TRACE: TOWARDS RELIABLE, AUDITABLE CLINICAL EVIDENCE THROUGH TOOL-AUGMENTED REASONING FOR CINE CMR DIAGNOSIS

Anonymous authors
Paper under double-blind review

ABSTRACT

Cine cardiac magnetic resonance (CMR) assesses cardiac structure and function. Despite advances in vision-language models (VLMs), automated diagnosis requires verifiable quantitative evidence and explicit checks of diagnostic sufficiency to avoid premature conclusions. We introduce **TRACE** (**T**owards **R**eliable, **A**uditable **C**linical **E**vidence), a training-free agentic framework integrating adaptive tool use, clinical guideline retrieval, and independent evidence verification. Guided by clinical knowledge, the agent acquires CMR measurements, updates diagnostic hypotheses, and selects the next measurement. An independent verifier checks diagnostic requirements and returns evidence gaps for targeted acquisition and reassessment. Unresolved gaps lead to deferral for clinician review; accepted decisions produce structured reports linking findings, diagnostic rationale, clinical sources, and follow-up recommendations. On a development cohort of 50 balanced ACDC cases, TRACE achieves 87.3% mean accuracy across eight seeds with 96.3% diagnosis coverage, compared with 85.3% for ACDC-finetuned CineMA and 16%–28% for eight single-pass VLMs. Its report compiler achieves 100% specification conformance versus 46% for an LLM given the same patient evidence. These results demonstrate the value of evidence-guided agentic reasoning for auditable cine CMR diagnosis and reporting without additional model training.

1 INTRODUCTION

Cine cardiac magnetic resonance (CMR) captures cardiac structure across short-axis slices and cardiac motion throughout the cardiac cycle. Interpreting a CMR study requires integrating image quality, ventricular volumes and function, myocardial morphology, and regional wall motion (Schulz-Menger et al., 2020; Kramer et al., 2020). Together, these findings help resolve diagnoses with overlapping presentations: a reduced left ventricular ejection fraction can accompany both dilated cardiomyopathy (DCM) and myocardial infarction (MINF), whereas regional motion patterns help distinguish global from focal dysfunction. Automated diagnosis must therefore determine which findings to acquire, how to interpret them, and when they are sufficient to support a conclusion.

Existing approaches, however, largely collapse these steps into a single prediction. Direct image-based classifiers output a disease label without exposing the measurements that underlie it (Figure 1a). Vision–language models (VLMs) go a step further by generating a textual rationale (Figure 1b), yet such text offers no guarantee that ventricular size, wall thickness, or contraction was actually measured across the relevant slices and phases. When patient observations and diagnostic interpretation are produced in the same step, even a plausible explanation remains difficult to verify.

Medical agents take a step toward separating observation from interpretation: they can invoke specialist image-analysis tools and reason over the resulting outputs (Li et al., 2024; Fallahpour et al., 2025). More recent systems further incorporate clinical guidelines and explicit visual evidence into this process (Wang et al., 2026; Du et al., 2026). However, acquiring some measurements does not ensure that the evidence required for a proposed diagnosis is complete. For example, normal ventricular volumes and function may favor a normal cine phenotype, yet a negative assessment remains incomplete without the corresponding wall-thickness and regional-motion findings. An agent therefore needs feedback that identifies such evidence gaps before its proposed diagnosis can be released (Figure 1c).

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.

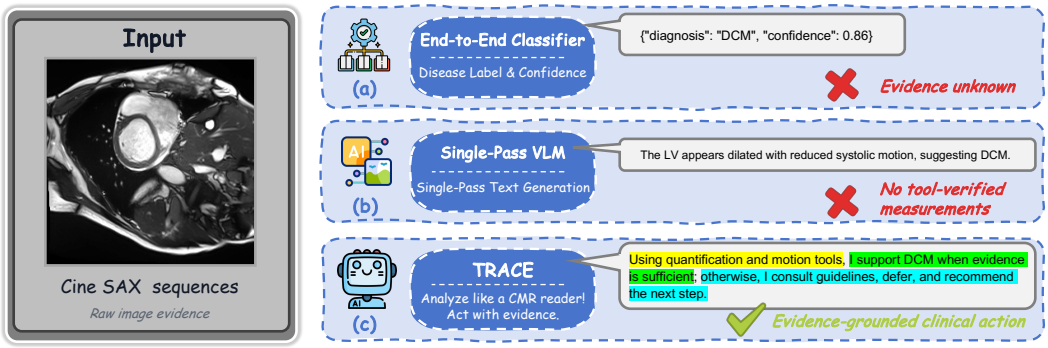

Figure 1: **Three routes from cine CMR to diagnosis.** (a) Image-based classifiers predict a disease label without an explicit evidence-acquisition process. (b) Single-pass VLMs generate a diagnosis and rationale without independently verified quantitative measurements. (c) TRACE adaptively acquires evidence, retrieves clinical guidance, and verifies diagnoses to produce structured reports and follow-up recommendations.

We introduce TRACE (Towards Reliable, Auditable Clinical Evidence), a training-free diagnostic agent that coordinates measurement, clinical guidance, and independent evidence verification. Tools derive findings from cine CMR segmentations with quality and provenance records; retrieved guidance interprets them and identifies diagnosis-specific requirements. The Agent updates its differential, selects measurements, and refreshes guidance as evidence accumulates. An independent verifier checks its proposal and returns gaps for targeted acquisition or, when unresolvable, deferral for clinician review. Accepted decisions become structured reports linking supporting measurements, diagnostic rationale, clinical sources, and follow-up recommendations.

We evaluate TRACE on 50 balanced ACDC cases using CineMA segmentations. Across eight runs, TRACE achieves 87.3% diagnostic accuracy at 96.3% diagnosis coverage, with the remaining trajectories ending in explicit deferral. By comparison, ACDC-finetuned CineMA achieves 85.3% over three runs, while eight single-pass VLMs, each run eight times, achieve only 16%–28%. Beyond accuracy, TRACE reliably recovers from rejection: in a separate three-seed analysis, all 84 trajectories containing a rejected proposal reach an admissible action. Finally, in matched post-lock recommendation generation, the deterministic selector meets the release specification in 100% of cases, compared with 46% for an LLM given the same locked records and applicable cards. Together, these results connect diagnostic performance with recovery from rejection and reliable report release.

In sum, our contributions are threefold:

- **A training-free agentic workflow for cine CMR diagnosis.** We introduce TRACE, which coordinates quantitative measurement, source-linked clinical guidance, and independent evidence verification to produce evidence-supported diagnoses and structured reports.
- **Guideline-informed evidence acquisition and verification.** TRACE maps Tool-derived findings and retrieved guidance to phenotype-specific Evidence Contracts. The Agent acquires missing measurements and reassesses the case after verifier feedback; accepted decisions and their supporting evidence are locked for report release.
- **Empirical results.** On 50 ACDC cases, TRACE achieves 87.3% diagnostic accuracy, comparable to the ACDC-finetuned CineMA classifier (85.3%). In matched recommendation generation after locking, selector conformance is 100%, versus 46% for the LLM.

2 Related Work

Sequential diagnosis and medical tool use. Sequential diagnosis reveals information over multiple turns (Nori et al., 2025). LA-CDM requests tests from a working hypothesis (Bani-Harouni et al., 2026), while KnowGuard examines external evidence before abstaining (Dang et al., 2026). Medical agents select specialist tools for clinical tasks (Li et al., 2024; Fallahpour et al., 2025), organize collaborating LLMs (Kim et al., 2024), and navigate pathology images across scales (Sun et al.,

2025). TRACE applies sequential control to cine CMR measurement: the Agent chooses which findings to acquire, and the final proposal is checked against the resulting patient evidence.

Clinical and visual evidence grounding. RAG connects generation to retrieved sources (Lewis et al., 2020), and Self-RAG learns when to retrieve and critique (Asai et al., 2024). RAD brings disease knowledge into multimodal diagnosis (Li et al., 2025); MedAgent-Pro incorporates guidance into agentic clinical reasoning (Wang et al., 2026). CARE localizes visual evidence and feeds it back into medical VQA through coordinated specialist models (Du et al., 2026). TRACE grounds cine CMR decisions in typed, segmentation-derived measurements. Retrieved guidance interprets those patient findings, identifies diagnosis-specific requirements, and is refreshed as the evidence changes.

Verification, deferral, and process evaluation. Selective prediction and learning to defer govern whether a model emits an answer or delegates it (Geifman & El-Yaniv, 2019; Mozannar & Sontag, 2020). Policy certificates bound an RL policy's expected performance (Dann et al., 2019). TRACE verifies a submitted diagnosis or deferral against the current evidence before release, returns resolvable gaps for further measurement, and binds an accepted decision to its supporting evidence. DrVD-Bench and Med-CMR assess medical evidence comprehension and reasoning (Zhou et al., 2025; Gong et al., 2026); our ACDC evaluation also measures acquisition, recovery, and report conformance.

3 TRACE: Evidence-Guided Diagnostic Reasoning

Reliable cine CMR diagnosis requires more than producing a label: the relevant evidence must be acquired, interpreted against clinical knowledge, and confirmed as sufficient before a conclusion is released. TRACE realizes this process by coupling adaptive measurement tools with clinical evidence grounding and independent verification under agentic control. As a result, diagnosis proceeds through three stages: evidence acquisition, evidence assessment, and report release.

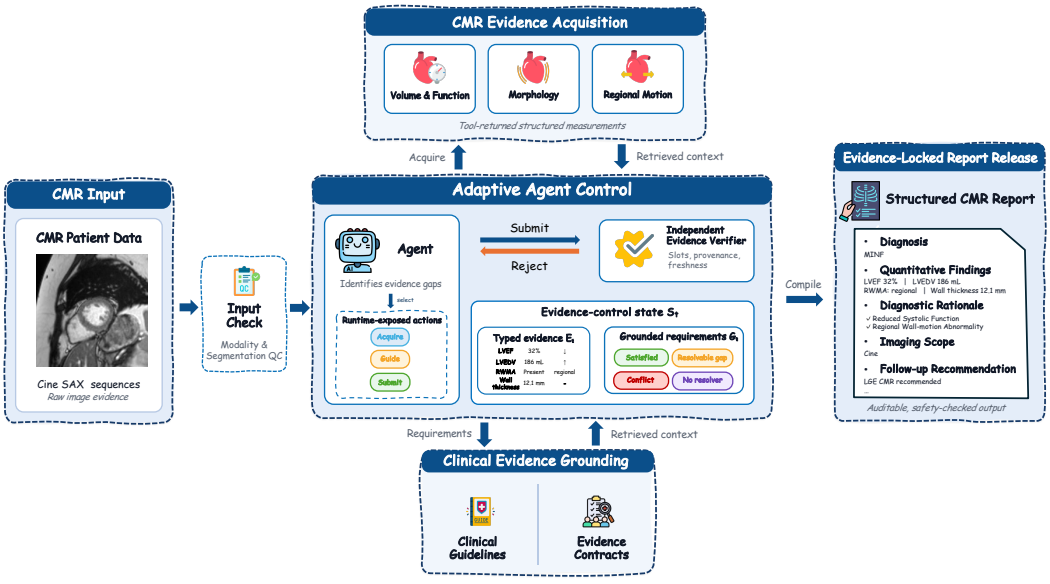

Figure 2: **TRACE framework.** TRACE comprises an adaptive Agent, measurement tools, clinical evidence grounding, and an independent evidence verifier. The Agent acquires quantitative CMR findings and retrieves diagnostic guidance, while the verifier checks submitted proposals and returns unmet requirements for further acquisition. Accepted decisions and supporting findings are compiled into structured reports with source-linked recommendations.

