart wall motion from multi-planar tagged mri. *arXiv preprint arXiv:2411.15233*, 2024.

Mengmeng Zhang, Xiaoping Wu, Hao Luo, Fan Wang, and Yisheng Lv. Medground: Bridging the evidence gap in medical vision-language models with verified grounding data. *arXiv preprint arXiv:2601.06847*, 2026.

Xiahai Zhuang. Multivariate mixture model for myocardial segmentation combining multi-source images. *IEEE transactions on pattern analysis and machine intelligence*, 41(12):2933–2946, 2018.

A Related work

General medical VQA models. General medical VLMs such as LLaVA-Med (Li et al., 2023), Med-Flamingo (Moor et al., 2023), HuatuoGPT-Vision (Chen et al., 2024), and MedGemma (Sellergren et al., 2026) have demonstrated strong medical visual question answering capabilities through large-scale multimodal pretraining and medical instruction tuning. However, their visual training is largely built around broad biomedical figures and common medical imaging domains, including radiography, pathology, dermatology, and ophthalmology (Lu & Wang, 2025; Li et al., 2025), rather than systematic supervision on cardiac MRI. This specialization gap is particularly consequential for CMR, where clinically meaningful answers depend on cardiac anatomy, cine dynamics, physical-scale measurements, and sequence-specific pathology (Schulz-Menger et al., 2020; Rajiah et al., 2023). Thus, strong general medical VQA capability does not necessarily translate to reliable CMR-specific clinical assessment.

Emerging CMR-specific foundation and vision–language models. Recent work has increasingly explored foundation and multimodal models tailored to cardiac MRI. CMR-CLIP (Nakashima et al., 2026) and Shad et al. (2026) both adopt CLIP-style image–text representation learning from cine CMR and associated clinical reports, targeting zero-shot disease recognition and downstream functional or diagnostic prediction. CineMA (Fu et al., 2026) scales self-supervised cine-CMR pretraining to support tasks including segmentation, landmark localization, diagnosis, and prognosis, while MARCUS (O'Sullivan et al., 2026) introduces modality-specific cardiac vision–language experts for ECG, echocardiography, and CMR within an agentic multimodal system. BAAI Cardiac Agent (Qu et al., 2026) combines CMR-specific VQA with expert-model orchestration for cardiac segmentation, functional quantification, disease diagnosis, and structured report generation. Together, these studies highlight growing interest in CMR-specific VLMs, but current approaches often emphasize contrastive representation learning, rely on heterogeneous or complex input settings, and lack a standardized fine-grained clinical QA formulation. Their dependence on large institutional or non-public cohorts also limits reproducibility and direct comparison.

B Clinical Background on Cardiac MRI

This appendix introduces cardiac anatomy, common cardiac MRI views and sequences, and the clinical concepts used to describe cardiac structure, function, and tissue characteristics. The left and right ventricular (LV and RV) cavities contain blood, while the myocardium forms the muscular walls responsible for contraction.

B.1 Imaging views and sequences

Figure 4 contrasts the imaging geometry and reference anatomy used across the three task families.

Short-axis cine. Short-axis (SAX) cine imaging provides cross-sectional views of the ventricles perpendicular to the left ventricular long axis. Cine acquisitions resolve changes in ventricular cavity size and myocardial wall thickening throughout the cardiac cycle. A basal-to-apical stack enables quantification of ventricular volumes, ejection fraction, and myocardial mass, while regional assessment characterises wall thickness and systolic wall motion (Kramer et al., 2020; Schulz-Menger et al., 2020).

Long-axis cine. Long-axis (LAX) cine imaging comprises standard two-, three-, and four-chamber views aligned with the left ventricular long axis. These planes delineate the ventricular apex, atrioventricular junctions, and chamber geometry, providing complementary assessment of global and regional systolic function. Temporal changes in myocardial position and ventricular length also support assessment of longitudinal shortening and mitral annular excursion (Kramer et al., 2020; Schulz-Menger et al., 2020).

Late gadolinium enhancement. Late gadolinium enhancement (LGE) imaging characterises myocardial injury using delayed contrast-enhanced acquisitions, with normal myocardial signal suppressed to delineate hyperenhanced tissue. The location, distribution, and transmural extent of

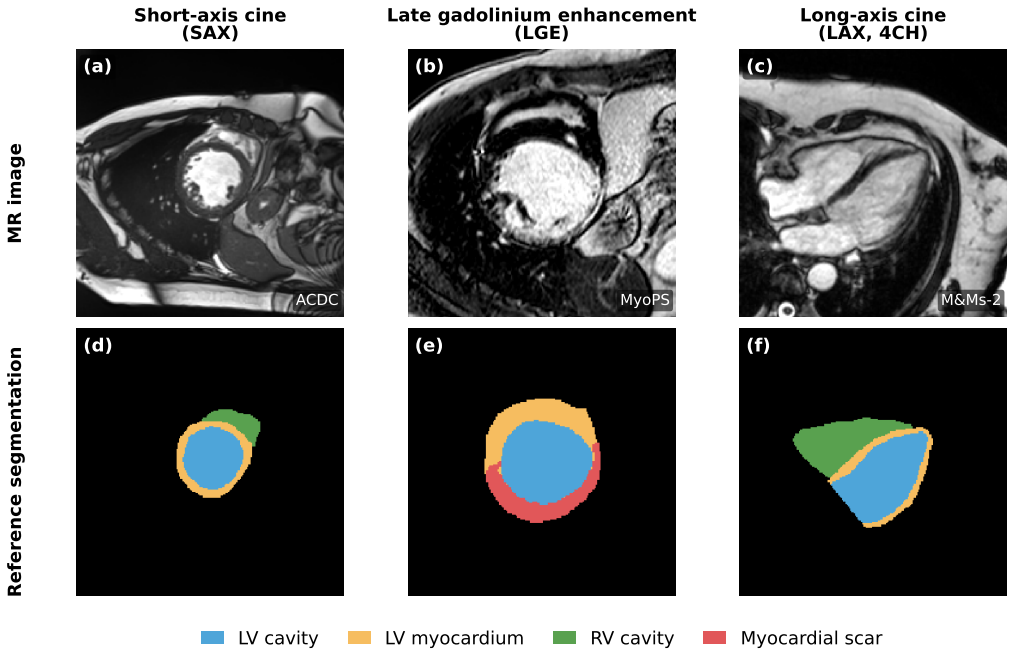


Figure 4: Representative CMR views and spatially matched reference annotations. Top: end-diastolic SAX cine (ACDC), short-axis LGE with myocardial scar (MyoPS), and end-diastolic four-chamber LAX cine (M&Ms-2). Bottom: the corresponding segmentations. SAX shows ventricular cross-sections, LAX displays chamber geometry along the ventricular long axis, and LGE delineates hyperenhanced scar. Colours denote the LV cavity (blue), LV myocardium (orange), RV cavity (green), and myocardial scar (red). These are reference annotations, not model predictions.

enhancement inform the assessment of myocardial scar and the differentiation of ischaemic and non-ischaemic injury patterns (Schulz-Menger et al., 2020).

B.2 Cardiac structure, function, and tissue characteristics

The reference masks in Figure 4 illustrate the ventricular cavities, surrounding LV myocardium, and myocardial scar.

Wall thickness and chamber size. Myocardial hypertrophy is characterised by increased wall thickness, which may be diffuse or confined to specific myocardial regions. Ventricular dilation instead refers to enlargement of the ventricular cavity. Wall thickness and cavity size therefore describe distinct aspects of cardiac morphology: a thickened myocardial wall does not necessarily imply an enlarged chamber. Their assessment requires clearly defined imaging planes and cardiac phases (Schulz-Menger et al., 2020).

AHA 17-segment model. The American Heart Association (AHA) 17-segment model standardises the anatomical subdivision of the left ventricular myocardium (on Myocardial Segmentation et al., 2002). It comprises six basal segments, six mid-ventricular segments, four apical segments, and the apical cap. The basal and mid-ventricular levels each include anterior, anteroseptal, inferoseptal, inferior, inferolateral, and anterolateral segments, whereas the apical level contains anterior, septal, inferior, and lateral segments. The apical cap represents the terminal myocardium beyond the ventricular cavity. This nomenclature specifies both the base-to-apex level and circumferential position, providing a consistent reference for reporting regional myocardial findings.

Ventricular volume and systolic function. In cine CMR, end-diastole and end-systole are conventionally identified as the phases of maximal and minimal ventricular cavity volume, respectively.

The corresponding volumes are termed end-diastolic volume (EDV) and end-systolic volume (ESV). Their difference defines stroke volume, and ejection fraction (EF) expresses stroke volume as a percentage of EDV (Schulz-Menger et al., 2020):

$$\mathrm{EF} = \frac{\mathrm{EDV} - \mathrm{ESV}}{\mathrm{EDV}} \times 100\%.$$

EDV characterises ventricular cavity size and is commonly indexed to body surface area for comparison with normal reference ranges (Kawel-Boehm et al., 2020). EF provides a global measure of systolic emptying, whereas regional wall-motion assessment evaluates myocardial displacement and systolic thickening within individual segments. Beyond these end-diastolic and end-systolic measures, ventricular volume–time curves further characterise the timing and rates of ventricular emptying and filling (Axel et al., 2026).

