Secure LLM-Assisted Labeling and Spatiotemporal CMR Representation for Sequence and View Recognition

Yixuan Liu YIXUAN.LIU@OSUMC.EDU
Zhenyu Bu ZHENYU.BU@OSUMC.EDU
Yi Yu YI.YU@OSUMC.EDU
Parker Martin PARKER.MARTIN@OSUMC.EDU
Yuchi Han YUCHI.HAN@OSUMC.EDU
Orlando Simonetti ORLANDO.SIMONETTI@OSUMC.EDU
Yuan Xue YUAN.XUE@OSUMC.EDU
The Ohio State University, Columbus, OH, USA

Editors: Under Review for MIDL 2026

Abstract

Cardiovascular magnetic resonance (CMR) studies combine diverse pulse sequences and imaging planes, which is clinically valuable but makes large scale data curation and automated analysis difficult. In routine practice, series descriptions in DICOM headers are heterogeneous across technologists, scanners, vendors, and time, so manual sequence and view labeling does not scale beyond small cohorts. We develop a secure labeling pipeline that uses a domain knowledge guided prompt for large language models (LLMs) with explicit CMR protocol based mapping rules to drive a locally deployed GPT-OSS model. From raw series descriptions, our prompt generates standardized pseudo labels for sequence type and cardiac view for approximately 76,000 CMR series from 1,000 patients entirely offline, preserving data security while capturing local naming conventions. These labels are used to train a spatiotemporal CMR encoder that combines a ConvNeXt image backbone with an xLSTM temporal module and maps heterogeneous series into a compact low dimensional embedding for multi-class sequence and view classification. On an expert annotated test set, the domain knowledge guided prompt reduces the number of unknown labels by two orders of magnitude and improves sequence and view label accuracy compared with a generic prompt. Models trained on these optimized pseudo labels achieve sequence and view classification accuracy of 0.983 and 0.989 respectively, outperforming existing 2D and Vision Transformer baselines. The proposed framework shows that clinically informed prompting and explicit spatiotemporal modeling together enable secure CMR curation and accurate sequence and view recognition at scale.

Keywords: cardiovascular magnetic resonance, sequence classification, view classification, spatiotemporal representation learning, large language models

1. Introduction

Cardiovascular magnetic resonance imaging (CMR) is a comprehensive non-invasive imaging technique that is central to the diagnosis and management of cardiovascular disease (Karamitsos et al., 2009). CMR is widely regarded as the reference modality for quantifying ventricular function, chamber volumes, and myocardial scar, and it provides rich tissue characterization for myocardial edema, iron overload, perfusion abnormalities, and diffuse fibrosis (Salerno et al., 2017b). Unlike brain or body MRI protocols that are often built

around a small number of relatively standardized three dimensional acquisitions, clinical CMR protocols must accommodate a rapidly moving organ with complex anatomy and physiology (Salerno et al., 2017a).

The heart is a dynamic structure composed of obliquely oriented chambers that move with both the cardiac cycle and respiration (Bogaert et al., 2012). As a result, despite recent progress in 3D whole-heart imaging, the clinical workhorse of CMR remains two dimensional multi-slice acquisitions that are frequently electrocardiogram-gated and breath-held (Lima and Desai, 2004). To visualize cardiac anatomy and function, images are acquired in a set of standardized long axis and short axis views rather than in orthogonal axial, coronal, and sagittal planes (Figure 1). These views are further combined with multiple sequence types, such as balanced steady state free precession cine for wall motion and function, phase contrast flow imaging, inversion recovery late gadolinium enhancement for scar, and quantitative T1, T2, and T2* mapping and fat water imaging for diffuse tissue characterization (Kramer et al., 2020). This diversity of sequences and views is clinically valuable but introduces substantial complexity for large scale data curation and automated analysis (Salerno et al., 2017a). In routine practice, technologists start from vendor specific protocol trees and then edit sequence names to reflect the imaging plane, heart rate adjustments, breath hold strategy, and minor parameter changes (Lim et al., 2022). Over time, local conventions evolve across scanners, software versions, and personnel. As a consequence, the DICOM SeriesDescription field is highly heterogeneous, and the same acquisition type may appear under many different textual variants. Manual labeling works for small research cohorts but does not scale to tens of thousands of series or multi-center datasets.