Method overview. Figure 2 illustrates the overall framework. Before diagnosis begins, cine CMR data pass through **CMR Input Control**, which checks modality availability and segmentation quality. Validated inputs then enter a diagnostic reasoning loop, which is organized around the following five components. **(1) Adaptive Agent Control.** At the center of the loop, the Agent maintains an evidence-control state S_t and, at each step, selects among the runtime-available actions: acquiring evidence,

requesting clinical grounding, or submitting a terminal proposal. **(2) CMR Evidence Acquisition.** When the Agent chooses to acquire evidence, Measurement Tools return patient-specific functional and structural findings as typed evidence $\mathcal{E}_t$, together with quality and provenance information. **(3) Clinical Evidence Grounding.** When the Agent requests guidance, clinical sources are applied to the current evidence to produce diagnostic guidance G_t, which summarizes supporting findings, conflicts, unresolved gaps, and candidate next actions. Both $\mathcal{E}_t$ and G_t then feed back into the Agent's subsequent action selection. **(4) Independent Evidence Verification.** Once the Agent submits a terminal proposal, either a diagnosis or a deferral, the proposal is checked against phenotype-specific Evidence Contracts. If it is rejected, the unmet requirements are returned to the Agent for targeted recovery. **(5) Evidence-Locked Report Release.** Finally, each accepted proposal is locked together with its supporting evidence and compiled into a structured CMR report.

3.1 CMR Evidence Acquisition and Typed Evidence

CMR Measurements. CMR Input Control first checks whether the required CMR inputs are available. For available cine studies, the evidence-acquisition backend preprocesses the short-axis (SAX) sequences and applies CineMA (Fu et al., 2026) to obtain ventricular segmentation masks. The processed cine images and predicted masks then undergo joint quality control before measurement actions are exposed. When requested by the Agent, deterministic measurement tools combine the segmentations with spatial and patient metadata to derive patient-specific evidence on ventricular volume and function, morphology, and regional contraction. These measurements include LV and RV volumes and ejection fractions, LV mass and wall thickness, and segmentation-derived regional wall-motion findings. The Agent receives these structured measurements, not images or masks.

Typed evidence. Each Tool return is converted into a typed evidence record containing the measured value and unit together with its producing action, Tool version, execution and quality status, and provenance. The runtime appends these records to the evidence store $\mathcal{E}_t$ as measurements are acquired and identifies the current evidence snapshot by σ_t, including its quality and provenance metadata. Diagnostic reasoning uses current records with complete provenance that satisfy the runtime's quality-admissibility policy:

$$\mathcal{E}_t^{\text{use}} = \text{Usable}(\mathcal{E}_t, \sigma_t). \tag{1}$$

Missing, failed, unavailable, or quality-blocked measurements remain explicit gaps, not negative findings. This yields traceable patient evidence for **Clinical Evidence Grounding**.

3.2 Clinical Evidence Grounding

Clinical Sources. The clinical corpus contains phenotype-specific source cards derived from clinical guidelines and reporting recommendations. When the Agent requests guidance, Clinical Evidence Grounding returns these cards together with interpretations of the current patient evidence. Each card links a clinical source to the findings, diagnostic rules, and measurements relevant to that phenotype.

Diagnostic Guidance. Grounding applies diagnostic rules to usable evidence $\mathcal{E}_t^{\text{use}}$, identifying support, conflicts, and unresolved requirements for candidate diagnoses. When possible, it links each gap to a measurement Tool that can acquire the missing patient evidence.

We denote this case-specific guidance by

$$G_t = \text{Ground}(\mathcal{E}_t^{\text{use}}, D_t; \mathcal{K}_{\text{diag}}). \tag{2}$$

For example, a NOR candidate with normal ventricular function but without wall-motion or wall-thickness assessment remains incomplete. Grounding retains these findings as unresolved and identifies the corresponding measurements as candidate next actions. Changes to patient evidence or the differential trigger a guidance refresh for the updated case state.

3.3 Adaptive Agent Control

The Agent controls the diagnostic process. At each step, it decides whether to collect more patient evidence, update the clinical guidance, or submit a conclusion.

Agent state. At step t, the Agent's state is

$$S_t = \big(\mathcal{E}_t^{\text{use}},\ G_t,\ D_t\big), \tag{3}$$

where $\mathcal{E}_t^{\text{use}}$ is the usable patient evidence acquired so far, G_t is the most recent diagnostic guidance (Section 3.2), and D_t is the current differential, i.e., the candidate diagnoses under consideration.

Action space. Not every action is valid at every step. Given S_t, the runtime determines the set of currently available actions $\mathcal{A}(S_t)$, and the Agent selects one of them:

$$a_t \in \mathcal{A}(S_t) \subseteq \{\textsc{Acquire}, \textsc{Guide}, \textsc{Submit}\}. \tag{4}$$

The three actions affect the state as follows:

- ACQUIRE requests one additional CMR measurement from the Measurement Tools. The returned findings are added to the patient evidence and filtered for usability, yielding $\mathcal{E}_{t+1}^{\text{use}}$.
- GUIDE re-applies clinical sources to the current evidence and differential, yielding updated guidance G_{t+1}, including the evidence gaps that remain unresolved.
- SUBMIT sends a terminal proposal for independent verification: either a diagnosis $y \in \mathcal{Y} = \{\text{NOR}, \text{DCM}, \text{HCM}, \text{MINF}, \text{RV}\}$ (normal, dilated cardiomyopathy, hypertrophic cardiomyopathy, myocardial infarction, and abnormal right ventricle) or a deferral that names the unresolved evidence gap.

Algorithm 1 TRACE diagnostic loop. State components not updated at step t carry over to $t{+}1$.

Require: Study x; corpus $\mathcal{K}_{\text{diag}}$; Contracts $\mathcal{C}$; budget T
Run input control on x; initialize snapshot σ_0
$\mathcal{E}_0 \leftarrow \emptyset$; initialize G_0, D_0
for $t = 0, \ldots, T-1$ **do**
  $\mathcal{E}_t^{\text{use}} \leftarrow \text{Usable}(\mathcal{E}_t, \sigma_t)$
  $S_t \leftarrow (\mathcal{E}_t^{\text{use}}, G_t, D_t)$
  $a_t \leftarrow \textsc{Agent}(S_t), \quad a_t \in \mathcal{A}(S_t)$
  if $a_t = \textsc{Acquire}(m)$: $\mathcal{E}_{t+1} \leftarrow \mathcal{E}_t \cup \textsc{Tool}_m(x)$
  if $a_t = \textsc{Guide}$:
  $G_{t+1} \leftarrow \text{Ground}(\mathcal{E}_t^{\text{use}}, D_t; \mathcal{K}_{\text{diag}})$
  if $a_t = \textsc{Submit}(p)$ **then**
    $(ok, U) \leftarrow \textsc{Verify}(p, S_t; \mathcal{C})$
    if ok **then**
      lock p and its supporting evidence
      return $\big(p, \textsc{Compile}(p, \mathcal{E}_t^{\text{use}})\big)$
    else
      provide U to the Agent and continue
    end if
  end if
  Agent revises D_{t+1}
end for
return INVALID ▷ budget exhausted

Decision loop. The Agent iteratively acquires measurements, refreshes guidance, and submits diagnostic proposals (Algorithm 1). After each ACQUIRE or GUIDE action, the Agent observes the updated state S_{t+1}, revises its differential D_{t+1} if needed, and selects the next action. This process repeats until the Agent chooses SUBMIT. The proposal is then passed to Independent Evidence Verification (Section 3.4). If the proposal is rejected, the unmet requirements are returned to the Agent, which resumes the loop to address them. Acceptance ends the loop and releases the report.

3.4 INDEPENDENT EVIDENCE VERIFICATION

Evidence Contracts. Each phenotype's Evidence Contract specifies the required evidence and findings that prevent acceptance. These operational criteria combine source-linked clinical guidance, ACDC phenotype conventions, and task-specific evidence requirements. The full criteria, their clinical sources, and the implemented phenotype thresholds are specified in the Appendix.

Proposal Verification. The Agent proposes a diagnosis or deferral, and an independent verifier checks the proposal against the corresponding Evidence Contract. A diagnosis is accepted when the required evidence is satisfied and no conflicting finding prevents acceptance. A deferral is accepted when a required evidence gap remains unresolved and cannot be addressed by the available actions.

When verification fails, the verifier returns the unmet requirements and the available actions associated with those gaps. The Agent can then acquire additional measurements, refresh clinical grounding, and reassess the case. Accepted proposals proceed to **Evidence-Locked Report Release**.

3.5 EVIDENCE-LOCKED REPORT RELEASE

Evidence Locking. Once a diagnosis or deferral passes verification, TRACE locks the accepted decision together with its supporting evidence. The report cannot change the diagnosis or replace Tool-derived measurements. The compiler uses locked evidence to assemble findings and rationale.

Structured Report. The compiler converts the locked case record into a structured CMR report with the final diagnosis or deferral, quantitative findings, diagnostic rationale, and supporting clinical

sources. Follow-up recommendations are selected only after the decision is locked. They use the accepted diagnosis and remaining evidence gaps to suggest further imaging, clinical workup, manual review, or clinical correlation when needed. These recommendations do not reopen the diagnostic decision. The report links the released decision to its supporting patient evidence.

4 EXPERIMENTS

We evaluate TRACE on cine CMR diagnosis and address four research questions:

RQ1: How does TRACE compare with image-based models and single-pass VLMs? **RQ2:** What do adaptive control and dynamic guidance contribute to diagnosis? **RQ3:** Can TRACE recover from rejected diagnostic proposals? **RQ4:** Does evidence-locked compilation improve report reliability?

Dataset and inputs. We evaluate all methods on a balanced 50-case cohort from the ACDC (Bernard et al., 2018) testing partition, with ten cases per phenotype. This cohort was also used during method development; the evaluation characterizes within-cohort performance and controlled workflow comparisons. Image-based baselines use all short-axis slices at end diastole (ED) and end systole (ES), with model-specific preprocessing and input formats. Single-pass VLMs receive these inputs as phase- and slice-identified images or paired ED/ES volumes, depending on their interfaces. TRACE receives Tool-derived measurements from these studies and adaptively requests measurements, without viewing images or masks.

Baselines. We compare TRACE with three groups of baselines. **Image-based classifiers** include 3D ResNets, zero-shot CMR-CLIP, and ACDC-finetuned CineMA models. **Single-pass VLMs** include general-purpose and medical models that directly predict one diagnosis from the ED/ES cine inputs in a single inference pass. **Measurement-based controls** use the same Tool-derived evidence as TRACE but remove key elements of the adaptive workflow, including sequential control or online diagnostic guidance. The Appendix details model-specific inputs, training, and inference.

Implementation Details. TRACE uses Qwen3-8B (Yang et al., 2025) for diagnosis, CineMA (Fu et al., 2026) segmentations for measurement, and a fixed clinical corpus for grounding, without additional model training. The Appendix specifies serving, inference, and action budgets.

4.1 MAIN RESULTS

Table 1 answers RQ1 by comparing TRACE with image-based classifiers and single-pass VLMs. Across eight runs, TRACE achieves 87.25% accuracy and 96.25% diagnosis coverage, with no invalid trajectories. Its diagnostic accuracy is comparable to ACDC-finetuned CineMA, which averages 85.33% across three checkpoints, and substantially exceeds the single-pass VLMs (16.00%–28.00%).

Tool-based evidence-guided reasoning can match strong specialist image classifiers while exposing the diagnostic trajectory through measurement, grounding, and verification.

4.2 ADAPTIVE CONTROL AND DIAGNOSTIC GUIDANCE

In the following comparisons, blue rows denote control configurations and green rows denote TRACE configurations or components.