Myocardial deformation. Myocardial strain quantifies tissue deformation relative to a reference configuration, conventionally end-diastole, and characterises regional and global myocardial mechanics (Pedrizzetti et al., 2016). Global radial strain (GRS) describes myocardial wall thickening, whereas global circumferential strain (GCS) and global longitudinal strain (GLS) describe shortening around the ventricular circumference and along the base-to-apex axis, respectively. "Global" refers to an aggregate measure over the analysed myocardium. Under the conventional sign convention, systolic radial strain is positive, while circumferential and longitudinal strain are negative; impaired shortening is therefore reflected by less negative values. CMR-based deformation assessment includes feature tracking of routine cine images and motion estimation from tagged MRI. Strain estimates depend on the segmentation procedure, slice coverage, and analysis software, making consistent measurement definitions important for interpretation (Lim et al., 2021; Ye et al., 2024).

Scar presence, transmurality, and extent. Myocardial scar represents replacement of injured myocardium by fibrotic tissue and can be assessed using LGE imaging. Scar presence indicates whether scar is identified within the myocardium. Transmurality describes the fraction of local myocardial wall thickness occupied by scar, ranging from partial-thickness to full-thickness involvement. Scar extent describes the amount or spatial spread of affected myocardium (Schulz-Menger et al., 2020). Circumferential extent describes the proportion of the myocardial circumference affected by scar. Transmurality and circumferential extent thus describe different dimensions of scar involvement: a lesion may span the full wall thickness while remaining confined to a small circumferential region. Segmentation of scar and myocardial boundaries supports quantitative assessment of these properties, as explored in automated scar analysis (Papetti et al., 2023; Guo et al., 2025).

C Data construction details

We build two training sets from one rendering pipeline (Fig. 1). The Stage-1 grounding set reads its answers directly from expert reference masks; the Stage-2 clinical set computes its answers from physical measurements on reference or automatically generated masks. Table 5 lists every source, its mask provenance, and its role in each stage.

C.1 From free-text reports to a closed question set

The 14 questions are not chosen a priori. They are the surviving intersection of what clinical reports actually state, what has a guideline-anchored threshold, and what can be measured from segmentation.

Report corpus and recurring findings. We analyse 3,215 de-identified CMR reports from three sources: 1,112 disease-labelled reports, 1,902 consecutive clinical reports, and 201 reports accompanied by structured post-processing results. A curated bilingual lexicon covering 20 clinical finding categories identifies whether each finding is mentioned in a report. Frequently discussed findings include myocardial enhancement, chamber dilation, reduced ejection fraction, wall thickening, and regional wall-motion abnormalities. These frequencies characterise recurring clinical concerns and inform candidate question selection. However, paired CMR images are unavailable for this report corpus. The reports therefore guide the scope of clinical questions rather than provide image–question–answer training pairs.

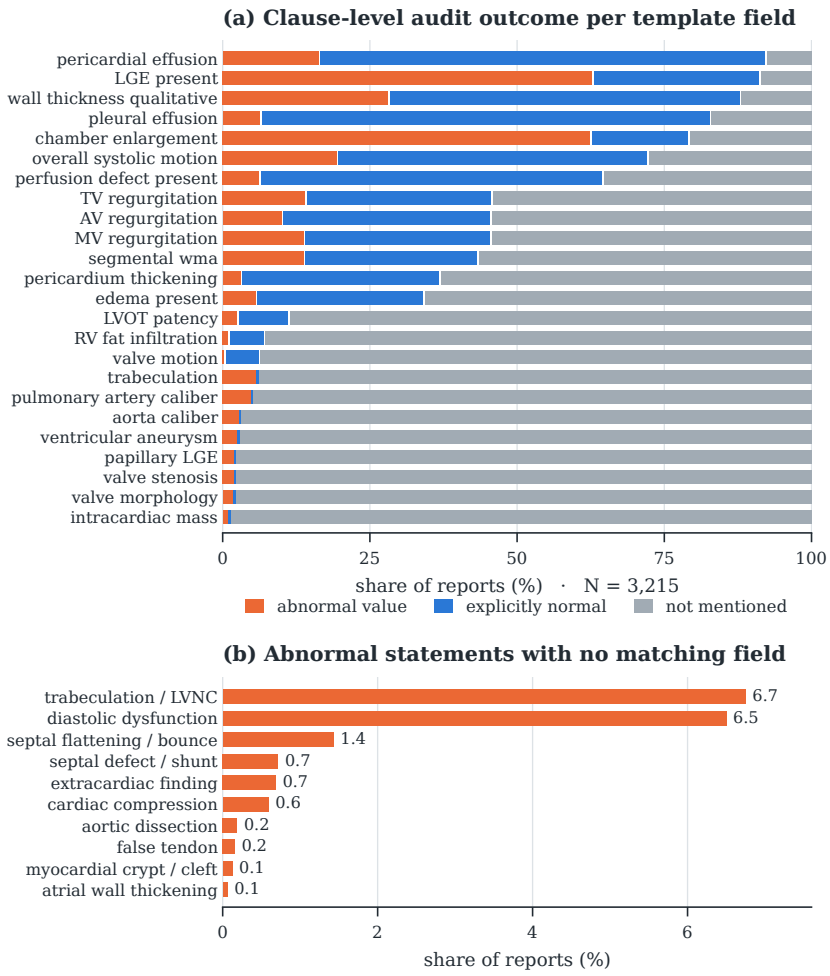


Figure 5: Structured report audit over 3,215 CMR reports. **(a)** For each template field, the share of reports in which the field is not mentioned, explicitly normal, or carries an abnormal value, ordered by mention rate (explicitly normal + abnormal). A high mention rate with a low abnormality rate marks a finding that radiologists address routinely but rarely find abnormal. **(b)** Abnormal statements matching no template field.

Structured report audit. We organise report findings into a structured template, with each field assigned an imaging sequence or study-level scope and a clinical information type, such as a measurement, visual finding, or anatomical location. A clause-level audit distinguishes findings that are not mentioned, explicitly normal, or abnormal, retaining reported measurements or grades where available. We iteratively review abnormal statements not captured by the template and add or refine fields to improve coverage. The resulting template contains 173 fields, including 59 with predefined answer categories. We then aggregate mention and abnormality frequencies separately, distinguishing frequently discussed findings from frequently reported abnormalities. Together, these statistics inform candidate question selection; the template serves as a report-analysis framework rather than an image-level labelling scheme. Figure 5 plots the audit.

Selecting clinical QA tasks. Candidate tasks are selected according to three criteria: they address recurring clinical findings in the reports, admit explicit measurement or grading criteria, and are supported by the imaging data and segmentation annotations available in our cohorts. Reports guide question selection, while reference answers are derived separately from image-based measurements and acquisition metadata. Valvular disease, perfusion defects, oedema, atrial findings, and paramet-

Table 4: Definitions and labelling criteria for the 14 clinical QA tasks. ED and ES denote end-diastole and end-systole, respectively. The table summarises answer categories, measurement rules, and exclusion intervals. Segmentation sources, quality control, measurement derivation, and training-set statistics are detailed in Appendix C.5.

Domain	Task and answer categories	Frames	Measurement and labelling rule
SAX	Hypertrophy: present / absent	ED	Present if any AHA segment has ED wall thickness $\geq 13\,\mathrm{mm}$; source-labelled normal cases are assigned negative.
	Regional hypertrophy: segment list	ED	Segments with ED wall thickness $\geq 13\,\mathrm{mm}$; six segments at basal and mid-ventricular levels, four at the apical level.
	LV size: normal / mild / moderate / severe dilation	ED+ES	Patient-level LV EDV: < 200, $[200, 240)$, $[240, 300)$, and $\geq 300\,\mathrm{mL}$, respectively.
	EF: normal / mild / moderate / severe reduction	ED+ES	Whole-stack area-ratio EF: $\geq 50\%$, $[40, 50)\%$, $[30, 40)\%$, and $< 30\%$, respectively.
	RV size: dilated / normal	ED+ES	Dilated if RV EDV $\geq 200\,\mathrm{mL}$ and RV/LV EDV ratio ≥ 1.3; single-criterion positives require visual confirmation. Normal if both are below threshold; unconfirmed candidates are excluded.
	GRS surrogate: normal / reduced	ED+ES	Mean systolic wall thickening: $\geq 45\%$ normal, $< 35\%$ reduced; exclude $[35, 45)\%$.
	GCS: normal / reduced	ED+ES	Feature-tracking peak circumferential strain: $\leq -16\%$ normal, $\geq -12\%$ reduced; exclude $(-16, -12)\%$.
	Regional motion: segment list	ED+ES	Segments with systolic wall thickening $< 20\%$. Inherits GRS sample filtering, with no additional segment-level exclusion interval.
LGE	Scar presence: present / absent	Single	Present if the reference scar mask is non-empty.
	Transmurality: transmural / non-transmural	Single	Transmural if any sliding 60° sector has mean transmural extent $\geq 50\%$. Automatically labelled positives are retained only at $\geq 60\%$; recorded review overrides are retained.
	Extent: localised / widespread	Single	Widespread if the circumferential scar fraction is ≥ 0.5. Automatically exclude $[0.40, 0.60)$; recorded review overrides are retained.
LAX	EF: normal / mild / moderate / severe reduction	ED+ES	Patient-level SAX EF grade where available; otherwise derived from Kaggle ventricular volumes or M&Ms-2 SAX reference masks.
	EDV: normal / mild / moderate / severe dilation	ED+ES	Patient-level SAX LV-size grade where available; otherwise derived from Kaggle EDV or M&Ms-2 SAX reference masks.
	GLS: normal / reduced	ED+ES	Feature-tracking GLS: $\leq -16\%$ normal, $\geq -12\%$ reduced; exclude $(-16, -12)\%$.

ric mapping are excluded because the available data do not support the required measurements. We also exclude ischaemic versus non-ischaemic enhancement classification because the available LGE cohort lacks sufficient variation for this distinction, and atrial-plane images without assessable ventricular myocardium. The resulting 14 clinical QA tasks are summarised in Table 4, with grading criteria detailed below.