Robust sequence and view classification is a critical enabling step for downstream CMR applications. Reliable labels allow automatic construction of analysis pipelines, for example selecting the correct cine series for ventricular function analysis, the appropriate late gadolinium enhancement images for scar quantification, or specific mapping sequences for quantitative tissue characterization (Flett et al., 2011). They also support quality control, protocol harmonization across scanners and sites, and retrospective cohort assembly for disease specific studies (Kramer et al., 2020). Existing work on MRI sequence identification has demonstrated that convolutional and transformer based models can classify sequence types directly from the images even when metadata are unreliable (Ranjbar et al., 2020; de Mello et al., 2021; Helm et al., 2024; Lim et al., 2022; Mahmutoglu et al., 2025; Wang et al., 2025). However, most prior studies focus on a limited set of brain or body sequences, where multi-slice information is either ignored or treated as independent channels rather than being modeled explicitly and rely on manually curated labels or rigid metadata rules that are difficult to maintain in real world clinical environments.

Large language models (LLMs) provide a natural way to interpret textual metadata such as DICOM series descriptions (Kamel et al., 2025). Off the shelf LLMs are not trained on cardiac specific terminology, local abbreviations or site-specific naming practices, and sending protected health information to external interfaces raises regulatory and security concerns. There is therefore a need for labeling strategies that pair clinical domain knowledge with LLMs while keeping all computation inside the institutional firewall.

In this work, we address these challenges with a unified framework for secure labeling and spatiotemporal representation learning for CMR. **First**, we design a domain knowledge guided prompt for a locally deployed GPT-OSS model (Agarwal et al., 2025) that

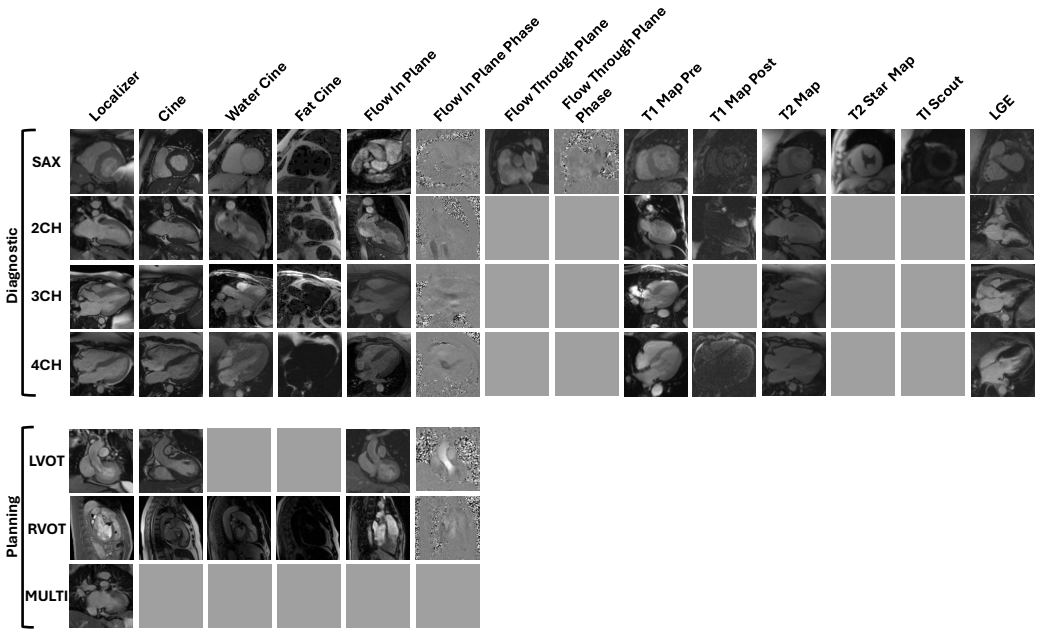

Figure 1: Comprehensive illustration of CMR sequences and imaging planes, showing all combinations included in this study. This variability motivates the need for robust sequence and view classification methods.

generates pseudo labels for both sequence type and imaging plane directly from raw series descriptions. This approach leverages expert knowledge of CMR protocols and site-specific naming patterns, and it produces high fidelity labels for approximately 76,000 series from 1,000 patients without any external data transfer. **Second**, we propose a spatiotemporal CMR encoder that combines a ConvNeXt (Liu et al., 2022) based spatial feature extractor with an xLSTM (Beck et al., 2024a) temporal module. The encoder operates directly on heterogeneous CMR inputs, including cine series, inversion recovery and mapping sequences, and multi slice localizers, and compresses them into a compact one dimensional embedding that captures sequence identity, view, and acquisition specific imaging physics. **Third**, we demonstrate that this embedding supports accurate multi-class sequence and view classification and that models trained on domain knowledge guided pseudo labels outperform strong 2D ConvNeXt and CNN–Transformer (Manzari et al., 2023) baselines. Together, these components provide a practical and scalable solution for CMR sequence and view recognition and lay the foundation for more advanced CMR representation learning and downstream analysis.