Single-pass versus agentic reasoning. To address RQ2, we compare the full TRACE workflow with single-pass Qwen3-8B (Table 2). The single-pass baseline receives complete measurements and a fixed diagnostic-rule summary specifying phenotype thresholds and evidence requirements. The full workflow improves accuracy from 73.3% to 86.7% and diagnosis coverage from 80.0% to 95.3%. To assess the backend's diagnostic capability, we also apply its rules and Evidence

Table 2: **Diagnosis from measurement evidence.** Qwen3-8B and TRACE use three matched runs on 50 cases; the rule-based reference uses one run. Invalid counts invalid outputs across all runs.

Method	Runs	Acc. (%) ↑	Cov. (%)	Invalid ↓
Qwen3-8B (Yang et al., 2025)	3	73.3	80.0	2/150
TRACE	3	86.7	95.3	0/150
Rule-based classifier	1	88.0	94.0	0/50

Table 1: **Diagnostic performance on ACDC.** We report accuracy and Macro-F1 averaged across the listed runs on 50 cases. Cov. is the proportion of case-runs with a diagnostic label; deferrals and invalid or missing labels count as incorrect in classification metrics. Medical VLMs are highlighted in gray and TRACE in green. Bold marks the best classification score within each group. Repetitions denote training seeds (3D classifiers), checkpoints (CineMA), or inference executions (CMR-CLIP and VLMs). Exact model releases and input adaptations are specified in Section A.

		Classification		Output
Method	**Runs**	**Acc. (%) ↑**	**Macro-F1 ↑**	**Cov. (%)**
(a) Direct image-based classifiers				
3D adaptation of ResNet-50 (random initialization) (He et al., 2016)	3	50.00	0.4656	100.00
3D ResNet-18 (MedicalNet initialization) (Chen et al., 2019)	3	58.67	0.5773	100.00
3D ResNet-50 (MedicalNet initialization) (Chen et al., 2019)	3	41.33	0.3656	100.00
CMR-CLIP (zero-shot matching) (Nakashima et al., 2026)	3	68.00	0.6617	100.00
CineMA (ACDC-finetuned) (Fu et al., 2026)	3	**85.33**	**0.8447**	100.00
(b) Single-pass generative VLMs				
Qwen3-VL-8B-Instruct (Bai et al., 2025)	8	20.00	0.0667	100.00
LLaVA-Med-v1.5-Mistral-7B (Li et al., 2023)	8	20.00	0.0667	100.00
HuatuoGPT-Vision-7B (Chen et al., 2024)	8	20.00	0.0690	100.00
Lingshu-7B (Xu et al., 2026)	8	20.00	0.0690	100.00
BiMediX2-8B (Mullappilly et al., 2025)	8	**28.00**	0.2108	98.00
MedVLM-R1 (Pan et al., 2025)	8	24.00	**0.2254**	78.00
Hulu-Med-7B (Jiang et al., 2025)	8	16.00	0.0788	100.00
M3D-LaMed-Phi-3-4B (dual-volume) (Bai et al., 2024)	8	20.00	0.0690	100.00
(c) Evidence-based agent workflow (Qwen3-8B)				
TRACE	8	**87.25**	**0.8772**	96.25

Contracts to complete measurements without an LLM, obtaining 88.0% accuracy. This reference demonstrates the diagnostic value of the measurement and rule backend, while the single-pass comparison shows the benefit of organizing diagnosis through TRACE's adaptive workflow.

Table 3: **Evidence acquisition policy.** Selective acquisition versus complete evidence exposure. Counts are averaged per trajectory: Opt., optional evidence actions; Guide, diagnostic guidance calls; Model, LLM calls.

		Counts / trajectory		
Policy	**Acc. (%) ↑**	**Opt. ↓**	**Guide ↓**	**Model ↓**
Complete	**92.0**	2.00	**1.73**	**12.67**
Selective	86.7	**1.23**	2.89	13.27

Table 4: **Dynamic diagnostic guidance.** Differences are dynamic minus static. Count differences compare means per trajectory; pp denotes percentage points.

Outcome	**Δ**
Accuracy	−2.0 pp
First-proposal acceptance	+24.0 pp
Proposal rejections	−0.95
Guidance calls	−1.39
Model calls	−2.91

Selective evidence acquisition. To examine evidence selection within RQ2, we compare selective acquisition with complete evidence exposure over 150 matched pairs (Table 3). Selective acquisition reduces optional evidence actions by 38.3%, but yields lower diagnostic accuracy. The trajectories correct only under complete exposure involve MINF cases. In these cases, the Agent stopped before acquiring regional-motion findings needed to distinguish the diagnosis.

Dynamic diagnostic guidance. We next examine whether updating guidance from the current evidence improves diagnostic proposals, comparing dynamic and static guidance over 150 matched pairs (Table 4). Dynamic guidance increases first-proposal acceptance from 18.7% to 42.7% and reduces rejections and model calls, helping the Agent meet diagnostic requirements with less repair.

4.3 VERIFICATION AND RECOVERY

Recovery after rejection. To address RQ3, we examine recovery from rejected proposals (Table 5). The verifier checks each proposal against its Evidence Contract; even a correct diagnosis requires complete evidence for release. The Agent recovers by revising an incorrect diagnosis or retaining a correct diagnosis while acquiring the missing evidence. All 84 trajectories encountering rejection reach an admissible action: 76 correct diagnoses and eight accepted deferrals. Separately, 12 incorrect first proposals are accepted without rejection.

Table 5: **Diagnostic transitions after rejection.** First diagnostic proposals and final outcomes for trajectories with rejected proposals. Diagnostic correctness is assessed against the reference label.

First diagnosis	Final outcome	Count
Incorrect	Correct	47
Correct	Correct	29
Incorrect	DEFER	8
Correct	Incorrect	0

Table 6: **Effect of verification feedback.** Correct denotes final correct diagnoses on 43 primary pairs. Other columns report means per trajectory across all 150 pairs: Rej., proposal rejections; Guide, diagnostic guidance calls; Model, LLM calls.

		Counts / trajectory ↓		
Feedback	**Correct ↑**	**Rej.**	**Guide**	**Model**
Generic	**39/43**	1.10	3.09	13.66
Structured	38/43	**0.83**	**2.79**	**13.03**

Table 7: **Recommendation release conformance.** Entries count cases passing each check out of 50. Conform. denotes overall conformance; Scope and Align. denote applicable-source and category-alignment checks, respectively.

	Cases / 50 ↑			
Generator	**Complete**	**Conform.**	**Scope**	**Align.**
LLM	42	23	31	29
Selector	**50**	**50**	**50**	**50**

Structured verification feedback. We next ask whether explaining a rejection helps the Agent recover (Table 6). With the Agent and verifier held fixed, generic feedback reports that the proposal was rejected, whereas structured feedback also identifies unmet evidence requirements and available actions to address them. The primary comparison uses 43 pairs matched on the pre-intervention state. Final diagnostic correctness is similar on this subset. Across all 150 whole-trajectory pairs, the structured-feedback configuration uses fewer rejections, guidance calls, and model calls. Call counts cover the full trajectories, which were run independently under the two feedback conditions.

4.4 EVIDENCE-LOCKED REPORT RELEASE

Report compilation. To address RQ4, we compare deterministic recommendation selection with Qwen3-8B generation using the same 50 locked records and applicable cards (Table 7). Under the shared operational checks, deterministic selection conforms in 100% of cases, compared with 46% for single-request LLM generation. Of the LLM outputs, eight reach the token limit and 19 completed outputs fail at least one check. Separately, all 400 full-workflow reports preserve the accepted decision and Tool-derived facts; detailed checks are reported in the Appendix (Table 21).

4.5 QUALITATIVE CASE STUDY

We showcase TRACE on patient123 from the ACDC testing partition in Figure 3. The Agent obtains measurements from cardiac segmentations and uses retrieved guidance to propose NOR. Verification identifies missing regional wall-motion evidence, prompting the Agent to acquire it and refresh guidance. The diagnosis remains NOR, now supported for release in a structured report with diagnostic rationale, clinical sources, and follow-up recommendations.

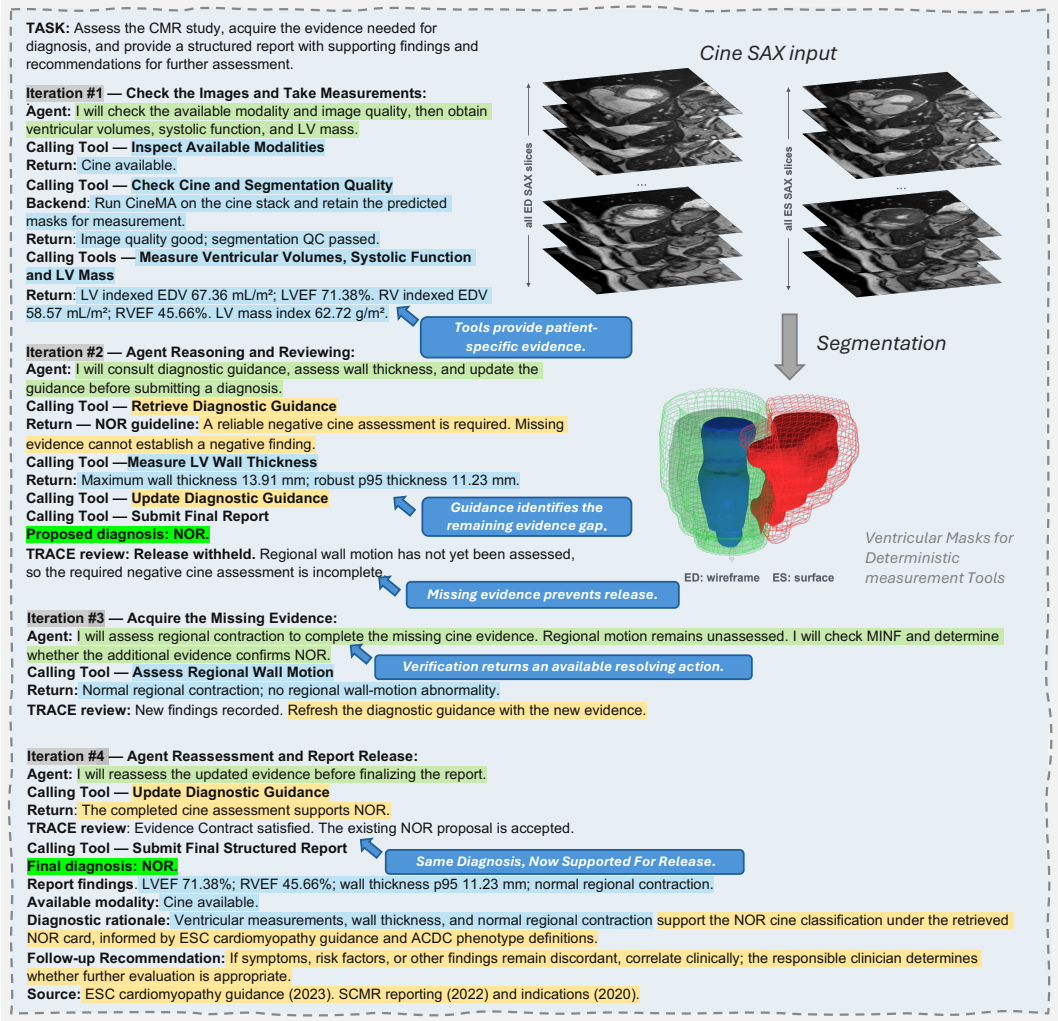

Figure 3: **Case study.** A summarized online diagnostic trajectory. Tool calls and patient-specific findings are highlighted in blue, Agent reasoning in green, and clinical guidance and diagnostic requirements in yellow.