Task-specific grading criteria. We derive categorical reference answers from image-based measurements using task-specific assessment criteria, informed by published clinical standards and the definitions of the measured quantities. For selected tasks, borderline measurements are excluded to reduce label ambiguity. Table 4 summarises the measurement rules, answer categories, and exclusion intervals used for clinical QA construction.

C.2 Data Sources

The corpus spans SAX cine, LGE, and LAX cine (two- and four-chamber views), with source contributions summarised in Table 5. Public datasets include ACDC (Bernard et al., 2018), M&Ms, M&Ms-2 (Campello et al., 2021; Martín-Isla et al., 2023), MyoPS-related cohorts (Zhuang, 2018; Qiu et al., 2023; Ding et al., 2023; 2025), EMIDEC (Lalande et al., 2020), the cine and LGE subsets of CMR-MULTI (Qu et al., 2026), and the Kaggle Data Science Bowl Challenge (Data Science Bowl, 2015). In-house A is an image-only training collection comprising an expert-annotated cine subset, a clinical image catalogue, and additional cine collections with automatically generated masks. In-house B is the image–report cohort used exclusively for external evaluation. Its final SAX set contains 612 examinations from 461 distinct patient identifiers and provides 14,336 SAX QA pairs. We use expert reference segmentation masks where available. For images without reference masks, we obtain masks from a deployed internal CMR segmentation pipeline and review their quality before clinical QA construction. Table 5 summarises the data sources and their contributions

to each training stage and evaluation set. Segmentation and quality-control procedures are detailed in Appendix C.5, and data-split procedures in §C.6. In-house B is reserved exclusively for external evaluation.

Table 5: Actual source contributions to training and held-out evaluation.

Domain	Source	Mask source	Stage-1	Stage-2 train	Held out
Public datasets					
SAX	ACDC	Expert	100 (8,845)	89 (1,803)	50
	M&Ms	Expert	320 (26,846)	281 (4,815)	–
	M&Ms-2	Expert	200 (16,449)	200 (3,841)	160
	CMR-MULTI cine	Expert	105 (10,093)	95 (1,459)	–
	MyoPS-related bSSFP cine	Auto. + QC	–	64 (994)	–
	Kaggle DSB	Auto. + QC	–	769 (6,924)	–
LGE	MyoPS-related cohorts	Expert	183 (10,900)	189 (3,618)	30
	EMIDEC	Expert	67 (2,815)	86 (1,143)	14
	CMR-MULTI LGE	Expert	31 (2,678)	41 (884)	6
LAX	M&Ms-2 (4CH)	Expert	200 (16,800)	200 (554)	160
	CMR-MULTI cine (4CH, 2CH)	Expert	240 series (18,240)	175 series (175)	–
	Kaggle DSB	Auto. + QC	–	1,064 (4,930)	–
	Public subtotal (QA)		113,666	31,140	
Internal datasets					
SAX	In-house A	Expert / Auto. + QC	153 (15,249)	9,278 QA	–
LAX	In-house A	Auto. + QC	–	2,381 QA	–
SAX	In-house B (with reports)	Auto. + QC	–	–	612 exams; 461 patients
	In-house subtotal (QA)		15,249	11,659	
Total QA			128,915	42,799	

Notes. Counts are not additive across stages or views. In-house A clinical training entries report QA counts; CMR-MULTI LAX entries report series rather than unique patients. Only expert masks are used for Stage-1.

C.3 Physical-scale rendering

We standardise spatial resolution and intensity range before QA generation. Using acquisition-specific pixel spacing, we resample SAX cine to 1.4 mm/pixel and LGE and LAX cine to 1.0 mm/pixel, with bilinear interpolation for images and nearest-neighbour interpolation for masks. Intensities are clipped to the 1st–99th percentiles for cine and the 2nd–98th percentiles of foreground pixels for LGE. Images and masks are cropped to 224×224 pixels, with segmentation-guided centring where needed to preserve cardiac coverage. Cases with unresolved spacing inconsistencies or truncated anatomy are excluded. The cropped images define the reference coordinate system and are subsequently upsampled bicubically to 448×448 for model input, with spatial coordinates scaled accordingly. All compared models receive the same preprocessed images.

C.4 Anatomical grounding supervision

Data and reference annotations. Stage-1 uses expert segmentation masks exclusively. The SAX pool contains 8,218 slices from 878 training patients across ACDC, M&Ms, M&Ms-2, CMR-MULTI, and the expert-annotated subset of In-house A. The LAX pool contains 440 series: 200 four-chamber series from M&Ms-2 and 240 two- and four-chamber series from CMR-MULTI. The LGE pool contains 2,241 scar-positive slices from MyoPS-related cohorts, EMIDEC, and CMR-MULTI. Held-out patients are excluded before QA generation. Annotations are mapped to a shared label set comprising the LV cavity, LV myocardium, and RV cavity, with scar retained for LGE; other annotation classes are excluded.

AHA-based myocardial partitioning. We use the 16 myocardial segments of the AHA model, comprising six basal, six mid-ventricular, and four apical segments, excluding the apical cap (on Myocardial Segmentation et al., 2002). For volumetric SAX data, slices containing both LV cavity and myocardium are assigned to approximately basal, mid-ventricular, and apical thirds, with the larger-cavity end designated basal. Within each slice, the LV-cavity centroid defines the polar origin. We

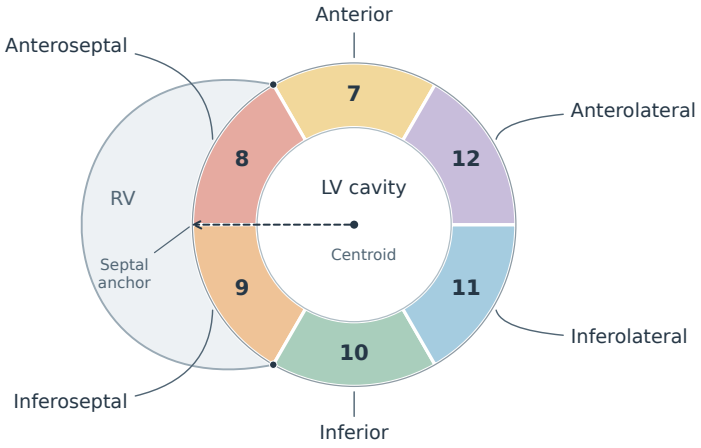


Figure 6: Mid-ventricular AHA partitioning in a canonical anatomical orientation. The six 60° myocardial sectors carry AHA numbers 7–12; these are anatomical identifiers, not the generator's local mask indices. The dashed ray runs from the LV-cavity centroid towards the septal anchor, separating the anteroseptal and inferoseptal sectors adjacent to the RV. Dots illustrate the two RV insertion locations. This is a geometric schematic, not a patient segmentation or a fixed scanner display orientation. Colours distinguish sectors only.

estimate the RV insertion locations from the myocardium adjacent to the RV and use their septal-arc midpoint to define the septal anchor. Acquisition orientation, where available, determines the anterior–inferior ordering of the segments. Basal and mid-ventricular slices are divided into six 60° sectors, with the anchor separating the anteroseptal and inferoseptal segments (Figure 6). Apical slices use four 90° sectors, with the septal segment centred on the anchor. Images, masks, and segment-label maps undergo the same geometric transformations during augmentation, preserving anatomical segment identities.

Grounding QA generation. We generate nine task types covering image context, anatomical recognition, localisation, and spatial relations (Table 6). Answers are derived from expert masks and the associated view, sequence, and phase information. Points and bounding boxes are represented by absolute pixel coordinates in the 448×448 reference image. Small annotated regions are filtered before localisation questions are generated, and scar bounding boxes are restricted to single connected components. For HIGHLIGHT and SEGMENT, the target region is highlighted at 50% opacity using one of eight colours. POINT questions include background probes. AHA segment questions use the labelled myocardial sectors described in §C.4. Task availability follows the annotations and imaging context of each source. Figure 7 shows one representative public-data example for each task.

Augmentation and consistency checks. For cine data, horizontal and vertical flips and rotations by 90°, 180°, and 270° are applied jointly to images, masks, and AHA segment maps. SAX slices receive one sampled transformation or the identity; LAX series retain the original images and receive three additional sampled transformations. The same transformation is applied to both images of a PHASE pair, and answers are regenerated after transformation. LGE images are not augmented. To reduce repetitive supervision, SAX VIEW and RELATION questions are retained with probabilities of 0.15 and 0.5, respectively. A consistency checker finds no mismatches in 14,311 sampled coordinate assertions across the three domains. This check assesses agreement between generated answers and transformed annotations, rather than annotation accuracy.