2. Related Work

Accurate identification of MRI sequence types is essential for ensuring consistent downstream analysis, and it becomes impractical as datasets grow larger and more diverse. Au-

tomatic MRI sequence identification can provide reliable inputs for downstream tasks and enables robust and scalable analysis pipelines. Early MRI sequence identification methods relied mainly on heuristics derived from DICOM tags such as field strength, echo time, and sequence name (Liang et al., 2021). However, these methods can be difficult to generalize in clinical archives where headers may be incomplete and series descriptions are highly inconsistent across technologists, scanners, and vendors.

Recent work has shifted toward image-based deep learning methods, which extract sequence features from images and remain reliable even without trustworthy metadata (Ranjbar et al., 2020; de Mello et al., 2021; Helm et al., 2024). For example, De Mello et al. achieved strong classification performance by training a ResNet-18 model on 3D brain MRI volumes with random slice selection as channel input (de Mello et al., 2021). In cardiac MRI, Lim et al. proposed a CNN tailored for classification that remains robust across vendors and protocols, with labels generated through DICOM-based metadata extraction followed by expert verification based on three selected slices as input (Lim et al., 2022). Ranjbar et al. built a 2D deep neural network to annotate the type of MR image sequence for scans of brain tumor patients (Ranjbar et al., 2020). Helm et al. developed a 3D DenseNet-121 model capable of classifying MRI sequences across chest, abdominal, and pelvic acquisitions (Helm et al., 2024). Recent work by Mahmutoglu et al. address domain-shift challenges in body MRI sequence classification by evaluating CNN–Transformer models on adult-to-pediatric transfer. Their findings show that MedViT, especially with expert-guided adjustments, substantially improves robustness under cross-population variability (Mahmutoglu et al., 2025). Despite these advances, CMR sequence classification remains substantially underexplored, where the imaging workflow shows greater variability and cardiac motion is more complex. View classification is also restricted to cine images, leaving other important modalities such as LGE underexplored. These constraints limit generalization of model and restrict downstream tasks like segmentation or localization to cine-only settings.

3. Methodology

3.1. Training label acquisition

To train the proposed model, we require supervisory labels that specify two key attributes of each CMR image: the sequence type (e.g., CINE, LGE, T2 MAP) and the anatomical view (e.g., SAX, 4CH, 2CH). These labels are essential for guiding the model to recognize clinically meaningful patterns. In clinical practice, CMR datasets usually include a sequence description—a short textual string automatically generated or edited during acquisition. During a CMR study, technologists queue default sequences from scanner preset protocol trees, but these sequence names are frequently modified to reflect the imaging plane, accommodate patient heart rate, or denote minor parameter adjustments. As a result, series descriptions become heterogeneous and non-standardized. Deriving accurate sequence and view labels from such DICOM metadata requires complex rule sets and expert validation, which becomes impractical at scale (Lim et al., 2022). Despite this variability, these descriptions still encode useful hints about both the imaging sequence and the view orientation, making them a valuable source for automated label extraction.

We propose a label-acquisition strategy that leverages sequence descriptions directly. Instead of relying on manual annotation, which is time-consuming and difficult to scale, we

use a large language model (LLM) to infer the desired labels from the textual descriptions. The LLM interprets the shorthand terms, abbreviations, and protocol-specific keywords embedded in each description and maps them to a unified set of predefined labels. This enables automatic, scalable, and reproducible label extraction from existing metadata without modifying the imaging pipeline.

Although LLMs provide an efficient mechanism for generating pseudo-labels in settings with limited annotations, general-purpose LLMs are not trained to interpret CMR-specific terminology or heterogeneous site-dependent naming conventions. Fine-tuning such models would require substantial curated data and computational resources, making it impractical for many clinical research environments. As a result, prompt optimization becomes the primary lever for improving pseudo-label quality and ensuring that the LLM can reliably interpret non-standardized CMR metadata.

Domain knowledge guided prompt. To further improve label quality, we introduce a domain knowledge guided prompting strategy that embeds cardiac MR domain knowledge directly into the LLM query. A naïve prompt simply instructs the model to "Select exactly one from: CINE, LGE, WATER CINE, T1 MAP PRE ...", relying solely on the sequence description for inference. However, many sequence descriptions contain cryptic abbreviations, scanner-specific shorthand, or institution-dependent naming patterns that can mislead a general LLM. To address this, we augment the prompt with explicit domain knowledge that captures how specific sequence types are commonly encoded in practice.

For example, we inform the LLM that LGE sequences are often denoted by keywords such as "DME", "DE", "SSHOT", "LGE", or "PSIR". By incorporating such guidance, the model can recognize semantically related but heterogeneous descriptors and map them to the correct standardized label. This domain-informed prompt design substantially increases the robustness of label inference, especially in real-world datasets where sequence names vary widely across vendors, scanners, and acquisition workflows.