5 Discussion and Conclusion

In this paper, we present TRACE, a training-free agentic framework that grounds cine CMR diagnosis in measurable evidence and clinical guidance. Using a general-purpose language model, TRACE substantially outperforms single-pass VLM inference and achieves diagnostic accuracy comparable to the ACDC-finetuned CineMA classifier. These findings highlight the value of evidence acquisition and decision control for automated cine CMR diagnosis. Decoupling evidence acquisition, diagnostic reasoning, and release verification makes diagnoses and deferrals traceable to supporting evidence and unresolved gaps. This supports accountable clinical review through evidence-linked diagnostic rationale and guideline-informed recommendations for further assessment.

Ethics Statement

This study uses the publicly available, anonymized ACDC dataset (Bernard et al., 2018). According to the dataset providers, the original clinical data were anonymized and handled under the regulations of the local ethics committee at the University Hospital of Dijon (ACDC Challenge, n.d.). Our study involves secondary analysis without new patient recruitment or clinical intervention. TRACE is evaluated as a research system; its reports and follow-up recommendations were not used to guide patient care.

AI Use Statement

We used generative AI tools to assist with translation and language polishing, literature search and summarization, feedback on experimental design, coding assistance, and drafting portions of the manuscript. The authors reviewed and revised the AI-assisted material, verified references and reported results against the original sources and experimental records, and take full responsibility for the final work.

Reproducibility Statement

The main text describes the TRACE workflow and evaluation setup. The appendices detail baseline configurations, measurement methods, clinical sources and Evidence Contracts, inference settings, experimental protocols, and statistical procedures. Experiments use the publicly available ACDC dataset (ACDC Challenge, n.d.).

References

ACDC Challenge. ACDC Dataset, n.d. URL https://www.creatis.insa-lyon.fr/Challenge/acdc/databases.html. Accessed September 26, 2026.

Elena Arbelo, Alexandros Protonotarios, Juan R. Gimeno, et al. 2023 ESC Guidelines for the management of cardiomyopathies. *European Heart Journal*, 44(37):3503–3626, 2023. doi:10.1093/eurheartj/ehad194.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avi Sil, and Hannaneh Hajishirzi. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In *International Conference on Learning Representations*, volume 2024, pp. 9112–9141, 2024.

Fan Bai, Yuxin Du, Tiejun Huang, Max Q-H Meng, and Bo Zhao. M3D: Advancing 3D medical image analysis with multi-modal large language models. *arXiv preprint arXiv:2404.00578*, 2024.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025.

David Bani-Harouni, Chantal Pellegrini, Ege Özsoy, Nassir Navab, and Matthias Keicher. Language Agents for Hypothesis-driven Clinical Decision Making with Reinforcement Learning. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 156240–156258, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/fcedb4cd8eea1ac0f139228376e4348d-Paper-Conference.pdf.

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. doi:10.1109/TMI.2018.2837502.

Junying Chen, Chi Gui, Ruyi Ouyang, Anningzhe Gao, Shunian Chen, Guiming Hardy Chen, Xidong Wang, Zhenyang Cai, Ke Ji, Guangjun Yu, Xiang Wan, and Benyou Wang. Towards

Injecting Medical Visual Knowledge into Multimodal LLMs at Scale. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 7346–7370, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi:10.18653/v1/2024.emnlp-main.418. URL https://aclanthology.org/2024.emnlp-main.418/.

Sihong Chen, Kai Ma, and Yefeng Zheng. Med3D: Transfer learning for 3D medical image analysis. *arXiv preprint arXiv:1904.00625*, 2019.

Xilin Dang, Kexin Chen, Xiaorui Su, Ayush Noori, Iñaki Arango, Lucas Vittor, Xinyi Long, Yuyang Du, Marinka Zitnik, and Pheng-Ann Heng. KnowGuard: Knowledge-driven abstention for multi-round clinical reasoning. In *International Conference on Learning Representations*, volume 2026, pp. 81773–81810, 2026.

Christoph Dann, Lihong Li, Wei Wei, and Emma Brunskill. Policy certificates: Towards accountable reinforcement learning. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 1507–1516. PMLR, 2019.

Yuexi Du, Jinglu Wang, Shujie Liu, Nicha C Dvornek, and Yan Lu. CARE: Towards Clinical Accountability in Multi-Modal Medical Reasoning with an Evidence-Grounded Agentic Framework. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 10666–10700, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/11fc8c98b46d4cbdfe8157267228f7d7-Paper-Conference.pdf.

Delafield Du Bois and Eugene F Du Bois. Clinical calorimetry: Tenth paper a formula to estimate the approximate surface area if height and weight be known. *Archives of Internal Medicine*, 17 (6_2):863–871, 1916. doi:10.1001/archinte.1916.00080130010002.

Adibvafa Fallahpour, Jun Ma, Alif Munim, Hongwei Lyu, and Bo Wang. MedRAX: Medical Reasoning Agent for Chest X-ray. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 15661–15676. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/fallahpour25a.html.

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*, 6:408, 2026. doi:10.1038/s43856-026-01636-0.

Yonatan Geifman and Ran El-Yaniv. SelectiveNet: A deep neural network with an integrated reject option. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2151–2159. PMLR, 2019.

Haozhen Gong, Xiaozhong Ji, Yuansen Liu, Wenbin Wu, Xiaoxiao Yan, Jingjing Liu, Kai Wu, Jiazhen Pan, Bailiang Jian, Jiangning Zhang, et al. Med-CMR: A fine-grained benchmark integrating visual evidence and clinical logic for medical complex multimodal reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 41224–41234, 2026.

David J. Hand and Robert J. Till. A Simple Generalisation of the Area Under the ROC Curve for Multiple Class Classification Problems. *Machine Learning*, 45:171–186, 2001. doi:10.1023/A:1010920819831.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016. doi:10.1109/CVPR.2016.90.

Sture Holm. A Simple Sequentially Rejective Multiple Test Procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979. URL https://www.jstor.org/stable/4615733.

W. Gregory Hundley, David A. Bluemke, Jan Bogaert, et al. Society for Cardiovascular Magnetic Resonance (SCMR) guidelines for reporting cardiovascular magnetic resonance examinations. *Journal of Cardiovascular Magnetic Resonance*, 24:29, 2022. doi:10.1186/s12968-021-00827-z.

Songtao Jiang, Yuan Wang, Sibo Song, Tianxiang Hu, Chenyi Zhou, Bin Pu, Yan Zhang, Zhibo Yang, Yang Feng, Joey Tianyi Zhou, et al. Hulu-Med: A transparent generalist model towards holistic medical vision-language understanding. *arXiv preprint arXiv:2510.08668*, 2025.

Yubin Kim, Chanwoo Park, Hyewon Jeong, Yik Siu Chan, Xuhai Xu, Daniel McDuff, Hyeonhoon Lee, Marzyeh Ghassemi, Cynthia Breazeal, and Hae Won Park. MDAgents: An adaptive collaboration of LLMs for medical decision-making. In *Advances in Neural Information Processing Systems*, volume 37, pp. 79410–79452, 2024. doi:10.52202/079017-2522.

Michael C. Kontos, James A. de Lemos, Steven B. Deitelzweig, Deborah B. Diercks, M. Odette Gore, Erik P. Hess, Cian P. McCarthy, James K. McCord, Paul I. Musey, Jr., Todd C. Villines, and Leesa J. Wright. 2022 ACC Expert Consensus Decision Pathway on the Evaluation and Disposition of Acute Chest Pain in the Emergency Department: A Report of the American College of Cardiology Solution Set Oversight Committee. *Journal of the American College of Cardiology*, 80(20):1925–1960, 2022. doi:10.1016/j.jacc.2022.08.750.

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. doi:10.1186/s12968-020-00607-1.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient Memory Management for Large Language Model Serving with PagedAttention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*, pp. 611–626, 2023. doi:10.1145/3600006.3613165.

Tim Leiner, Jan Bogaert, Matthias G Friedrich, Raad Mohiaddin, Vivek Muthurangu, Saul Myerson, Andrew J Powell, Subha V Raman, and Dudley J Pennell. SCMR Position Paper (2020) on clinical indications for cardiovascular magnetic resonance. *Journal of Cardiovascular Magnetic Resonance*, 22(1):76, 2020. doi:10.1186/s12968-020-00682-4.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, volume 33, pp. 9459–9474, 2020.

Binxu Li, Tiankai Yan, Yuanting Pan, Jie Luo, Ruiyang Ji, Jiayuan Ding, Zhe Xu, Shilong Liu, Haoyu Dong, Zihao Lin, et al. MMedAgent: Learning to use medical tools with multi-modal agent. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 8745–8760, 2024. doi:10.18653/v1/2024.findings-emnlp.510.

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. In *Advances in Neural Information Processing Systems*, volume 36, pp. 28541–28564, 2023. doi:10.52202/075280-1240.

Haolin Li, Tianjie Dai, Zhe Chen, Siyuan Du, Jiangchao Yao, Ya Zhang, and Yanfeng Wang. RAD: Towards Trustworthy Retrieval-Augmented Multi-modal Clinical Diagnosis. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 123233–123267. Curran Associates, Inc., 2025. doi:10.52202/085713-4109. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/b28ca52a43881fd87196c71145f30092-Paper-Conference.pdf.

Frank I Marcus, William J McKenna, Duane Sherrill, Cristina Basso, Barbara Bauce, David A Bluemke, Hugh Calkins, Domenico Corrado, Moniek GPJ Cox, James P Daubert, et al. Diagnosis of arrhythmogenic right ventricular cardiomyopathy/dysplasia: Proposed modification of the Task Force Criteria. *Circulation*, 121(13):1533–1541, 2010. doi:10.1161/CIRCULATIONAHA.108.840827.

Joseph P. Mathew, Madhav Swaminathan, and Chakib M. Ayoub (eds.). *Clinical Manual and Review of Transesophageal Echocardiography*. McGraw-Hill, 2 edition,

2010. URL https://accessanesthesiology.mhmedical.com/content.aspx?bookid=417&sectionid=40109095. Appendix B: Wall Motion and Coronary Perfusion.

Hussein Mozannar and David Sontag. Consistent estimators for learning to defer to an expert. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 7076–7087. PMLR, 2020.

Sahal Shaji Mullappilly, Mohammed Irfan Kurpath, Sara Pieri, Saeed Yahya Alseiari, Shanavas Cholakkal, Khaled M Aldahmani, Fahad Shahbaz Khan, Rao Muhammad Anwer, Salman Khan, Timothy Baldwin, and Hisham Cholakkal. BiMediX2: Bio-Medical EXpert LMM for Diverse Medical Modalities. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 14051–14071. Association for Computational Linguistics, 2025. doi:10.18653/v1/2025.findings-emnlp.756. URL https://aclanthology.org/2025.findings-emnlp.756/.

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. doi:10.1038/s41467-026-73022-2.

Harsha Nori, Mayank Daswani, Christopher Kelly, Scott Lundberg, Marco Tulio Ribeiro, Marc Wilson, Xiaoxuan Liu, Viknesh Sounderajah, Jonathan Carlson, Matthew P Lungren, et al. Sequential diagnosis with language models. *arXiv preprint arXiv:2506.22405*, 2025.

Steve R Ommen, Seema Mital, Michael A Burke, Sharlene M Day, Anita Deswal, Perry Elliott, Lauren L Evanovich, Judy Hung, José A Joglar, et al. 2020 AHA/ACC guideline for the diagnosis and treatment of patients with hypertrophic cardiomyopathy: A report of the American College of Cardiology/American Heart Association Joint Committee on Clinical Practice Guidelines. *Circulation*, 142(25):e558–e631, 2020. doi:10.1161/CIR.0000000000000937.