Dataset composition. The resulting dataset contains 128,915 QA pairs: 77,482 from SAX, 35,040 from LAX, and 16,393 from LGE. These pairs use 45,773 unique images, including 23,576 with highlight overlays; the 9,978 PHASE pairs each contain two images.

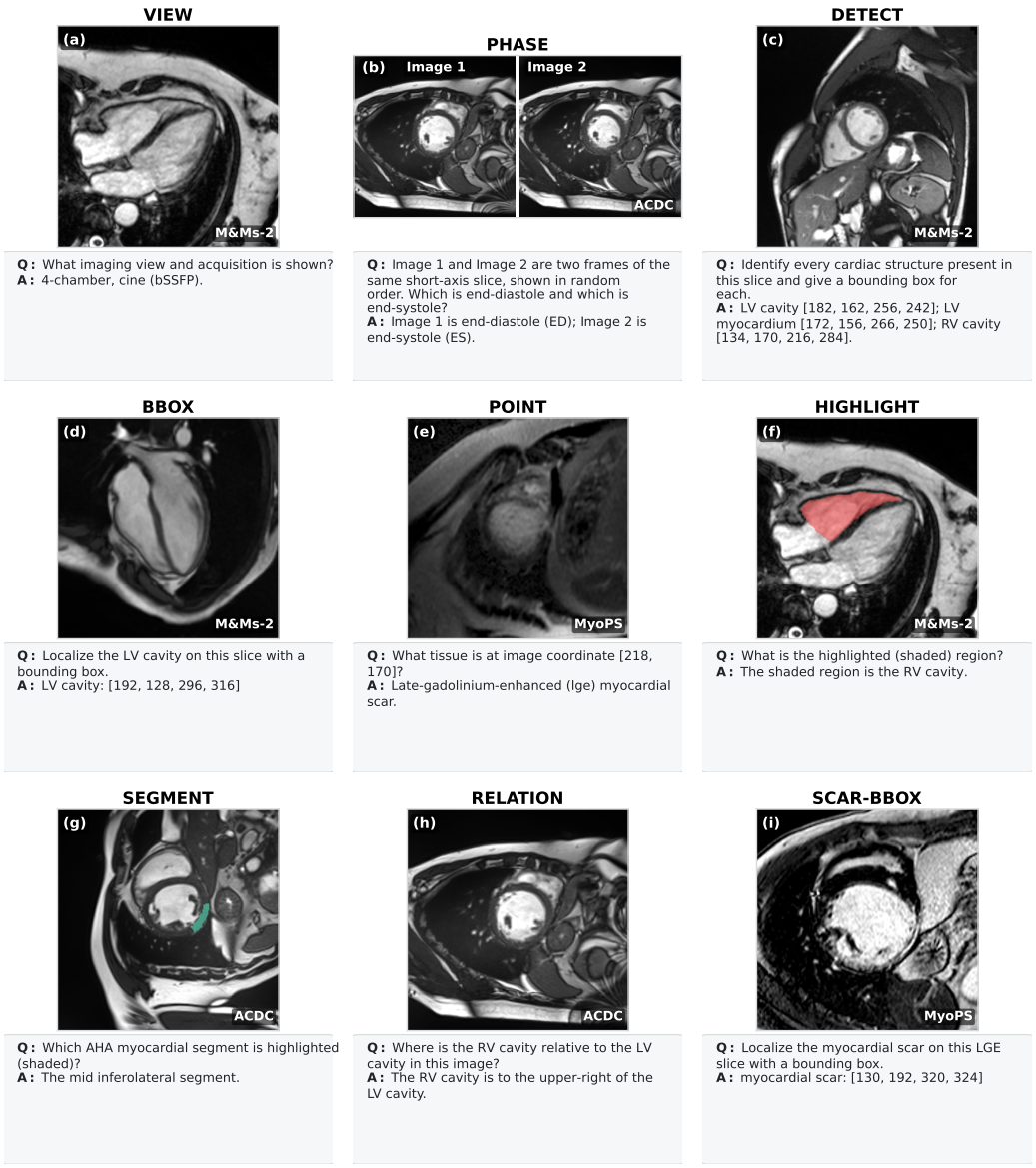


Figure 7: Examples of the nine Stage-1 grounding tasks from public ACDC, M&Ms-2, and MyoPS data. Each panel shows the frozen 448×448 input, generated question, and target answer; PHASE uses two input frames. Coordinate targets appear only in text, whereas the highlights in HIGHLIGHT and SEGMENT are part of the input.

C.5 Clinical QA construction

We construct clinical QA pairs from image-derived measurements and reference annotations, using the task-specific criteria in Table 4. The following sections describe reference eligibility, quality control, measurement derivation, and training-pool assembly.

C.5.1 Segmentation and quality control

Segmentation and label harmonisation. Cardiac segmentation provides an anatomical basis for quantitative assessment of cardiac structure and function (Gao et al., 2021; Guo et al., 2026). We retain expert reference masks where available and generate masks for the remaining images using a deployed internal CMR segmentation pipeline. Its native labels are mapped to a shared scheme

Table 6: Stage-1 grounding tasks. Answers are generated from the ground-truth mask.

Task	Question → answer	Domains	QA
VIEW	imaging view and sequence → "Short-axis, cine (bSSFP)."	SAX, LAX, LGE	7,005
PHASE	two shuffled frames of one slice → which frame is ED and which is ES	SAX, LAX	9,978
DETECT	enumerate all structures → name and box for each	SAX, LAX	11,738
BBOX	named structure → $[x_{\min}, y_{\min}, x_{\max}, y_{\max}]$	SAX, LAX	33,024
POINT	coordinate $[x, y]$ → structure name, or background	SAX, LAX, LGE	35,796
HIGHLIGHT	region highlighted in a random colour → structure name	SAX, LAX, LGE	15,038
SEGMENT	highlighted myocardial sector → AHA level and segment	SAX, LGE	8,538
RELATION	position of RV cavity relative to LV cavity (centroids, 8-pixel margin)	SAX, LAX	6,076
SCAR-BBOX	single connected scar component ($\geq$15 pixels) → box	LGE	1,722
Total			128,915

comprising the LV cavity, myocardium, and RV cavity, with papillary muscle merged into the LV cavity. Expert scar annotations are retained for LGE. Missing annotations are treated as unavailable rather than as evidence that a structure is absent.

Mask screening and QA slice selection. For SAX cine, masks across the full stack undergo a rapid visual check for obvious, substantial segmentation failures, such as missing structures or gross anatomical misalignment. This screening does not assess fine-grained boundary accuracy. Stack-level measurements, including EDV and EF, use the available masks that pass screening across the full anatomical coverage, including basal and apical levels. Middle-level slices are selected as image inputs for SAX QA, except for the global and regional hypertrophy tasks, which also retain eligible basal and apical slices. This input selection does not restrict the stack used to derive volume-based reference labels. QA pairs requiring both ED and ES use corresponding images whose masks pass screening at both phases.

From reviewed masks to clinical QA. Reviewed masks provide the anatomical regions for task-specific measurements, supplemented by acquisition metadata, source-provided values, or tracking outputs where required. Measurements are computed in physical units and converted into reference answers using Table 4. For SAX LV-size, RV-size, and EF questions, the reference answer is derived from the stack-level measurement and paired with an eligible middle-level image or ED/ES image pair. Slice-specific questions, such as hypertrophy and regional wall motion, use measurements from the displayed slice. An image can support multiple questions when the required annotations and measurements are available. The following subsections describe reference construction for each imaging domain and clinical task.

C.5.2 SHORT-AXIS CINE REFERENCE CONSTRUCTION

Hypertrophy and regional hypertrophy. We estimate end-diastolic wall thickness from the LV-cavity and myocardial masks by sampling 36 radial directions from the LV-cavity centroid. Along each ray, the contiguous myocardial thickness is converted to millimetres, and valid measurements are averaged within anatomically anchored AHA segments, using the partition described in §C.4. A segment is labelled hypertrophic when its mean thickness is at least 13 mm. The global hypertrophy question receives a positive answer if any segment in the displayed slice is hypertrophic. The regional question names the affected segments or states that none is hypertrophic. When at least five of six segments, or three of four segments, are affected, the answer uses the predefined description "Concentric left-ventricular hypertrophy in this slice." Both tasks use the same segment-level measurements and therefore have consistent presence labels.

SAX LV size. We compute LV-cavity areas from the ED segmentation masks by counting cavity pixels and multiplying by the corresponding pixel area. These areas are summed across the SAX stack, weighted by acquisition-specific slice separation, to estimate LV end-diastolic volume (EDV) in millilitres. Reference EDV values are used where available. The absolute EDV is then mapped to normal, mild, moderate, or severe LV dilation according to the thresholds in Table 4.

SAX ejection fraction. We compute LV-cavity areas from the ED and ES segmentation masks by counting cavity pixels and multiplying by the corresponding pixel area. Areas are summed across

matched SAX slices to estimate LV ejection fraction (EF):

$$\mathrm{EF} = 100 \left(1 - \frac{\sum_{z \in \mathcal{Z}} A_z^{\mathrm{ES}}}{\sum_{z \in \mathcal{Z}} A_z^{\mathrm{ED}}} \right), \tag{8}$$

where A_z^{ED} and A_z^{ES} are mask-derived LV-cavity areas, and $\mathcal{Z}$ contains the matched slices passing screening across the SAX stack. For uniformly spaced slices, the common slice separation cancels in the volume ratio. Study-level reference EF values are used where available. The resulting EF is classified as normal, mildly reduced, moderately reduced, or severely reduced according to Table 4.