To ensure maximum data security, all labeling was executed locally without external API calls. A GPT-OSS-20B model was used to generate sequence and view pseudo-labels. Images from 20 patients were manually reviewed in a DICOM viewer to verify labeling accuracy and establish a reliable benchmark for evaluation.

3.2. Network architecture

Clinical CMR acquisitions are inherently multidimensional, varying across temporal phases (e.g., CINE), inversion times (e.g., T1 mapping), and spatial positions (e.g., multi-slice LOCALIZER). Multiple sequences can exhibit similar image contrast, making single-frame classification unreliable. To address this, we developed a spatial–temporal xLSTM architecture to extract compact 1D representations from heterogeneous CMR inputs.

Each CMR image—such as 2D+T cine, 2D+TI T1 mapping, or 2D+D localizer—is first decomposed into a sequence of 2D frames. Each frame is independently encoded by a ConvNeXt backbone (Liu et al., 2022) pretrained on ImageNet, which extracts spatial features and produces a sequence of fixed-dimensional latent vectors. The xLSTM module captures long-range dependencies and integrates spatial–temporal context through gated recurrent transformations (Beck et al., 2024b). It compresses the fixed-dimensional latent vectors into a single continuous 1D embedding that consolidates information across time,

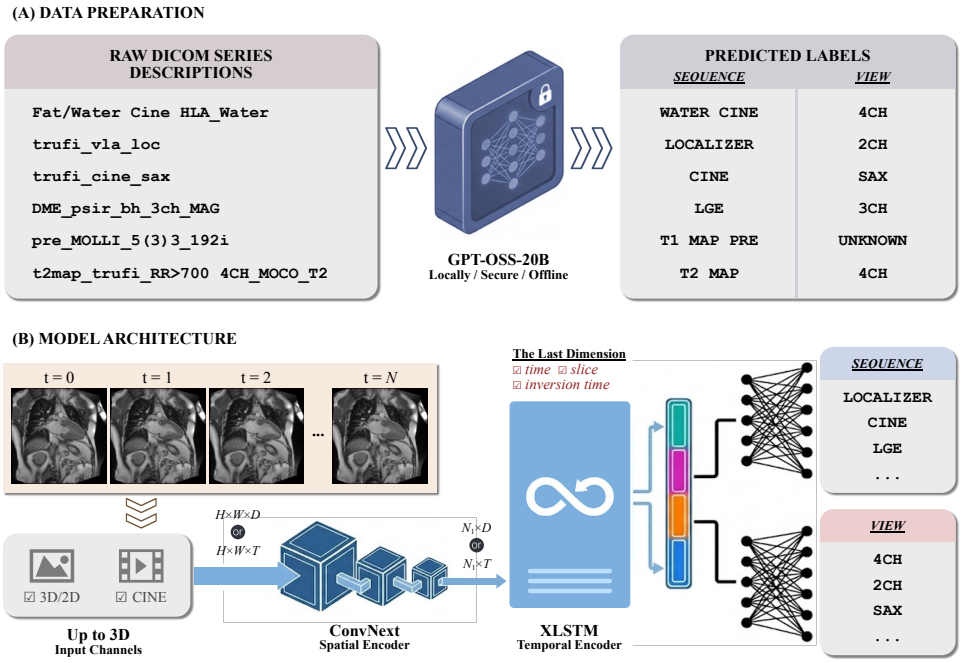

Figure 2: Overview of our proposed framework. Panel (A) shows the data preparation pipeline, highlighting the variation of DICOM series descriptions and. Panel (B) shows the architecture of proposed model.

inversion time, or spatial slice dimensions. This final representation is then passed to task-specific linear heads for sequence and view classification.

Let the input image be denoted by $I \in \mathbb{R}^{H\times W\times K}$, where the third dimension K may represent temporal frames T, inversion times TI, or slice index D. We first apply a ConvNeXt encoder $\mathcal{F}_{\text{ConvNeXt}}$ to each of the K feature planes, yielding a sequence of feature vectors $X = \mathcal{F}_{\text{ConvNeXt}}(I) \in \mathbb{R}^{N_1\times K}$. This sequence is then fed into an xLSTM module $\mathcal{F}_{\text{xLSTM}}$, which aggregates information along the K-dimension and produces a final representation $E = \mathcal{F}_{\text{xLSTM}}(X) \in \mathbb{R}^{N_2}$. In our implementation, K is fixed at 15. For images with fewer than 15 frames, the images are repeated; for those with more, a random subset of 15 is sampled during training.

Loss function. For both sequence classification and view classification, we adopt the standard cross-entropy loss, which is well suited for multi-class prediction tasks. The total training loss is computed as the sum of the sequence and the view cross-entropy losses.