Steve R Ommen, Carolyn Y Ho, Irfan M Asif, Seshadri Balaji, Michael A Burke, Sharlene M Day, Joseph A Dearani, Kelly C Epps-Anderson, Lauren Evanovich, Victor A Ferrari, et al. 2024 AHA/ACC/AMSSM/HRS/PACES/SCMR guideline for the management of hypertrophic cardiomyopathy: A report of the American Heart Association/American College of Cardiology Joint Committee on Clinical Practice Guidelines. *Circulation*, 149(23):e1239–e1311, 2024. doi:10.1161/CIR.0000000000001250.

Jiazhen Pan, Che Liu, Junde Wu, Fenglin Liu, Jiayuan Zhu, Hongwei Bran Li, Chen Chen, Cheng Ouyang, and Daniel Rueckert. MedVLM-R1: Incentivizing Medical Reasoning Capability of Vision-Language Models (VLMs) via Reinforcement Learning. In James C. Gee, Daniel C. Alexander, Jaesung Hong, Juan Eugenio Iglesias, Carole H. Sudre, Archana Venkataraman, Polina Golland, Jong Hyo Kim, and Jinah Park (eds.), *Proceedings of Medical Image Computing and Computer Assisted Intervention – MICCAI 2025*, volume 15966 of *Lecture Notes in Computer Science*, pp. 337–347, Cham, September 2025. Springer Nature Switzerland. ISBN 978-3-032-04981-0. doi:10.1007/978-3-032-04981-0_32.

Jeanette Schulz-Menger, David A. Bluemke, Jens Bremerich, et al. Standardized image interpretation and post-processing in cardiovascular magnetic resonance—2020 update. *Journal of Cardiovascular Magnetic Resonance*, 22:19, 2020. doi:10.1186/s12968-020-00610-6.

Yuxuan Sun, Yixuan Si, Chenglu Zhu, Kai Zhang, Zhongyi Shui, Bowen Ding, Tao Lin, and Lin Yang. CPathAgent: An agent-based foundation model for interpretable high-resolution pathology image analysis mimicking pathologists' diagnostic logic. In *Advances in Neural Information Processing Systems*, volume 38, pp. 101673–101731, 2025. doi:10.52202/085713-3404.

Kristian Thygesen, Joseph S. Alpert, Allan S. Jaffe, Bernard R. Chaitman, Jeroen J. Bax, David A. Morrow, and Harvey D. White. Fourth universal definition of myocardial infarction (2018). *Journal of the American College of Cardiology*, 72(18):2231–2264, 2018. doi:10.1016/j.jacc.2018.08.1038.

Martin Ugander, Peter A. Cain, Per Johnsson, John Palmer, and Håkan Arheden. Chronic non-transmural infarction has a delayed recovery of function following revascularization. *BMC Cardiovascular Disorders*, 10:4, 2010. doi:10.1186/1471-2261-10-4.

Ziyue Wang, Junde Wu, Linghan Cai, Chang Han Low, Xihong Yang, Qiaxuan Li, and Yueming Jin. MedAgent-Pro: Towards evidence-based multi-modal medical diagnosis via reasoning agentic workflow. In *International Conference on Learning Representations*, volume 2026, pp. 114360–114389, 2026.

David E. Winchester, David J. Maron, Ron Blankstein, Ian C. Chang, Ajay J. Kirtane, Raymond Y. Kwong, Patricia A. Pellikka, Jordan M. Prutkin, Raymond Russell, and Alexander T. Sandhu. ACC/AHA/ASE/ASNC/ASPC/HFSA/HRS/SCAI/SCCT/SCMR/STS 2023 Multimodality Appropriate Use Criteria for the Detection and Risk Assessment of Chronic Coronary Disease. *Journal of Cardiovascular Magnetic Resonance*, 25(1):58, 2023. doi:10.1186/s12968-023-00958-5. URL https://doi.org/10.1186/s12968-023-00958-5.

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. doi:10.1109/TPAMI.2026.3730310.

An Yang, Anfeng Li, Baosong Yang, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Katja Zeppenfeld, Jacob Tfelt-Hansen, Marta de Riva, et al. 2022 ESC Guidelines for the management of patients with ventricular arrhythmias and the prevention of sudden cardiac death. *European Heart Journal*, 43(40):3997–4126, 2022. doi:10.1093/eurheartj/ehac262.

Tianhong Zhou, Yin Xu, Yingtao Zhu, Chuxi Xiao, Haiyang Bian, Lei Wei, and Xuegong Zhang. DrVD-Bench: Do vision-language models reason like human doctors in medical image diagnosis? In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, volume 38, 2025. doi:10.52202/085713-0420.

APPENDIX CONTENTS

A Baseline Details **16**

B CMR Measurements and Tool Definitions **17**

B.1 Inputs and Measurement Tools 17

B.2 Ventricular Volume, Function, and Mass 17

B.3 Myocardial Thickness and Regional Contraction 18

C Clinical Grounding and Evidence Contracts **19**

C.1 Clinical Knowledge Cards 19

C.2 Diagnostic Guidance and Evidence Contracts 19

C.3 Guideline-Informed Follow-up Recommendations 20

D Implementation Details **21**

E Experimental Protocols **22**

E.1 Metrics and Statistical Analysis 23

F Additional Experimental Results **24**

F.1 Repeated Full-Workflow Evaluation 24

F.2 Diagnostic Endpoint Comparisons 24

F.3 Dynamic Diagnostic Guidance 25

F.4 Verification and Recovery 25

F.5 Evidence-Locked Report Release 26

F.6 Implementation Checks 27

G Scope and Future Work **27**

A BASELINE DETAILS

All baselines use the same balanced 50-case ACDC cohort as TRACE. Image-based methods receive all short-axis ED/ES slices, without masks or Tool-derived measurements. Resampling, normalization, and input packing follow each model's released implementation or adapter.

Direct image classifiers. Three 3D classifiers use a shared 90/10 split of the 100 ACDC training cases. Each has a five-class head; checkpoint selection maximizes validation Matthews correlation coefficient.

We use a 3D adaptation of ResNet-50 trained from random initialization, while the remaining MedicalNet-based classifiers initialize their backbones from MedicalNet pretraining. The three training runs use seeds 0, 1, and 2. ED and ES images are packed as two input channels, with all available short-axis slices participating in resampling before center cropping and padding.

CineMA classifier. We evaluate three released ACDC-finetuned CineMA classification checkpoints on the same 50 cases and average their metrics.

CMR-CLIP. CMR-CLIP is evaluated in a zero-shot image–text matching setting without ACDC task fine-tuning. Each spatial position forms an ED/ES two-frame clip. Slice-level embeddings are normalized, averaged across the stack, and normalized again to obtain one case-level representation. The predicted label is the class description with the highest cosine similarity to this case embedding.

Two-dimensional generative VLMs. We use the released LLaVA-Med-v1.5-Mistral-7B (Li et al., 2023) and HuatuoGPT-Vision-7B (Chen et al., 2024) checkpoints. For VLMs with two-dimensional visual interfaces, every ED and ES short-axis slice is included in the input. Each image is accompanied by its cardiac phase and slice identifier, resulting in 12–42 images per case in the evaluated cohort. The prompt asks the model to select one of the five ACDC diagnostic labels and provide a brief supporting rationale.

Qwen3-VL and LLaVA-Med use model-specific concise diagnostic prompts compatible with their native conversation templates. The remaining medical VLMs use a shared concise medical diagnostic prompt. Each case is evaluated with one generation per run. No Tool interaction, iterative refinement, answer voting, or external evidence is used.

M3D-LaMed. We use the official M3D-LaMed-Phi-3-4B release (Bai et al., 2024) and adapt it to the ED/ES cine input through a dual-volume interface. All ED slices are assembled into one volume and all ES slices into a second volume. Each volume is resampled to $32 \times 256 \times 256$ and independently encoded using the released visual encoder and projector. The ED and ES visual token blocks, together with explicit phase identifiers, are then provided in a single generation request.

Scoring. Diagnostic accuracy uses all 50 cases as the denominator. Macro-F1 averages F1 across the five ACDC phenotypes. A response without a valid class label is counted as incorrect for accuracy and contributes a false negative to its reference class. Labels are parsed from the final generated response without requiring JSON formatting or a standardized rationale template.

BiMediX2 produces one response without a class decision, and MedVLM-R1 produces 11 such responses. AUROC is reported only for methods that provide continuous five-class scores and is computed using macro one-vs-one aggregation (Hand & Till, 2001). Generative VLMs and TRACE do not record comparable continuous class scores and are therefore not assigned AUROC values.

Repetitions comprise three training seeds for 3D classifiers, three CineMA checkpoints, three CMR-CLIP executions, and eight executions per VLM. Metrics are averaged across runs or checkpoints. Each VLM's eight runs give identical case-level labels. TRACE and its measurement-based controls are scored on final actions, counting DEFER and INVALID as incorrect.

B CMR MEASUREMENTS AND TOOL DEFINITIONS

Tools compute requested patient measurements from predicted cardiac masks. Volumes, ejection fractions, and mass follow standard CMR definitions (Schulz-Menger et al., 2020); thickness and contraction use the segmentation-derived geometry and summaries defined below.

B.1 INPUTS AND MEASUREMENT TOOLS

Segmentation backend. We use CineMA (Fu et al., 2026) to obtain short-axis cine segmentations of the LV cavity, RV cavity, and LV myocardium. ED and ES are taken from the ACDC phase metadata. All downstream measurements are deterministic functions of the predicted masks, image spacing, and patient metadata. The Agent receives only the resulting structured measurements, not images or segmentation masks.

Quality control and evidence records. Before using mask-derived measurements, we check image and segmentation quality. Image QC checks finite values, intensity range and variation, and frame count. Segmentation QC checks temporal label presence, ventricular-volume consistency, and physiologic plausibility. Failed or unavailable QC blocks use of the measurements. Both PASS and SUSPECT are admissible when measurement provenance is complete; SUSPECT retains a quality advisory rather than blocking measurements. Provenance and QC status are checked separately. Each Tool returns a typed record containing the value, unit, availability, quality, producing Tool, and provenance. The Agent receives only the results of its requested Tools (Table 8).

Table 8: **CMR measurement Tools.** Inputs and returned measurements for each Tool. ED/ES denote end diastole/end systole; indexed values additionally require height and weight.

Tool	Input	Output
LV/RV volume	ED/ES cavity masks; voxel spacing	EDV, ESV (mL); EDVi, ESVi (mL/m^2)
LV/RV function	EDV and ESV	LVEF, RVEF (%)
LV mass	ED myocardial mask; voxel spacing	LVM (g); LVMi (g/m^2)
LV wall thickness	ED cavity and myocardial masks; in-plane spacing	Maximum and p95 thickness (mm)
Regional contraction	ED/ES masks and spacing; LVEF and LVEDVi for pattern assignment	Shortening, thickening, pattern, burdens, support index

B.2 VENTRICULAR VOLUME, FUNCTION, AND MASS

Mask-derived measurements. For chamber $c \in \{\mathrm{LV}, \mathrm{RV}\}$ and phase $p \in \{\mathrm{ED}, \mathrm{ES}\}$, let $N_{c,p}$ be the number of cavity voxels and $(\Delta x, \Delta y, \Delta z)$ the voxel spacing in mm. We compute chamber volume in mL and ejection fraction in percent as

$$V_{c,p} = \frac{N_{c,p}\Delta x \Delta y \Delta z}{1000}, \qquad \mathrm{EF}_c = 100\frac{\mathrm{EDV}_c - \mathrm{ESV}_c}{\mathrm{EDV}_c}, \tag{5}$$

Here, $\mathrm{EDV}_c = V_{c,\mathrm{ED}}$ and $\mathrm{ESV}_c = V_{c,\mathrm{ES}}$. We compute LV mass in grams from the ED myocardial volume using a myocardial density of 1.05 g/mL:

$$\mathrm{LVM} = 1.05\, V_{\mathrm{myo,ED}}. \tag{6}$$

These measurements follow standard CMR post-processing recommendations (Schulz-Menger et al., 2020).