SAX RV size. We compute study-level RV EDV and the RV-to-LV EDV ratio from ED cavity masks across the SAX stack, using the physical-volume calculation described for LV size. The pipeline automatically assigns a dilated label when both RV EDV $\geq 200\,\mathrm{mL}$ and the RV-to-LV EDV ratio ≥ 1.3 are satisfied. Studies meeting only one criterion are flagged for visual review of ED images and cavity-mask overlays: confirmed cases are labelled dilated, while unconfirmed candidates are excluded. Studies below both thresholds form the normal sampling pool. Each accepted study-level answer, "RV cavity dilated" or "RV cavity normal", is paired with eligible middle-level ED/ES image pairs to generate RV-size QA.

Radial motion and the GRS surrogate. We use mean systolic wall thickening derived from ED and ES myocardial masks as a surrogate for GRS. For each radial sampling direction j, the relative thickness change is computed as

$$h_j = 100\, \frac{t_j^{\mathrm{ES}} - t_j^{\mathrm{ED}}}{\max(t_j^{\mathrm{ED}}, 1)}, \tag{9}$$

where t_j^{ED} and t_j^{ES} denote mask-derived myocardial thicknesses in pixels. Only directions with positive ED thickness are included, and the denominator is bounded below by one pixel. Relative thickness changes are averaged within six myocardial sectors, followed by an average across valid sectors to obtain a slice-level reference. The resulting value is classified as normal or reduced using the thresholds and exclusion interval in Table 4.

Regional wall motion. We derive regional wall-motion labels from the ED and ES myocardial masks using the AHA 16-segment partition described in §C.4. Each middle-level SAX slice contains six AHA segments. For each segment, we average the mask-derived radial thickening values defined in Equation 9. A segment is labelled as having reduced motion when its mean systolic thickening is below 20%. The QA answer lists the affected AHA segments, states that none has reduced motion, or indicates that all six segments are affected.

Circumferential motion. We obtain GCS from the feature-tracking module of the deployed CMR analysis pipeline. A learned registration network estimates inter-frame displacement fields, which propagate myocardial contour points through the cine sequence. Mid-wall circumferential strain is computed from changes in the lengths of consecutive contour segments:

$$\varepsilon_{\mathrm{cc}}(t) = \frac{L(t) - L_0}{L_0}, \tag{10}$$

where L_0 and $L(t)$ denote segment lengths in the reference frame and at time t, respectively. The module reports the peak mid-wall circumferential strain as GCS, with more negative values indicating greater circumferential shortening. Cases without usable tracking results are excluded. GCS is classified as normal or reduced using the thresholds and exclusion interval in Table 4, with normal-control overrides described below.

C.5.3 LGE REFERENCE CONSTRUCTION

Scar presence. We derive scar-presence labels from expert scar masks provided by the MyoPS-related cohorts, EMIDEC, and CMR-MULTI. After harmonising annotation labels, an image is labelled scar-positive if its reference scar mask is non-empty, and scar-negative otherwise. The label is paired with a question asking whether myocardial scar is present in the displayed image. Presence QA includes both SAX and LAX LGE images; transmurality and extent QA require scar-positive SAX images with usable myocardial and scar annotations.

Regional scar transmurality. We derive transmurality labels from expert myocardial and scar masks. Along each radial direction, we compute the scar extent relative to the myocardial wall thickness and average these ratios within sliding 60° sectors. A slice is labelled as containing regionally transmural scar if any sector has a mean ratio of at least 50%; otherwise, it is labelled non-transmural.

Circumferential scar extent. We estimate scar extent from the myocardial and scar masks by sampling 180 radial directions from the LV-cavity centroid. Circumferential involvement is computed as

$$c = \frac{N_{\text{scar}}}{N_{\text{myocardium}}}, \tag{11}$$

where $N_{\text{myocardium}}$ counts rays intersecting the myocardial wall and N_{scar} counts those also intersecting scar within the wall. This measures circumferential coverage on the displayed slice rather than scar area or volume. The QA answer is "Localised scar." when $c < 0.5$ and "Widespread scar." when $c \geq 0.5$. Automatically labelled examples with $c \in [0.40, 0.60)$ are excluded, while recorded manual-review decisions can retain cases within this interval.

C.5.4 Long-axis cine reference construction

LAX ejection fraction and ventricular volume. LAX EF and EDV questions use two- or four-chamber ED/ES image pairs, with the view specified in the question. Reference labels are obtained from the same patient's SAX grades, available reference ventricular volumes, or measurements derived from SAX segmentation masks. We use the same four-grade categories as the SAX tasks (Table 4) and pair each patient-level label with the corresponding LAX images, maintaining consistent reference answers across views.

LAX longitudinal motion. Training GLS references are computed from the full cine sequence of each LAX view using the deployed feature-tracking module. Anatomical landmarks and myocardial segmentation define contour points, and a pretrained tracking network estimates dense displacement fields from the first frame to subsequent frames. Longitudinal strain at myocardial points is derived from local deformation and myocardial orientation, then temporally smoothed. We average strain across valid points at each frame and take the minimum of the resulting curve:

$$\text{GLS} = 100 \min_t \left(\frac{1}{|\mathcal{V}_t|} \sum_{i \in \mathcal{V}_t} \varepsilon_{\ell,i}(t) \right), \tag{12}$$

where $\mathcal{V}_t$ contains the valid myocardial points at frame t. Automatic myocardial masks support tracking; for M&Ms-2, expert ED masks replace the automatic segmentation, while motion is still estimated from the complete image sequence. Training values are classified as normal at $\leq -16\%$ and reduced at $\geq -12\%$, excluding intermediate, missing, and sign-inconsistent positive values. Each accepted label is paired with the corresponding ED/ES images for QA input; these two frames do not limit the sequence used for GLS computation. Eligibility is independent of EF/EDV reference availability.

C.5.5 QA assembly and dataset summary

The measurements described above are converted into binary, four-grade, or regional answers using Table 4, subject to task-specific exclusions and recorded review corrections. For the ACDC cohort, patients clinically diagnosed as normal are assigned normal GRS and GCS labels before thresholding and exclusion, with corresponding normal-reference handling applied to hypertrophy labels. Each accepted answer is paired with a task-specific question specifying the imaging view and anatomical context. Hypertrophy tasks use one ED image, the remaining SAX tasks and all LAX tasks use ED/ES pairs, and LGE tasks use one image. The model receives images without overlays and the question; masks support label construction and the anatomical supervision described in Section 3.1. The training sources in Table 5 yield 42,799 clinical QA pairs across 14 tasks: 29,114 SAX, 5,645 LGE, and 8,040 LAX pairs. Table 7 summarises task-level training counts and answer distributions. Task counts vary with annotation availability, measurement validity, and quality filtering. The final RV-size training subset contains 1,221 QA pairs: 227 dilated and 994 normal. The

203 non-transmural examples are oversampled to 1,249 records for training, producing 43,845 rows without adding independent examples.

Table 7: Clinical training QA counts before oversampling. Four-grade counts follow the order normal, mild, moderate, and severe.

Domain	Task	QA pairs	Answer counts
SAX	Hypertrophy	4,733	Absent / present: 4,243 / 490
	Regional hypertrophy	4,733	Unaffected / affected: 4,243 / 490
	LV size	3,985	2,963 / 412 / 304 / 306
	EF	3,971	2,291 / 732 / 418 / 530
	RV size	1,221	Normal / dilated: 994 / 227
	GRS surrogate	3,713	Normal / reduced: 2,939 / 774
	GCS	3,062	Normal / reduced: 2,216 / 846
	Regional motion	3,696	Unaffected / affected: 2,656 / 1,040
LGE	Scar presence	2,917	Absent / present: 786 / 2,131
	Transmurality	1,452	Non-transmural / transmural: 203 / 1,249
	Extent	1,276	Localised / widespread: 873 / 403
LAX	EF	2,725	1,752 / 493 / 189 / 291
	EDV	2,747	2,011 / 304 / 242 / 190
	GLS	2,568	Normal / reduced: 2,125 / 443
Total		42,799	

Note. Regional counts group QA pairs by whether no segment or at least one segment is affected.

Table 8: Clinical test QA counts and label distributions. Four-grade order follows Table 7; external counts use the existing model-scoring labels.