4. Experiments

4.1. Data collection

We collected CMR studies from 1,000 patients at the Ohio State University Wexner Medical Center. All data were anonymized before analysis. The dataset contains approxi-

Table 1: Evaluation of sequence and view labels generated by GPT-OSS-20B under two prompting strategies, using expert annotations as reference. "Unknown" indicates series where the model failed to assign a label and is included in the accuracy calculation, contributing to lower overall metrics.

Evaluation Target	**Sequence**			**View**		
	Unknown	Accuracy	F1-Score	Unknown	Accuracy	F1-Score
General Prompt	104	0.8589	0.7680	352	0.6297	0.5536
Domain-informed Prompt	**2**	**0.9962**	**0.9318**	**228**	**0.7737**	**0.7906**

mately 76,000 CMR series, which were divided into training, validation, and testing subsets (882/98/20 patients). For each series, the DICOM tag `SeriesDescription` (0008,103E) was extracted and processed by the domain knowledge guided prompt to generate sequence and view pseudo labels. An expert reader manually reviewed all series from the 20 test patients in a DICOM viewer to establish the reference labels used for evaluation.

4.2. Training and implementation details

All models were trained on a single NVIDIA A100 GPU. Training was performed for 25 epochs using a batch size of 8 and a learning rate of 1×10^{-4} with ADAM optimizer. We saved both the checkpoint with the highest validation accuracy and the checkpoint from the final epoch. For configurations that exceeded the memory capacity of one A100 GPU, the batch size was reduced to 4 and gradient accumulation of 2 was applied to maintain an effective batch size of 8.

4.3. Effect of domain knowledge guided prompting

We evaluated a locally deployed GPT-OSS-20B model for automatic extraction of CMR sequence and view labels from Series Description fields. Two prompting strategies were compared: (1) General prompt, which provides minimal task specification; and (2) Domain-informed prompt, which incorporates CMR-specific mapping rules, common abbreviations, and disambiguation logic derived from expert practice. The comparison of two types of prompts is provided in the Appendix (Figure 5).

To quantify label quality, both prompt outputs were produced for the entire dataset and evaluated against an expert-reviewed test set. Accuracy and F1 score were computed for sequence and view categories. We then trained the proposed spatial–temporal classifier using pseudo-labels from each prompting strategy and assessed model performance on the same expert-annotated test set.

The results (Table 1) show that domain-informed prompting substantially improves pseudo-label quality over the general prompt. Models trained on domain-optimized pseudo-labels also achieved higher accuracy and F1 scores, indicating that improvements in label fidelity translate directly into downstream model performance (Table 2). These findings confirm that prompt engineering is an effective and efficient mechanism to adapt LLMs for CMR metadata interpretation.