Body-size indexing. We index volumes and LV mass by body surface area (BSA), calculated using the Du Bois formula (Du Bois & Du Bois, 1916):

$$\mathrm{BSA} = 0.007184\, H^{0.725} W^{0.425}, \qquad X_i = \frac{X}{\mathrm{BSA}}, \tag{7}$$

where H is height in cm, W is weight in kg, BSA is in m^2, and $X \in \{\mathrm{EDV}, \mathrm{ESV}, \mathrm{LVM}\}$. Indexed values are returned when both height and weight are available.

B.3 Myocardial Thickness and Regional Contraction

We derive myocardial thickness and regional contraction measurements from short-axis cine segmentations. Clinical CMR recommendations motivate the measured quantities, while deterministic algorithms aggregate slice- and region-level measurements into evidence descriptors for the Agent. These include thickness estimates, contraction burdens, and regional contraction patterns.

LV wall thickness. Clinical CMR assessment measures maximal wall thickness perpendicular to the LV wall (Schulz-Menger et al., 2020). For automated extraction, we sample the outer myocardial boundary on eligible ED slices and measure the in-plane physical distance from each boundary point to its nearest LV-cavity point. Pooling all distances $\mathcal{D}$ across eligible slices, we report

$$T_{\max} = \max \mathcal{D}, \qquad T_{95} = Q_{0.95}(\mathcal{D}). \tag{8}$$

Here, $Q_{0.95}$ is the empirical 95th percentile. The segmentation-derived thickness proxy T_{95} provides a robust summary that reduces the influence of extreme distances arising from segmentation boundary irregularities.

Radial shortening. SCMR recommends assessing segmental motion through systolic wall thickening (Schulz-Menger et al., 2020). We compute two segmentation-derived measures: radial shortening and wall thickening. For radial shortening, both phases use the ED cavity centroid as the origin of 12 equal angular sectors, with angles measured from the positive image x axis. Radii are Euclidean distances in pixel coordinates, without spacing correction. We retain a slice–sector pair only when it contains cavity pixels in both phases. Let $r^p_{z,s}$ be its 95th-percentile radius. Radial shortening is

$$S_{z,s} = \frac{r^{\mathrm{ED}}_{z,s} - r^{\mathrm{ES}}_{z,s}}{\max(r^{\mathrm{ED}}_{z,s}, 10^{-3})}. \tag{9}$$

For the set $\mathcal{R}$ of retained slice–sector pairs, let $N_r = |\mathcal{R}|$. Using operational shortening thresholds, we compute the abnormal and severe fractions as

$$a_r = \frac{1}{N_r} \sum_{(z,s)\in\mathcal{R}} \mathbf{1}[S_{z,s} < 0.10], \qquad s_r = \frac{1}{N_r} \sum_{(z,s)\in\mathcal{R}} \mathbf{1}[S_{z,s} < 0.05]. \tag{10}$$

Here, $\mathbf{1}[\cdot]$ is the indicator function. Each pair has equal weight; μ_r and σ_r are the mean and population standard deviation of shortening. If no pair remains, the Tool returns unavailable measurements.

Wall thickening. We construct a geometric basal–mid–apical $6/6/4$ partition separately in ED and ES and pair regions by index. In each phase, slices containing both LV cavity and myocardium are ordered by slice index and divided into thirds. Measurement slices require at least 80 pixels in each of these structures and an outer myocardial boundary.

Each slice uses its phase-specific cavity centroid and the angular convention above to define 12 sectors. Sector thickness is the p95 of boundary-to-cavity distances in mm, obtained with a distance transform using the ED in-plane spacing. Adjacent sectors are merged in pairs for the basal and mid groups and in triples for the apical group. Region thickness T^p_k is the median of its contributing slice–sector values. For indices present in both phases, relative wall thickening is

$$W_k = 100 \frac{T^{\mathrm{ES}}_k - T^{\mathrm{ED}}_k}{\max(T^{\mathrm{ED}}_k, 10^{-3})}. \tag{11}$$

For the common region set $\mathcal{K}$, let $N_w = |\mathcal{K}|$. We use $W_k < 30\%$ to identify reduced thickening, consistent with prior cine MRI assessment (Ugander et al., 2010), and $W_k < 10\%$ for severely reduced thickening, following echocardiographic wall-motion grading (Mathew et al., 2010). The corresponding regional fractions are:

$$a_w = \frac{1}{N_w} \sum_{k\in\mathcal{K}} \mathbf{1}[W_k < 30], \qquad s_w = \frac{1}{N_w} \sum_{k\in\mathcal{K}} \mathbf{1}[W_k < 10]. \tag{12}$$

Each region has equal weight. W_k is expressed in percent, and its population standard deviation is in percentage points. This measurement requires at least eight common regions ($N_w \geq 8$); otherwise it is unavailable.

Contraction summary. The Tool combines radial shortening, relative wall thickening, LVEF, and LVEDVi to summarize contraction as NORMAL, FOCAL, or DIFFUSE. These patterns describe regional function and enter diagnostic evidence assessment alongside the quantitative measurements. Let $e = \text{LVEF}$, $v = \text{LVEDVi}$, $B_r = [a_r \geq 0.50]$, $B_w = [a_w \geq 0.55]$, and $P = [\mu_r < 0.18 \wedge \sigma_r < 0.15]$. The global-motion context is

$$G = [e < 40] \wedge \begin{cases} [v \geq 100] \wedge (B_r \vee B_w \vee P), & v \text{ available}, \\ (B_r \vee P) \wedge (B_w \vee [a_r \geq 0.60]), & v \text{ unavailable}. \end{cases} \tag{13}$$

Table 9 summarizes pattern assignment. Radial shortening is required; available wall-thickening findings refine the initial assessment. Comparisons involving unavailable values are false. Diagnostic grounding uses the larger burden from the two branches: $a = \max(a_r, a_w)$ and $s = \max(s_r, s_w)$. Unavailable wall terms contribute zero, with measurement availability retained in the evidence record.

Table 9: Ordered rules for the segmentation-derived contraction summary. The wall-thickening update follows radial initialization; the final row sets the output priority.

Stage	Rule
Initial pattern criteria	Focal flag $F = [e < 50] \wedge [\sigma_r \geq 0.20] \wedge [0.10 \leq a_r \leq 0.55] \wedge \neg G$; diffuse flag $D = [e < 45] \wedge (G \vee [a_r > 0.75])$.
Wall thickening	No clear reduction if $a_w < 0.10$; otherwise diffuse reduction if $a_w \geq 0.55$, $s_w \geq 0.25$, and $\text{SD}(W) < 25$; otherwise regional reduction.
Pattern refinement	Diffuse reduction adds diffuse support (D). Regional reduction adds diffuse support and removes focal support (F) when $G \wedge B_w$; otherwise it adds focal support. No clear reduction removes focal support. All other initial support is retained.
Final pattern	DIFFUSE if D; otherwise FOCAL if F; otherwise NORMAL.

Evidence-support index. A bounded, rule-based index summarizes support for the selected contraction pattern and is used in the phenotype-specific acceptance criteria in Table 11. The index combines shortening statistics, wall-thickening burdens, and ventricular dilatation, with an upper bound of 0.99 for focal and diffuse patterns and 0.95 for the normal pattern. Unavailable wall terms contribute zero, and the returned value is rounded to three decimals.

C CLINICAL GROUNDING AND EVIDENCE CONTRACTS

Clinical grounding links patient measurements to source-backed diagnostic requirements; separate cards support follow-up recommendations after acceptance.

C.1 CLINICAL KNOWLEDGE CARDS

We organize clinical knowledge into six diagnostic cards and ten follow-up recommendation cards (Table 10). Each diagnostic card summarizes relevant clinical points, links them to measurement Tools, and records its sources. Recommendation cards provide source-linked suggestions for further assessment after diagnosis. We store both sets as structured records and load them before inference. The cards supply predefined clinical content; inference updates the patient-specific interpretation, not the cards.

C.2 DIAGNOSTIC GUIDANCE AND EVIDENCE CONTRACTS

Diagnostic criteria. Clinical sources define the phenotype features, which we translate into measurable support conditions in Table 11. For HCM, the AHA/ACC guideline uses maximal end-diastolic wall thickness $\geq$ 15 mm (Ommen et al., 2020); TRACE applies this cutoff to the T_{95} proxy in Section B. For DCM, ESC guidance identifies LV dilatation and systolic dysfunction (Arbelo et al., 2023); TRACE uses LVEDVi ≥ 100 mL/m^2 and LVEF $< 45\%$ for this task. For RV, TRACE applies a uniform operational rule: RVEDVi > 110 mL/m^2 or RVEF $< 40\%$, following the phenotype definition on the official ACDC dataset webpage (ACDC Challenge, n.d.).

Table 10: Clinical sources for diagnostic and follow-up cards, grouped by phenotype. Deferral combines clinical context with the TRACE workflow policy.

Card scope	Diagnostic basis and source	Follow-up and diagnostic context
HCM	Wall-thickness phenotype: AHA/ACC HCM guideline (Ommen et al., 2020)	HCM guideline (Ommen et al., 2024); SCMR indications (Leiner et al., 2020)
DCM	Dilatation and systolic dysfunction: ESC cardiomyopathy guideline (Arbelo et al., 2023)	ESC cardiomyopathy guideline (Arbelo et al., 2023); SCMR indications (Leiner et al., 2020)
MINF	Regional dysfunction as imaging evidence: Universal Definition of MI (Thygesen et al., 2018)	MI definition (Thygesen et al., 2018); chest-pain pathway (Kontos et al., 2022); imaging appropriateness criteria (Winchester et al., 2023); SCMR indications (Leiner et al., 2020)
RV	Abnormal-RV phenotype: ACDC category definition (Bernard et al., 2018)	ACDC (Bernard et al., 2018); ESC guidance (Arbelo et al., 2023; Zeppenfeld et al., 2022); ARVC criteria (Marcus et al., 2010); SCMR reporting (Hundley et al., 2022)
NOR	Integrated assessment context: ESC cardiomyopathy guideline (Arbelo et al., 2023); evidence completion: TRACE	SCMR reporting and indications (Hundley et al., 2022; Leiner et al., 2020)
DEFER	ESC clinical context (Arbelo et al., 2023); TRACE deferral policy	TRACE deferral policy; SCMR reporting (Hundley et al., 2022); ACDC context for RV-related review (Bernard et al., 2018)

ACDC describes the MINF phenotype in terms of impaired LV function and regional contraction abnormalities (Bernard et al., 2018). To quantify regional contraction, we draw on SCMR's assessment of systolic wall thickening (Schulz-Menger et al., 2020); the Universal Definition of Myocardial Infarction provides the clinical context for regional wall-motion abnormalities (Thygesen et al., 2018). TRACE combines the focal-contraction summary with LVEF $< 50\%$, the ESC threshold for LV systolic dysfunction (Arbelo et al., 2023). The resulting Contract classifies the cine phenotype; clinical confirmation of infarction integrates additional clinical evidence. NOR and deferral use TRACE's evidence-completion requirements.

Guidance and acceptance. Each guidance call returns the five fixed phenotype cards with interpretations updated from the current evidence; the differential does not rank or filter the cards. The deferral card is added when unresolved deferral reasons are present. The verifier applies the support and conflict conditions in Table 11 to quality-admissible evidence. Missing required support blocks the proposed label; an unobserved competing predicate does not count as a positive conflict. Consequently, DCM and RV may be accepted without regional-motion evidence when their support conditions hold and no observed competitor blocks them. Rejected proposals return unmet requirements and available resolving actions, as illustrated in Figure 3.