A. Internal: ACDC SAX, held-out LGE, and M&Ms-2 LAX

Domain	Task	QA pairs	Answer counts
SAX	Hypertrophy	457	Absent / present: 407 / 50
	Regional hypertrophy	457	Unaffected / affected: 406 / 51
	LV size	171	116 / 21 / 22 / 12
	EF	171	105 / 7 / 11 / 48
	RV size	170	Normal / dilated: 125 / 45
	GRS surrogate	160	Normal / reduced: 111 / 49
	GCS	148	Normal / reduced: 94 / 54
	Regional motion	171	Unaffected / affected: 109 / 62
LGE	Scar presence	463	Absent / present: 123 / 340
	Transmurality	209	Non-transmural / transmural: 29 / 180
	Extent	202	Localised / widespread: 138 / 64
LAX	EF	160	93 / 32 / 13 / 22
	EDV	160	115 / 19 / 19 / 7
	GLS	121	Normal / reduced: 87 / 34
Total		3,220	SAX: 1,905; LGE: 874; LAX: 441

B. External In-house B: same QA pool, two reference standards

		Rule-based		Report-based	
Task	QA pool	Labelled QA	Negative / positive	Labelled QA	Negative / positive
Hypertrophy	1,945	1,945	1,646 / 299	1,936	1,359 / 577
Regional hypertrophy	1,831	1,831	1,646 / 185	1,822	1,357 / 465
LV size	1,783	1,783	1,561 / 222	1,780	741 / 1,039
EF	1,795	1,795	1,226 / 569	1,795	1,404 / 391
RV size	1,795	1,795	1,753 / 42	1,778	1,203 / 575
GRS surrogate	1,708	1,708	1,551 / 157	1,708	1,280 / 428
GCS	1,684	1,684	1,587 / 97	1,684	1,272 / 412
Regional motion	1,795	1,795	1,458 / 337	1,795	1,314 / 481
Total	14,336	14,336	—	14,298	—

Notes. Counts are per QA; external LV size/EF are binary, and regional positives mean at least one affected region. Report-based counts exclude 38 QA pairs without report answers. Paired-reference counts in Table 9 additionally require documented, comparable fields (Appendix C.6).

Clinical QA examples. Figures 8 and 9 show contrasting examples of all 14 tasks from public ACDC, MyoPS, and M&Ms-2 data, with unmarked input images and verbatim training QA. Each pair shares one question; four-grade tasks show normal and severe cases.

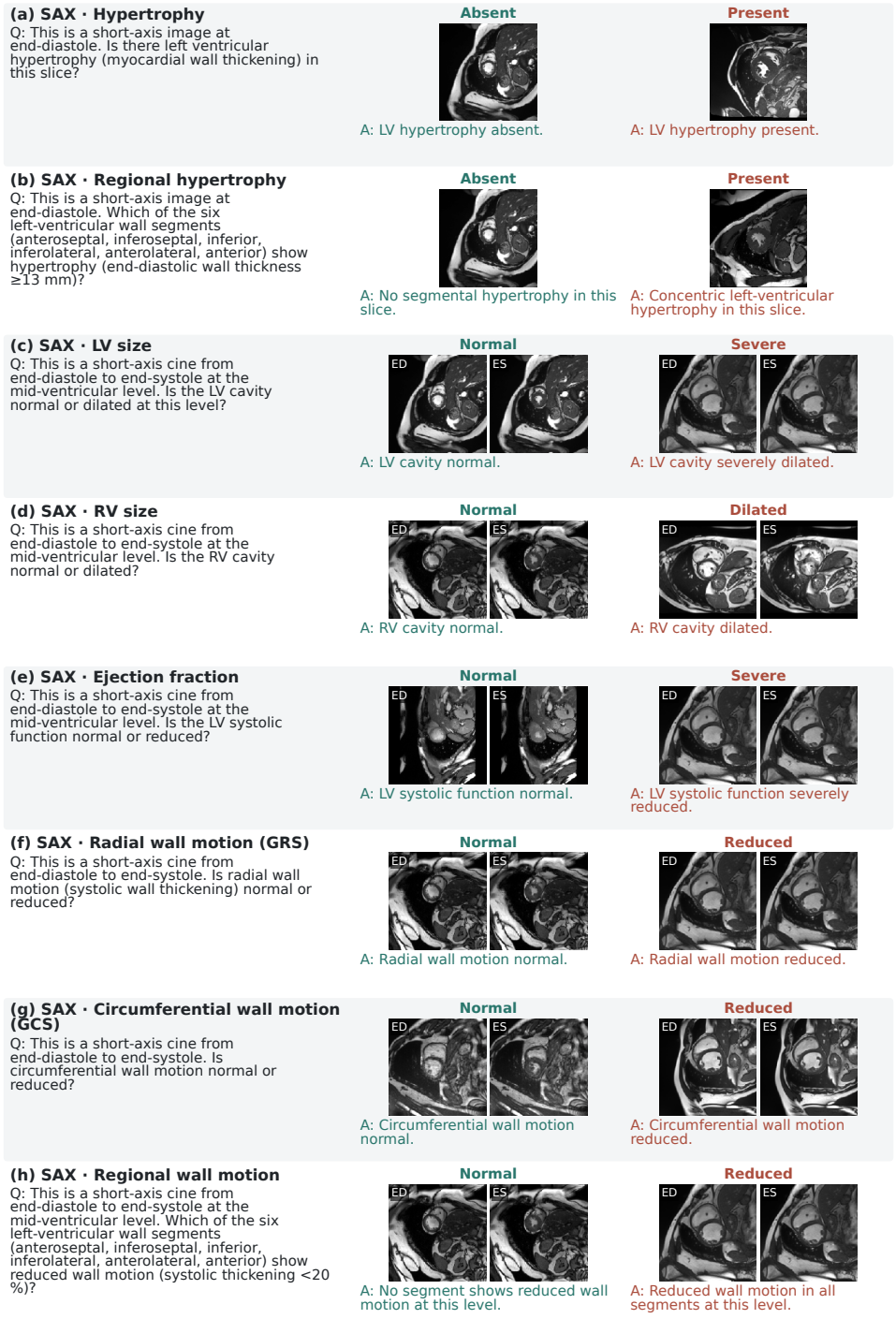


Figure 8: Clinical QA examples for the eight SAX tasks from ACDC, covering cardiac structure, ventricular function, and regional findings. Each row presents a task-specific question alongside two example inputs and their corresponding reference answers.

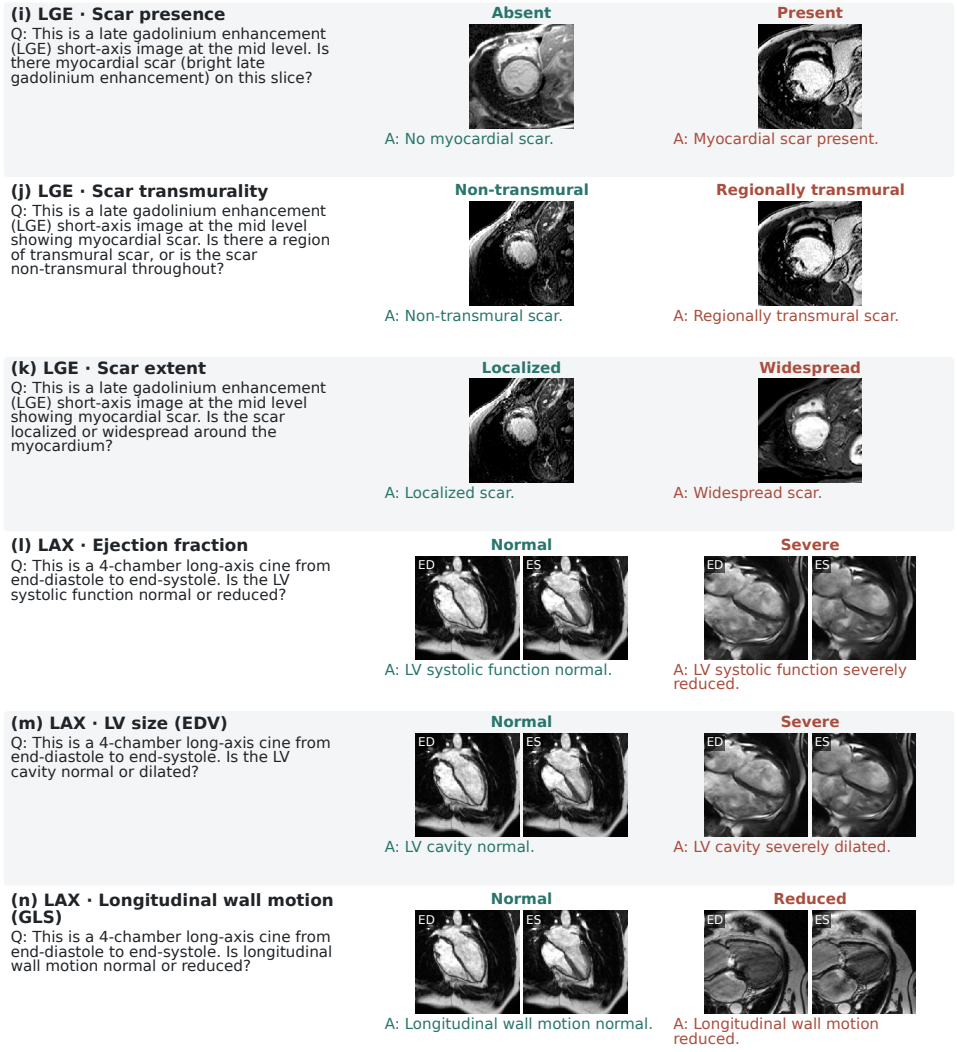


Figure 9: Clinical QA examples for the three LGE tasks from MyoPS and three LAX tasks from M&Ms-2, covering scar assessment and ventricular function. Each row presents a task-specific question alongside two example inputs and their corresponding reference answers.