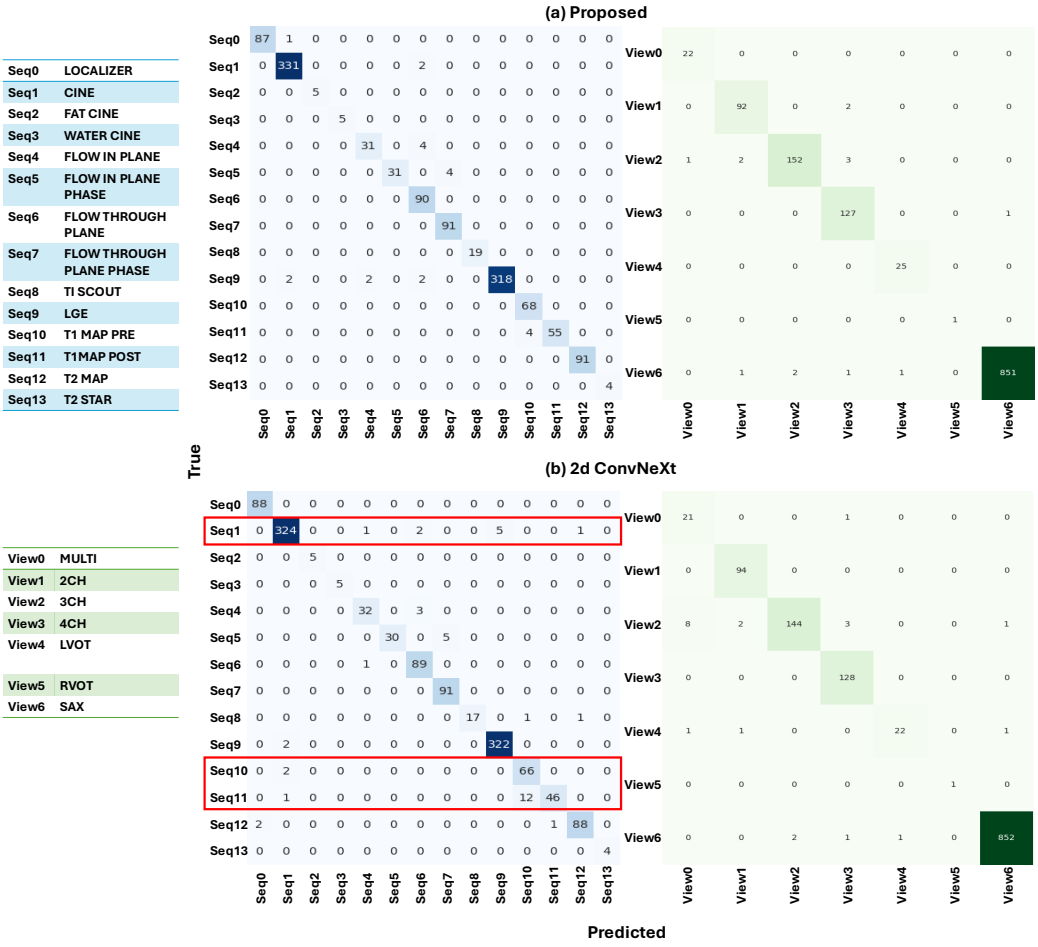

Figure 3: Confusion matrices for sequence and view classification. Panel (a) shows results from the proposed model, and panel (b) shows results from the baseline 2D ConvNeXt model. The corresponding categories are listed on the left.

4.4. Effect of spatiotemporal modeling

To evaluate the importance of modeling this multidimensional structure, we compared our proposed architecture against a strong 2D baseline model. The proposed model employs a pretrained ConvNeXt as the spatial encoder to extract features from each temporal frame, inversion frame, or slice. An xLSTM temporal encoder then models dependencies across frames through spatiotemporal sequence learning. This design enables the network to incorporate temporal and slice-wise information that cannot be recovered from individual 2D frames. In contrast, the 2D baseline uses a pretrained ConvNeXt trained on a single representative frame sampled from each series following standard 2D classification protocols.

The comparative results in Table 2 show that the proposed spatiotemporal model achieves higher accuracy in sequence classification, while both models achieve similar perfor-

mance in view classification. This outcome is consistent with expectations. Many sequences share similar contrast, making it challenging for a 2D model to differentiate among them without temporal context. Cine imaging is a representative example. Individual frames are often visually ambiguous, leading to frequent misclassification. The red-highlighted confusion in Figure 3 illustrates common failure modes of the 2D baseline, underscoring its limitations in capturing temporal or contextual cues. In contrast, view classification relies primarily on anatomical localization, which can typically be inferred from a single frame. Therefore, adding temporal modeling provides limited additional benefit for this task. These findings highlight that explicit spatiotemporal modeling is important for accurate CMR sequence classification, while view recognition depends less on temporal information.

4.5. Baselines comparison

Recent CMR sequence classification studies commonly adopt hybrid architectures that combine a convolutional feature extractor with a Vision Transformer (ViT) for global feature aggregation. These models aim to retain local spatial detail while enabling global token-to-token reasoning. Prior work has reported strong performance with these CNN–ViT hybrids, but, to our knowledge, none have publicly released their training pipelines, making direct comparison difficult. To establish a fair baseline, we implemented a representative CNN–ViT model following standard design practices in medical image analysis.

A CNN encoder produces 2D feature embeddings for each frame in a CMR series. These frame-wise embeddings are then treated as a sequence of tokens and passed through a ViT encoder that performs self-attention across time. Two linear classification heads, identical to those used in our framework, predict sequence type and imaging view. This design tests whether generic transformer-based temporal modeling on frame-level embeddings is sufficient for CMR classification, or whether the explicit spatiotemporal inductive biases in our xLSTM architecture provide measurable gains.

Training and evaluation followed the identical protocol described earlier. Results (Table 2) show that the CNN–ViT hybrid underperforms relative to our proposed model, highlighting the importance of incorporating explicit temporal and slice-wise structure rather than relying solely on 2D tokenization and global attention.

A t-SNE visualization (Figure 4) of the learned embedding further illustrates these effects. When projected into 2D, view classes form clear and coherent clusters, indicating that the model captures anatomical geometry reliably. Sequence clusters, however, appear partially overlapping in the 2D projection, reflecting the limitations of t-SNE in preserving high-dimensional structure rather than deficiencies in the representation itself. Despite this visual overlap, the high-dimensional embedding supports near-perfect sequence classification, consistent with the quantitative results.