Deferral covers two unresolved states: a named evidence gap that no available action can resolve, or diagnostic exhaustion, where all five labels have been rejected without new evidence. The latter means that the current evidence supports no admissible candidate, rather than that a measurement is missing. A single rejected proposal is insufficient for deferral.

C.3 Guideline-Informed Follow-up Recommendations

After acceptance, we lock the decision and supporting evidence and select follow-up recommendations from the ten recommendation cards. Each card specifies applicable diagnoses, triggering conditions, recommendation text, priority, and clinical sources. We match cards to the accepted diagnosis, LGE availability, and remaining evidence gaps, then select up to two by priority and a fixed card order. The report includes the selected text and references with their clinical conditions intact. For example, the NOR card recommends further assessment if symptoms or other findings are discordant; the clinician evaluates this condition, which is not inferred from cine data.

Table 11: Implemented Evidence Contracts for ACDC. Conditions are conjunctive unless marked "or"; all diagnoses require usable QC. Sources and operational thresholds are described in the text.

Proposal	Required support	Additional conditions
DCM	LVEF $< 45\%$; LVEDVi ≥ 100 mL/m^2.	Diffuse dysfunction is supportive, not required. The MINF predicate is a direct conflict; the HCM predicate is an unresolved competitor.
HCM	ED wall thickness $T_{95} \geq 15$ mm.	LV mass index ≥ 90 g/m^2 is supportive only. $T_{\max} \geq 15$ mm with $T_{95} < 15$ mm is borderline. Positive DCM or MINF predicates prevent acceptance.
MINF	Focal contraction pattern; support index ≥ 0.65; abnormal burden ≥ 0.15; severe burden ≥ 0.10; LVEF $< 50\%$.	Diffuse or normal contraction patterns conflict with acceptance.
RV	RVEDVi > 110 mL/m^2 or RVEF $< 40\%$.	The volume branch requires valid indexing provenance. Equality does not pass. Positive DCM and MINF predicates are unresolved competitors.
NOR	LVEF $\geq 45\%$; RVEF $\geq 40\%$; RVEDVi ≤ 110 mL/m^2 with valid indexing; below-threshold wall thickness; assessed motion without the positive MINF predicate.	Wall thickness must be available with neither T_{95} nor $T_{\max}$ reaching 15 mm. Borderline focal motion is permitted by the operational rule. Missing evidence or a positive HCM, DCM, MINF, or RV predicate prevents acceptance.
DEFER	A named unresolved gap with no permitted resolving action, or diagnostic exhaustion after all five labels are rejected without new evidence.	Generic uncertainty or rejection of one label is insufficient. A gap resolvable by an available Tool requires further acquisition.

D IMPLEMENTATION DETAILS

Agent model and serving. The diagnostic Agent uses Qwen3-8B (Yang et al., 2025), served with vLLM 0.19.0 (Kwon et al., 2023) through an OpenAI-compatible API. Initial planning uses temperature 0.6, while diagnostic reconsideration and terminal submission use temperature 0. The maximum context length is 40,960 tokens with a 1,024-token safety margin, and each model response is limited to 2,048 tokens. Parallel Tool calls are disabled, and only the first public Tool call in each model completion is executed.

Each case allows up to 15 clinical model completions, with an additional four completions reserved for protocol repair. Diagnostic reconsideration and terminal submission are separately bounded within each evidence state. No additional training or fine-tuning of Qwen3-8B is performed.

Segmentation and measurement preprocessing. Cine CMR images are center-cropped before being passed to CineMA. Predicted segmentations are cached and reused by the deterministic measurement Tools. TRACE and its measurement-based controls share these masks to hold the segmentation backend fixed.

Action and write permissions. The runtime exposes measurement, guidance, and terminal actions according to the current state. The Agent chooses among them and owns the differential and submitted decision. Evidence changes or rejection require refreshed guidance before release. Table 12 summarizes the component permissions; recommendation selection follows Section C.

Report construction. The compiler uses the accepted action and its associated Tool evidence to assemble measurements, qualitative findings, provenance, diagnostic rationale, and unresolved gaps. The post-lock selector adds at most two source-linked recommendation cards without model calls. These recommendations cover further imaging, clinical workup, manual review, or clinical correlation.

Table 12: Component responsibilities and authority boundaries in TRACE.

Component	Controls	Cannot
Agent	Differential, Tool selection, guidance requests, terminal proposals.	Write measurements, approve proposals, or alter locked facts.
Measurement Tools	Measurements, contraction summaries, QC, provenance.	Select diagnoses or change acceptance rules.
Clinical grounding	Source-linked interpretation, requirements, gaps, resolving actions.	Create or modify measurements, or select diagnoses.
Verifier and compiler	Proposal acceptance, evidence locking, factual report compilation.	Substitute diagnoses or create patient evidence.
Post-lock selector	Selection and ordering of follow-up cards.	Reopen diagnoses, modify evidence, or recommend treatment.

E EXPERIMENTAL PROTOCOLS

Overview. All experiments use the balanced 50-case ACDC cohort. Table 13 maps each condition to its evaluation question.

Table 13: Experimental conditions and the questions they address.

ID	Condition	Scale	Evaluation focus
E0	Full TRACE workflow	8×50	Full-workflow performance across runs.
E1	Full workflow vs. single-pass reasoning	paired 3×50	Adaptive workflow vs. single-pass reasoning with complete evidence.
E2	Deterministic evidence-to-label reference	50 offline	Classification from rules and Contracts without an Agent.
E3	Selective evidence acquisition	paired 3×50	Adaptive vs. complete optional evidence exposure.
E4	Dynamic diagnostic guidance	paired 3×50	Evidence-updated vs. static guidance.
E5	Verification and recovery	3×50 audit; paired $2 \times 3 \times 50$ intervention	Recovery outcomes and structured-feedback effects.
E6	Evidence-locked recommendation release	3×50 locked outputs	Recommendation conformance: selection vs. generation.
INV	Authority and lock enforcement	10 synthetic checks	Component permissions and post-lock restrictions.

E0: full-workflow evaluation. E0 runs the full workflow eight times per case, measuring accuracy, coverage, invalid termination, and run-to-run consistency.

E1: full workflow vs. single-pass reasoning. E1 compares the full workflow with the same Qwen3-8B model performing a single inference pass over complete cached measurements. Along with the full measurement set, the single-pass condition receives diagnostic guidance including phenotype thresholds and evidence requirements. It does not perform adaptive Tool acquisition, dynamic guidance updates, proposal verification, or recovery.

E2: deterministic evidence-to-label reference. E2 removes the language-model control loop entirely. The same deterministic interpretation rules used by Clinical Evidence Grounding are first applied to the complete measurement set, after which every phenotype Evidence Contract is evaluated. If exactly one Contract closes, its phenotype is emitted. If none or more than one Contract closes, the reference abstains. We count abstentions as DEFER for scoring, although this reference does not identify an unresolved evidence gap. The inferential comparison with TRACE uses each patient's eight E0 runs; the comparison with single-pass reasoning uses its three E1 runs. The descriptive three-run comparison in Table 2 is reported separately.

E3: selective evidence acquisition. E3 holds the segmentation cache and measurement backend fixed. The selective condition acquires wall thickness and regional contraction at the Agent's request; complete exposure supplies both before terminal reasoning.

E4: dynamic diagnostic guidance. E4 compares guidance recomputed from the current patient evidence with static diagnostic context. Both conditions use the same clinical source material, measurement backend, model, cases, and action space. The dynamic condition updates satisfied requirements, contradictions, and unresolved gaps as evidence changes; the static condition does not update these interpretations from the evolving case state.

E5: recovery after verification. E5 contains two complementary analyses. First, a trajectory audit follows the first submitted diagnostic proposal and the final admissible action across three runs of each case. This analysis measures how often rejected proposals are followed by diagnosis revision, additional evidence acquisition, or deferral. Its outcome counts are reported in Tables 18 and 19; E1 endpoint counts are reported separately in Table 2.

Second, a paired intervention compares structured rejection feedback with a generic rejection message. Structured feedback identifies unmet evidence requirements and available resolving actions, whereas generic feedback withholds these details. The two arms are executed separately for each case and seed, with the same Agent model, verification rules, public action schema, and budget limits. The arms do not branch from a shared state immediately before feedback. Both retain the runtime's state-dependent action restrictions; the contrast changes the model-visible rejection message. Intervention outcomes are reported separately in Table 20, not pooled with the recovery audit.

E6: recommendation release from locked case records. E6 evaluates the recommendation component of report release using 50 locked records from E0 repeat 1 (sampling seed 1009). The primary comparison gives the deterministic selector and Qwen3-8B the same case-applicable recommendation cards. A third, source-pack-only condition gives the LLM the same locked state and reference registry, but withholds the typed cards. This is an information-exposure stress condition. The registry supplies reference IDs, citations, titles, authorities, years, support scopes, and limitations.

Each LLM condition uses one JSON-schema-constrained request per case, temperature 0.6, seed 1009, and a 2,048-token output limit, with no Tool calls or retries. The request does not override the server's thinking-template setting; a complete leading thinking block is removed before JSON parsing. The output contains at most two recommendations, each with text, basis, scope, source IDs, and a card ID or null. The diagnostic factual core remains locked throughout this comparison.

The same automatic evaluator applies the ten checks in Table 14 to all conditions. Source applicability is defined by the references attached to cards eligible for the locked state; alignment and label checks use predefined text rules. Shared conformance is the conjunction of all ten checks. Card identity and exact-copy checks are separate. Thus E6 measures adherence to this operational recommendation specification.

Length-limited or otherwise invalid generations enter evaluation with an empty recommendation list. Empty outputs fail completion and content checks, even if they contain no prohibited labels or treatment terms, and therefore fail shared conformance.

Synthetic authority and lock tests. Two positive controls and eight negative tests check software enforcement. Negative tests cover measurement writes by grounding, changes to locked values, diagnosis bindings, labels or evidence, and missing-cine, failed-QC or backend-failure paths. Expected outcomes are predefined; these tests are separate from patient-level diagnosis evaluation.

E.1 Metrics and Statistical Analysis

Diagnostic metrics. Diagnostic accuracy uses all evaluated cases as the denominator; DEFER and INVALID are counted as incorrect. Diagnosis coverage is the fraction of trajectories ending in a concrete diagnosis. Accuracy among concrete diagnoses uses only those diagnosed trajectories as its denominator. Macro-F1 averages F1 across the five ACDC phenotypes.

Table 14: Operational checks shared by all E6 conditions. Source and text checks use the predefined registry, card applicability rules, and term lists.

Check	Passing condition
Completion	Response completes and validates under the output schema.
Source IDs	Every recommendation has nonempty source IDs, all in the registry.
Applicable sources	Cited IDs belong to the union of sources of case-applicable cards, and each recommendation has a nonempty basis.
Category alignment	Every recommendation contains a term from the locked phenotype's term list; DEFER uses evidence-completion and quality terms.
Locked state and label	Locked input identity is preserved and recommendation text contains no canonical diagnostic label other than the locked label.
Treatment exclusion	No predefined treatment/intervention term is detected.
Allowed scope	Each scope is one of the four permitted categories, and no treatment term is detected.
Count limit	At most two recommendations.
No duplicates	No repeated pair of normalized recommendation text and source-ID set.
Nonempty content	At least one recommendation, with nonempty text in each entry.

Process metrics. Workflow analyses additionally record first-proposal acceptance, proposal rejections, optional evidence actions, guidance calls, public Tool actions, measurement-backend executions, model requests, valid terminal actions, and recovery outcomes. Report-release analyses measure conformance with the shared release specification and its component checks.