C.6 Splits and evaluation sets

Patient-level partitions. Held-out patients are excluded across tasks, views, and both training stages. Internal evaluation comprises 3,220 QA pairs: 1,905 from ACDC SAX, 874 from 50 held-out LGE patients, and 441 from M&Ms-2 LAX. Table 8 lists task-level test counts and answer distributions, including the existing external scoring populations under both reference standards. Patient identities are shared across views and reconstructions when assigning splits. In-house B is reserved exclusively for external evaluation.

External cohort and dual reference labels. After quality filtering, In-house B contains 612 examinations from 461 distinct patients, yielding 14,336 SAX QA pairs. Repeated examinations are retained, and evaluation is performed per QA instance. We maintain two reference-label sources where available. **Rule-based labels** are generated by applying task-specific criteria to segmentation-derived measurements and other image-analysis outputs. **Report-based labels** are extracted from the findings, measurements, and grades documented in the corresponding clinical reports. QA instances without an archived task-level report answer are excluded from evaluation. All clinical

Table 9: Reference-standard differences and paired-label disagreement in In-house B . These are comparisons between reference labels, not model scores. Panel B uses only documented, comparable reference units; rates are percentages.

A. Implemented criteria and reference scope

Task	Rule-based reference	Report-based reference	Difference
Hypertrophy	Any slice AHA wall thickness ≥ 13 mm; reviewed basal/apical positives retained.	Septal or posterior wall thickness ≥ 11 mm.	Cut-off gap: 2 mm; slice-wide vs. two report regions.
LV size	Product LV EDVi: normal ≤ 105; mild $(105, 130]$; moderate $(130, 160]$; severe > 160.	Report LV EDVi above its recorded upper limit: 62 (female) or 75 (male); binary dilation.	Normal-limit gap: 43/30 mL/m^2. No matched report four-grade scale.
RV size	Product RV EDVi > 110 mL/m^2; if BSA is absent, RV EDV > 205 mL.	Explicit normal/enlarged RV description in the report conclusion.	Quantitative rule vs. qualitative finding; not numerically comparable.
EF	Product EF: normal $\geq 50\%$; mild $[40, 50)$; moderate $[30, 40)$; severe $< 30\%$.	Report EF value mapped using the same 50/40/30% bands.	Cut-off gap: 0 percentage points; different measurement sources.
GRS	Mean mask-derived systolic wall thickening: reduced $< 35\%$; normal $\geq 45\%$; $[35, 45)$ excluded.	Any report-mapped middle-segment WMS ≥ 2 indicates abnormal motion.	Thickening vs. WMS proxy; not equivalent strain references.
GCS	Feature-tracking peak circumferential strain: normal $\leq -16\%$; reduced $\geq -12\%$; $(-16, -12)$ excluded.	Same WMS proxy as above, not measured circumferential strain.	Strain vs. motion score; not numerically comparable.
Regional hypertrophy	Six middle-slice AHA segments; thickness ≥ 13 mm.	Two regions: septum and posterior wall; thickness ≥ 11 mm.	2 mm gap; map AS or IS to septum, IL to posterior.
Regional motion	Six middle-slice segments; systolic thickening $< 20\%$.	WMS ≥ 2 on five mapped segments; no anterolateral reference.	Different quantities; compare only observed shared segments.

B. Agreement between reference labels

Task	Compared labels (N)	Rule-based abnormal (%)	Report-based abnormal (%)	Disagreement (%)
Hypertrophy	1,936	15.4	29.8	15.9
LV size	1,780	12.3	58.4	46.9
RV size	1,778	2.4	32.3	31.3
EF	1,795	31.7	21.8	13.3
GRS vs. WMS proxy	1,553	10.0	27.6	23.6
GCS vs. WMS proxy	1,528	6.3	27.0	24.5
Regional hypertrophy	3,644	5.6	19.2	14.5
Regional motion	8,766	7.2	24.8	23.5

Notes. N counts paired QA labels for global tasks and paired region or segment labels for regional tasks. LV size and EF are evaluated as normal/abnormal. Disagreement is the proportion of compared labels that differ. Missing or indeterminate reference units are excluded. WMS denotes wall-motion score; its use for GRS/GCS provides a motion proxy rather than a direct strain reference. Report-field availability and comparison criteria are detailed in Appendix C.6. AS/IS/IL: anteroseptal/inferoseptal/inferolateral; BSA: body surface area. EDVi is expressed in mL/m^2.

reports in In-house B follow echocardiographic conventions for anatomical localisation and grading. For hypertrophy and LV dilation, the adopted thresholds are lower than the corresponding CMR thresholds, allowing the same numerical measurement to receive different normal/abnormal labels under the two standards. The references also differ in anatomical coverage: report-based regional hypertrophy covers septal and posterior regions, while regional motion covers five mid-ventricular segments. Report-based GRS and GCS labels use wall-motion-score proxies rather than direct strain measurements. We therefore evaluate the same model predictions separately against each reference source. Table 9 details the implemented criteria, reference coverage, and paired-label disagree-

ment. Disagreement reflects differences in measurement sources and anatomical coverage as well as thresholds.

Report-derived reference availability. The availability of report-derived references depends on whether relevant findings are documented and can be unambiguously extracted. Model evaluation uses task-level report labels, whereas paired-reference analysis additionally requires documented, comparable entries from both sources and excludes missing or indeterminate entries. The resulting sample counts therefore differ between Tables 8 and 9. Regional comparisons count matched regions or segments rather than QA instances.

D Implementation details

D.1 Model architecture

Anatomical occupancy head. The shared occupancy head maps each image's $16 \times 16 \times 2048$ post-merger features to three occupancy maps. It comprises a 1×1 convolution from 2048 to 256 channels, a 3×3 convolution with 256 channels, and a 3×3 output convolution. The first two layers use GELU, with dropout ($p = 0.1$) after the second. All convolutions preserve the spatial grid. Independent sigmoid activations produce LV-cavity, myocardial, and RV-cavity occupancies.

Routing and attention injection. Table 10 specifies the anatomical prior for each task. Each prior is peak-normalised per image with $\epsilon = 10^{-6}$ and scaled by an independent task-specific strength β_k, initialised to 2.0 without a sign constraint. The resulting bias is added at visual-key positions across decoder layers, attention heads, and query positions, preserving the original causal and padding masks (Equation 6).

Table 10: Anatomical priors for the 14 clinical tasks. LV and RV denote ventricular cavities, and MYO denotes myocardium. Each task has an independent learned injection strength.

Domain	Clinical task	Anatomical prior
SAX	Hypertrophy	MYO
SAX	Regional hypertrophy	MYO
SAX	LV size	LV
SAX	EF	LV
SAX	RV size	RV
SAX	Radial motion (GRS)	MYO
SAX	Circumferential motion (GCS)	MYO
SAX	Regional wall motion	MYO
LGE	Scar presence	MYO
LGE	Scar transmurality	MYO
LGE	Scar extent	MYO
LAX	EF	LV
LAX	EDV	LV
LAX	Longitudinal motion (GLS)	MYO

D.2 Training losses

During Stage 2, we optimise $\mathcal{L}_{\mathrm{QA}} + 0.5\,\mathcal{L}_{\mathrm{region}}$, where $\mathcal{L}_{\mathrm{QA}}$ is next-token cross-entropy averaged over answer tokens, excluding prompt and padding tokens, and $\mathcal{L}_{\mathrm{region}} = \mathcal{L}_{\mathrm{WBCE}} + \mathcal{L}_{\mathrm{Dice}}$.

Occupancy targets. Reference masks are average-pooled onto the visual-token grid to obtain soft occupancy targets. Let p_{ir} and R_{ir} denote the predicted and target occupancies for cell i in the local batch and structure $r \in \{\mathrm{LV}, \mathrm{MYO}, \mathrm{RV}\}$. The validity indicator $v_{ir} \in \{0, 1\}$ specifies whether the

image–structure target contributes to supervision. LGE supervision uses LV and MYO targets only. Images without valid anatomical targets contribute only to the QA loss.

Weighted binary cross-entropy. We use positive-class weights $w_r = (15, 40, 15)$ for LV, MYO, and RV, respectively, and average over valid cell–structure pairs:

$$\mathcal{L}_{\mathrm{WBCE}} = \frac{\sum_{i,r} v_{ir} \left[-w_r R_{ir} \log p_{ir} - (1 - R_{ir}) \log(1 - p_{ir})\right]}{\max\left(1, \sum_{i,r} v_{ir}\right)}. \tag{13}$$

The loss is evaluated from logits for numerical stability.

Soft Dice. Dice is computed across the local batch for each structure and averaged over valid structures, with a smoothing constant of one:

$$D_r = 1 - \frac{2 \sum_i v_{ir} p_{ir} R_{ir} + 1}{\sum_i v_{ir} p_{ir} + \sum_i v_{ir} R_{ir} + 1}, \qquad \mathcal{L}_{\mathrm{Dice}} = \frac{1}{|\mathcal{R}_v|} \sum_{r \in \mathcal{R}_v} D_r, \tag{14}$$

where $\mathcal{R}_v = \{r : \sum_i v_{ir} > 0\}$. If no valid annotation is available, $\mathcal{L}_{\mathrm{region}} = 0$.

Features are detached before entering g_ϕ, so the region loss updates the occupancy head without affecting the visual encoder. The QA loss also updates g_ϕ and β_k through the differentiable attention bias.