5. Discussion and Conclusion

We presented an integrated framework for automated CMR sequence and view classification that targets two persistent obstacles in large scale CMR analysis: nonstandard metadata in clinical workflows and the multidimensional structure of CMR acquisitions. Our approach combines a ConvNeXt spatial encoder with an xLSTM temporal module to map heterogeneous CMR series into a compact one dimensional embedding, and uses a locally

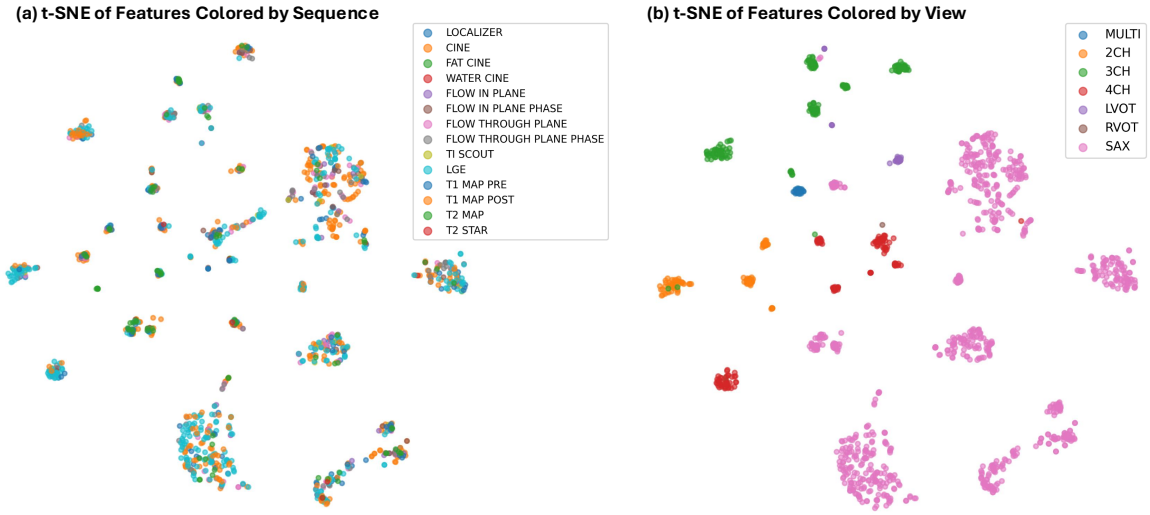

Figure 4: Two-dimensional t-SNE visualization of the shared embedding. (a) Colored by sequence label, showing partial overlap due to projection limitations. (b) Colored by view label, showing clear anatomical clustering.

Table 2: Summary of quantitative results for all model-training experiments. The table reports sequence and view classification performance for each model configuration

Evaluation Target	**Sequence**		**View**	
	Accuracy	F1-Score	Accuracy	F1-Score
2D ConvNeXT	0.9679	0.9593	0.9829	0.9476
MedViT	0.9404	0.9052	0.9533	0.8532
Proposed Model (General Prompt labels)	0.8869	0.8363	0.9283	0.8668
Proposed Model (Domain-informed prompt labels)	**0.9832**	**0.9786**	**0.9891**	**0.9822**

deployed GPT-OSS model with a domain knowledge guided prompt to generate secure and standardized pseudo labels from DICOM series descriptions.

On a large institutional dataset, this framework consistently outperformed strong two dimensional and CNN-Vision Transformer baselines, highlighting the importance of explicit spatiotemporal inductive biases when distinguishing sequences with similar contrast or acquisition parameters. The unified embedding supports accurate sequence and view recognition and offers a reusable representation for downstream applications such as protocol harmonization, quality control, quantitative parameter estimation, and disease specific phenotyping. Future work will include multi-center validation, extension to additional vendors and pathologies, and integration with downstream tasks such as segmentation and prognostic modeling to further assess how secure large scale labeling coupled with multidimensional representation learning generalizes across CMR workflows.

References

Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K Arora, Yu Bai, Bowen Baker, Haiming Bao, et al. gpt-oss-120b & gpt-oss-20b model card. *arXiv preprint arXiv:2508.10925*, 2025.

Maximilian Beck, Korbinian Pöppel, Markus Spanring, Andreas Auer, Oleksandra Prudnikova, Michael Kopp, Günter Klambauer, Johannes Brandstetter, and Sepp Hochreiter. xlstm: Extended long short-term memory. *Advances in Neural Information Processing Systems*, 37:107547–107603, 2024a.

Maximilian Beck, Korbinian Pöppel, Markus Spanring, Andreas Auer, Oleksandra Prudnikova, Michael Kopp, Günter Klambauer, Johannes Brandstetter, and Sepp Hochreiter. xlstm: Extended long short-term memory, 2024b. URL https://arxiv.org/abs/2405.04517.

Jan Bogaert, Steven Dymarkowski, Andrew M Taylor, and Vivek Muthurangu. *Clinical cardiac MRI*. Springer Science & Business Media, 2012.