Statistical unit and intervals. The patient is the independent unit. For complete repeated-run comparisons, we average paired differences within each patient, then average over the 50 patients. We compute percentile bootstrap intervals from 10,000 resampled datasets. Each dataset draws 50 patient clusters with replacement from the original 50 patients and retains each selected patient's repeated runs (seed 20260811 for the original suite; 20260819 for the feedback intervention). For the 43-pair feedback subset, 24 patients contribute eligible pairs. The estimate is the difference summed over eligible pairs divided by their count. Its bootstrap samples the original 50 patient clusters, retaining each cluster's eligible pairs and recomputing that ratio; draws with no eligible pair are omitted.

Paired tests and multiplicity. E1–E5 use exact two-sided sign-flip tests on patient-level summed paired differences, enumerating both signs for every nonzero patient difference. Zero-difference patients do not affect the randomization distribution. These discrete tests and percentile bootstrap intervals use different constructions and need not give identical threshold decisions. The global Holm family (Holm, 1979) covers the five primary contrasts: E1 accuracy, E2 accuracy, E3 accuracy, E4 accuracy, and E5 first-to-final accuracy. Secondary outcomes are adjusted within each experiment. The feedback intervention uses its 43-pair subset for primary correctness and all 150 pairs for operational outcomes, with the latter forming a separate Holm family.

F Additional Experimental Results

Additional results follow the protocols and statistical procedures in Sections E and E.1.

F.1 Repeated Full-Workflow Evaluation

Table 15 reports the complete E0 results across eight runs of the 50-case cohort. TRACE produces no invalid trajectories across the 400 executions. The patient-level 95% bootstrap confidence interval for diagnostic accuracy is $[77.8, 95.5]\%$. Of the 50 patients, 45 are correctly diagnosed in at least one of eight runs and 41 in all eight runs.

F.2 Diagnostic Endpoint Comparisons

The full workflow improves accuracy by 13.3 percentage points over single-pass reasoning with complete measurements (Table 16).

Table 15: E0 results across eight repeated runs of the 50-case cohort.

Metric	Count	Value
Accuracy	349/400	87.3%
Diagnosis coverage	385/400	96.3%
Accuracy among concrete diagnoses	349/385	90.6%
Deferral	15/400	3.8%
Invalid termination	0/400	0.0%
Correct in at least one of eight runs	45/50	90.0%
Correct in all eight runs	41/50	82.0%

Table 16: Endpoint contrasts, in the order shown, with runs clustered by patient. E2 uses eight E0 runs for TRACE and three E1 runs for single-pass reasoning.

Contrast	Effect	95% CI	Exact p	Holm q
E1: Full workflow − single pass	$+13.3$ pp	$[4.0, 24.0]$	0.0137	0.0547 (global)
E2: Full workflow − deterministic reference	-0.8 pp	$[-6.3, 5.0]$	0.7500	1.0000 (global)
E2: Deterministic reference − single pass	$+14.7$ pp	$[4.0, 26.7]$	0.0215	0.0430 (within E2)

F.3 Dynamic Diagnostic Guidance

Updating guidance increases first-proposal acceptance and reduces rejections and calls; the accuracy difference is -2.0 percentage points (Table 17).

Table 17: E4 paired inferential statistics. Effects are dynamic guidance minus static guidance. Holm q is within E4 unless marked global.

Outcome	Effect	95% CI	Exact p	Holm q
Accuracy	-2.0 pp	$[-7.3, 2.7]$	0.625	1.000 (global)
First-proposal acceptance	$+24.0$ pp	$[15.3, 33.3]$	3.10×10^{-6}	3.10×10^{-6}
Proposal rejections	-0.95	$[-1.23, -0.68]$	1.83×10^{-9}	5.50×10^{-9}
Guidance calls	-1.39	$[-1.79, -1.00]$	4.70×10^{-9}	9.40×10^{-9}
Model requests	-2.91	$[-3.73, -2.12]$	1.12×10^{-9}	4.50×10^{-9}

F.4 Verification and Recovery

Recovery audit. Table 18 follows first proposals to final actions. All 84 rejected trajectories recover: 47 change an incorrect diagnosis to the correct label, 29 retain a correct diagnosis and complete its evidence, and eight defer.

Table 18: Recovery across 50 patients and three seeds.

Measure	Result
Accuracy if the first proposal were released	83/150 (55.3%)
Accuracy of the final admissible action	130/150 (86.7%)
Paired accuracy difference	$+31.3$ pp; 95% CI $[20.0, 43.3]$ pp
Exact patient-level p / global Holm q	9.54×10^{-7} / 4.77×10^{-6}
Trajectories with rejection	84/150
Patients with rejection	32/50
Patients rejected under 1 / 2 / 3 seeds	4 / 4 / 24
Admissible action after rejection	84/84
Correct→correct / incorrect→correct	29 / 47
Correct→incorrect / incorrect→DEFER	0 / 8

The phenotype decomposition is shown in Table 19. Rejected trajectories occur most frequently for NOR, HCM, and MINF in this cohort; no RV trajectory receives a rejected proposal.

Table 19: Recovery by reference phenotype. Each patient contributes up to three trajectories with rejection.

Phenotype	Patients	Trajectories	Final correct	Final DEFER
DCM	5	10	10	0
HCM	10	26	23	3
MINF	7	18	13	5
NOR	10	30	30	0
RV	0	0	0	0
Total	32	84	76	8

Across the rejected trajectories, the 123 rejected proposals contain nonexclusive reasons including unsupported predicates (106), unresolved contradictions (47), and unresolved competing phenotypes (15). Sixty-one of the 84 trajectories receive one rejection, 17 receive two, four receive three, and two receive eight. Seventy-two trajectories are offered an action that can acquire missing evidence. All 84 subsequently refresh diagnostic guidance, and 72 acquire regional-contraction evidence; five of these also acquire wall-thickness evidence.

Across all 150 trajectories, the final actions comprise 142 concrete diagnoses, eight deferrals, and no invalid outputs. The 12 incorrect trajectories ending in a concrete diagnosis arise from five MINF patients: nine end as DCM and three as RV. All 12 of these incorrect first proposals are accepted without a preceding rejection. These cases therefore represent accepted diagnostic errors rather than failed recovery after verification.

Structured verification feedback. The feedback intervention reports its own complete executions, separately from the recovery audit above. Structured and generic feedback give similar correctness on the 43 primary pairs; structured feedback uses fewer rejections and calls across all 150 trajectory pairs (Table 20).

Table 20: Structured minus generic feedback. The primary comparison uses 43 pairs from 24 patients with matched first rejected proposals and verification outcomes; operational measures cover all 150 complete trajectory pairs.

Outcome	Structured	Generic	Effect	95% CI	Holm q
Primary final accuracy	38/43	39/43	-2.33 pp	$[-7.89, 0]$	1.000
Accuracy across all pairs	130/150	131/150	-0.67 pp	$[-2.67, 1.33]$	1.000
Valid action rate	149/150	148/150	$+0.67$ pp	$[0, 2.00]$	1.000
Diagnosis coverage	141/150	142/150	-0.67 pp	$[-2.67, 1.33]$	1.000
Invalid rate	1/150	2/150	-0.67 pp	$[-2.00, 0]$	1.000
Rejected proposals	0.827	1.100	-0.273	$[-0.433, -0.133]$	0.0037
Optional evidence actions	1.213	1.247	-0.033	$[-0.087, 0.013]$	1.000
Guidance calls	2.793	3.093	-0.300	$[-0.467, -0.147]$	0.0052
Public Tool actions	10.993	11.327	-0.333	$[-0.547, -0.133]$	0.0160
Backend executions	6.787	7.087	-0.300	$[-0.480, -0.133]$	0.0079
Model requests	13.027	13.660	-0.633	$[-1.033, -0.280]$	0.0135

The exact paired patient-level p value for the primary contrast is 1.0. Forty-two primary pairs tie; the only discordant pair favors generic feedback. Resource outcomes count calls and backend executions over the entire trajectory, including the pre-feedback portion.

F.5 EVIDENCE-LOCKED REPORT RELEASE

Report integrity in the full workflow. All 400 reports preserve the accepted decision, use current evidence, ground numerical claims in Tool records, and retain complete provenance (Table 21).

Shared release conformance. Shared conformance is 50/50 for deterministic selection, 23/50 for the LLM given the same records and cards, and 0/50 for the source-pack-only LLM (Table 22).

Table 21: E0 audit of report integrity across the 400 full-workflow trajectories.

Release invariant	Result
Submitted action equals released action	400/400
Current evidence state used at release	400/400
Numerical claims grounded in Tool records	400/400
Compiled factual core with complete provenance	400/400
Controller label substitution	0/400

Table 22: E6 recommendation-release checks after locking. The source-pack-only arm withholds typed cards. Shared conformance requires all ten checks in Table 14; card identity and exact copying are excluded.

Check	Deterministic	LLM with cards	Source pack only
Completed generation	50/50	42/50	50/50
Shared conformance	50/50	23/50	0/50
Conformance among completed outputs	50/50	23/42	0/50
Applicable source scope	50/50	31/50	0/50
Category alignment	50/50	29/50	20/50
Locked state and label check	50/50	50/50	13/50
No treatment leakage	50/50	50/50	50/50
No duplicate recommendations	50/50	44/50	50/50
Valid source IDs	50/50	42/50	49/50
Allowed scope	50/50	50/50	50/50
Within count limit	50/50	50/50	50/50
Nonempty content	50/50	42/50	50/50

The LLM-with-cards condition produces eight outputs that exceed the output length limit. Among completed outputs, 19 fail at least one shared release requirement. Its nonexclusive check failures include category misalignment in 21/50 cases, invalid source scope in 19/50, and duplicate recommendations in 6/50. The source-pack-only LLM completes all 50 generations but fails the applicable-source check in 50/50 cases, the locked-state/label check in 37/50, and category alignment in 30/50. These are failures of the specified registry and text checks, not independently adjudicated clinical errors.

F.6 Implementation Checks

TRACE passes all ten synthetic checks (Table 23): two valid controls, five unauthorized writes or post-lock modifications, and three missing-cine, failed-QC, or backend-failure paths. The 400 E0 trajectories also preserve component responsibilities (Table 12): Tools supply measurements; grounding supplies diagnostic interpretation and deferral-policy information.

Table 23: Synthetic authority and lock tests. Matched denotes tests producing the expected outcome. Errors are false rejections for controls and false acceptances otherwise.

Test group	**Matched**	**Errors**
Valid controls	2/2	0
Write / lock violations	5/5	0
Input / QC / backend failures	3/3	0
All checks	**10/10**	**0**

G Scope and Future Work

This study examines evidence acquisition, diagnostic verification, and structured reporting for five cine phenotypes using CineMA segmentations. The 50-case ACDC cohort supported iterative method development and evaluation, with observed outcomes informing method refinement. The results characterize diagnostic performance and controlled workflow comparisons within this development cohort.

TRACE motivates three directions for further research. First, evidence acquisition and stopping can be studied jointly: which additional measurement best distinguishes competing diagnoses, and when should the system conclude, acquire more evidence, or defer? Policies that account for diagnostic uncertainty and acquisition cost could be evaluated through their trade-offs among diagnostic error, coverage, and resource use.

Second, extending TRACE to additional CMR sequences and clinical information would enable investigation of multimodal diagnostic workflows. A central question is how evidence requirements and verification should handle missing or conflicting findings across modalities. Evaluation on independent, multi-center cohorts, with measurement rules and Contracts specified before testing, would assess transfer across acquisition settings and segmentation methods. Clinician-reviewed reference measurements would support validation of the geometric proxies and acceptance criteria.

Third, clinician reader studies could examine whether structured evidence and follow-up recommendations improve diagnostic accuracy, review time, and the appropriateness of further assessment. The present report evaluation measures factual preservation and operational conformance; these studies would establish whether those properties translate into clinical decision support.