D.3 Training protocol

Stage 2 combines all 14 tasks into a shuffled training pool. The 203 non-transmural examples are oversampled to 1,249 records, increasing the pool from 42,799 QA pairs to 43,845 training records. No other task is oversampled. Task frequencies follow the resulting pool composition, and all records have unit loss weight. Stage 2 uses AdamW with learning rates of 10^{-4} for the LoRA parameters and occupancy head, and 10^{-2} for the injection strengths. Both groups use 5% linear warm-up followed by cosine decay. Training uses four NVIDIA B200 GPUs with four QA instances per GPU, giving an effective batch size of 16.

D.4 Evaluation metrics

Scores are computed over QA instances and reported as percentages.

Classification. For class c, true positives (TP_c), false positives (FP_c), and false negatives (FN_c) define precision, recall, and F_1:

$$P_c = \frac{\mathrm{TP}_c}{\mathrm{TP}_c + \mathrm{FP}_c}, \qquad R_c = \frac{\mathrm{TP}_c}{\mathrm{TP}_c + \mathrm{FN}_c}, \qquad F_{1,c} = \frac{2\mathrm{TP}_c}{2\mathrm{TP}_c + \mathrm{FP}_c + \mathrm{FN}_c}. \tag{15}$$

Balanced accuracy and macro-F_1 average over the C classes:

$$\mathrm{BA} = \frac{1}{C} \sum_{c=1}^{C} R_c, \qquad \text{Macro-}F_1 = \frac{1}{C} \sum_{c=1}^{C} F_{1,c}. \tag{16}$$

Binary tasks use $C = 2$; sensitivity and specificity are positive- and negative-class recall, respectively. PPV is positive-class precision, and ACC is the proportion of correct predictions. For maintable LV size, EF, and EDV, both predictions and references are folded into normal versus abnormal before scoring. Four-grade results use $C = 4$. SAX averages give equal weight to the six global classification tasks, using unrounded task scores and excluding regional tasks. External averages are computed separately for each reference. GRS and GCS remain separate QA tasks in this average although their Report-based labels share wall-motion-score proxies.

Regional localisation. Let G_q and $\widehat{G}_q$ contain the reference and predicted affected regions for question q, within the scored anatomical coverage. Pooled counts and micro-F_1 are

$$\mathrm{TP} = \sum_q |G_q \cap \widehat{G}_q|, \quad \mathrm{FP} = \sum_q |\widehat{G}_q \setminus G_q|, \quad \mathrm{FN} = \sum_q |G_q \setminus \widehat{G}_q|, \tag{17}$$

$$\text{Micro-}F_1 = \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}}. \tag{18}$$

Regional precision and recall use these pooled counts. Exact match is the proportion of questions with $\widehat{G}_q = G_q$. Mean Jaccard averages $|G_q \cap \widehat{G}_q|/|G_q \cup \widehat{G}_q|$ over questions, assigning one when both sets are empty. For both classification and regional metrics, an F_1 with a zero denominator is set to zero.

ROI attention. For visual token j, let a_j be its attention weight and $o_j \in [0, 1]$ its reference occupancy for the target anatomy:

$$\text{ROI attention} = 100\,\frac{\sum_{j \in \mathcal{I}_{\text{visual}}} a_j o_j}{\sum_{j \in \mathcal{I}_{\text{visual}}} a_j}, \tag{19}$$

where $\mathcal{I}_{\text{visual}}$ indexes visual tokens.

E Additional experimental results

We provide additional results for CARA-VL using the same test sets and reference labels as the main evaluation. Table 11 reports four-grade classification results for LV size, EF, and EDV. Tables 12 and 13 provide detailed binary classification metrics, with external results reported separately against Rule-based and Report-based references. Table 14 further evaluates regional hypertrophy and wall-motion localisation.

Table 11: Four-grade QA classification by CARA-VL on the internal and external test sets. Grades are normal, mild, moderate, and severe. All scores are percentages.

Cohort	View	Reference	Task	N	ACC	BA	Macro-F_1
ACDC-50	SAX	Internal	LV size	171	78.4	63.9	62.1
ACDC-50	SAX	Internal	EF	171	79.5	59.3	53.4
M&Ms-2	LAX	Internal	EF	160	56.9	53.5	51.3
M&Ms-2	LAX	Internal	EDV	160	75.0	71.5	62.6
In-house B	SAX	Rule-based	LV size	1,783	70.1	41.4	31.6
In-house B	SAX	Rule-based	EF	1,795	52.4	51.6	44.4
In-house B	SAX	Report-based	EF	1,795	47.6	51.5	35.5

BA is macro recall over four grades; macro-F_1 gives equal weight to all four classes. Report-based LV size is binary and therefore has no four-grade result. Internal SAX scores retain the archived reporting precision; the remaining scores are recomputed from the corresponding saved predictions.

Table 12: Additional binary classification metrics for CARA-VL on the internal test sets, using the same evaluations as Table 1. All scores are percentages.

Task	N	ACC	Sens.	Spec.	PPV	BA	Macro-F_1
ACDC-50 (SAX)							
Hypertrophy	457	93.4	78.0	95.3	67.2	86.7	84.3
LV size	171	89.5	90.9	88.8	79.4	89.9	88.4
RV size	170	91.8	75.6	97.6	91.9	86.6	88.8
EF	171	90.6	98.5	85.7	81.2	92.1	90.4
GRS	160	93.8	89.8	95.5	89.8	92.6	92.6
GCS	148	95.9	98.1	94.7	91.4	96.4	95.7
LGE test (LGE)							
Scar presence	463	85.5	88.5	77.2	91.5	82.9	82.0
Scar transmurality	209	88.0	95.6	41.4	91.0	68.5	71.1
Scar extent	202	84.7	76.6	88.4	75.4	82.5	82.3
M&Ms-2 (LAX)							
EF	160	75.0	97.0	59.1	63.1	78.1	74.9
EDV	160	83.1	88.9	80.9	64.5	84.9	81.0
GLS	121	85.1	58.8	95.4	83.3	77.1	79.6

ACC denotes accuracy; Sens. and Spec. denote sensitivity and specificity; PPV denotes positive predictive value (precision). LV size, EF, and EDV are folded to normal versus any abnormal grade. Classification F_1 is macro-averaged over both classes. Positive labels are hypertrophy, cavity dilation, reduced function/motion, scar presence, transmural scar, and widespread scar, respectively.

Table 13: Additional binary classification metrics for CARA-VL on In-house B under the two reference standards. All scores are percentages.

Task	N	ACC	Sens.	Spec.	PPV	BA	Macro-F_1
Rule-based							
Hypertrophy	1,945	95.6	74.9	99.3	95.3	87.1	90.7
LV size	1,783	76.1	91.0	74.0	33.2	82.5	66.6
RV size	1,795	94.9	28.6	96.5	16.4	62.5	59.1
EF	1,795	66.5	92.6	54.4	48.5	73.5	66.3
GRS	1,708	95.1	77.7	96.9	71.8	87.3	86.0
GCS	1,684	85.6	82.5	85.8	26.2	84.1	65.8
Report-based							
Hypertrophy	1,936	81.9	40.0	99.7	98.3	69.9	72.7
LV size	1,780	60.9	45.6	82.3	78.3	64.0	60.7
RV size	1,778	68.4	7.3	97.6	59.2	52.4	46.8
EF	1,795	59.5	95.9	49.4	34.5	72.6	58.2
GRS*	1,708	79.5	29.0	96.4	72.9	62.7	64.5
GCS*	1,684	84.7	55.8	94.1	75.4	75.0	77.2

ACC denotes accuracy; Sens. and Spec. denote sensitivity and specificity; PPV denotes positive predictive value (precision). LV size, EF, and EDV are folded to normal versus any abnormal grade. Classification F_1 is macro-averaged over both classes. The two reference blocks score the same predictions using the task-specific coverage of Table 2. *Report GRS/GCS labels are wall-motion-score proxies, not measured strain. The paired-reference analysis additionally excludes missing or indeterminate reference fields (Appendix C.6).

Table 14: Additional regional-task metrics for CARA-VL. Precision, recall, and micro-F_1 are pooled over question–region/segment pairs; exact match and mean Jaccard compare the affected-region sets per QA. All scores are percentages.

Cohort	Reference	Task	N	Prec.	Rec.	Micro-F_1	Exact	Jaccard
ACDC-50	Internal	Hypertrophy	457	65.8	58.3	61.8	88.2	91.3
ACDC-50	Internal	Wall motion	171	51.1	61.5	55.8	57.9	71.1
In-house B	Rule-based	Hypertrophy	1,831	71.6	44.8	55.1	91.2	93.3
In-house B	Rule-based	Wall motion	1,795	33.9	40.1	36.7	76.7	80.2
In-house B	Report-based	Hypertrophy	1,822	94.9	18.5	30.9	74.8	78.0
In-house B	Report-based	Wall motion	1,795	69.4	22.9	34.5	69.4	74.6

N counts QA instances. Report-based hypertrophy covers septal and posterior regions; report-based wall motion covers the five mapped middle segments. Exact match requires equality of the predicted and reference affected-region sets within the scored coverage. Jaccard is one when both sets are empty. These are regional localisation metrics, not pixel-level segmentation scores. No regional metric is included in classification averages.