Jean Pablo Vieira de Mello, Thiago M Paixão, Rodrigo Berriel, Mauricio Reyes, Claudine Badue, Alberto F De Souza, and Thiago Oliveira-Santos. Deep learning-based type identification of volumetric mri sequences. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 1–8. IEEE, 2021.

Andrew S Flett, Jonathan Hasleton, Christopher Cook, Derek Hausenloy, Giovanni Quarta, Cono Ariti, Vivek Muthurangu, and James C Moon. Evaluation of techniques for the quantification of myocardial scar of differing etiology using cardiac magnetic resonance. *JACC: cardiovascular imaging*, 4(2):150–156, 2011.

Kimberly Helm, Tejas Sudharshan Mathai, Boah Kim, Pritam Mukherjee, Jianfei Liu, and Ronald M Summers. Automated classification of body mri sequence type using convolutional neural networks. In *Medical Imaging 2024: Computer-Aided Diagnosis*, volume 12927, pages 119–123. SPIE, 2024.

Peter I Kamel, Florence X Doo, Dharmam Savani, Adway Kanhere, Paul H Yi, and Vishwa S Parekh. Standardizing heterogeneous mri series description metadata using large language models. *Journal of Imaging Informatics in Medicine*, pages 1–11, 2025.

Theodoros D Karamitsos, Jane M Francis, Saul Myerson, Joseph B Selvanayagam, and Stefan Neubauer. The role of cardiovascular magnetic resonance imaging in heart failure. *Journal of the American College of Cardiology*, 54(15):1407–1424, 2009.

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.

Shuai Liang, Derek Beaton, Stephen R Arnott, Tom Gee, Mojdeh Zamyadi, Robert Bartha, Sean Symons, Glenda M MacQueen, Stefanie Hassel, Jason P Lerch, et al. Magnetic

resonance imaging sequence identification using a metadata learning approach. *Frontiers in Neuroinformatics*, 15:622951, 2021.

Ruth P Lim, Stefan Kachel, Adriana DM Villa, Leighton Kearney, Nuno Bettencourt, Alistair A Young, Amedeo Chiribiri, and Cian M Scannell. Cardisort: a convolutional neural network for cross vendor automated sorting of cardiac mr images. *European radiology*, 32 (9):5907–5920, 2022.

João AC Lima and Milind Y Desai. Cardiovascular magnetic resonance imaging: current and emerging applications. *Journal of the American College of Cardiology*, 44(6):1164–1171, 2004.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11976–11986, 2022.

Mustafa Ahmed Mahmutoglu, Aditya Rastogi, Gianluca Brugnara, Philipp Vollmuth, Martha Foltyn-Dumitru, Felix Sahm, Stefan Pfister, Dominik Sturm, Martin Bendszus, and Marianne Schell. Optimizing mri sequence classification performance: insights from domain shift analysis. *European Radiology*, pages 1–9, 2025.

Omid Nejati Manzari, Hamid Ahmadabadi, Hossein Kashiani, Shahriar B Shokouhi, and Ahmad Ayatollahi. Medvit: a robust vision transformer for generalized medical image classification. *Computers in biology and medicine*, 157:106791, 2023.

Sara Ranjbar, Kyle W Singleton, Pamela R Jackson, Cassandra R Rickertsen, Scott A Whitmire, Kamala R Clark-Swanson, J Ross Mitchell, Kristin R Swanson, and Leland S Hu. A deep convolutional neural network for annotation of magnetic resonance imaging sequence type. *Journal of digital imaging*, 33(2):439–446, 2020.

Michael Salerno, Behzad Sharif, Håkan Arheden, Andreas Kumar, Leon Axel, Debiao Li, and Stefan Neubauer. Recent advances in cardiovascular magnetic resonance: techniques and applications. *Circulation: Cardiovascular Imaging*, 10(6):e003951, 2017a.

Michael Salerno, Behzad Sharif, Håkan Arheden, Andreas Kumar, Leon Axel, Debiao Li, and Stefan Neubauer. Recent Advances in Cardiovascular Magnetic Resonance Techniques and Applications. *Circulation. Cardiovascular imaging*, 10(6):e003951, June 2017b. ISSN 1941-9651. doi: 10.1161/CIRCIMAGING.116.003951. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC5777859/.

Yuli Wang, Kritika Iyer, Sep Farhand, and Yoshihisa Shinagawa. Evaluating unsupervised contrastive learning framework for mri sequences classification. *arXiv preprint arXiv:2501.06938*, 2025.

Appendix A. Prompts

	Sequence Extraction	View Extraction
	System: You are an expert in CMR sequence classification.	**System:** You are an expert in CMR sequence classification.
General Prompt	**Task**: Extract the sequence name from the Series Description. **Choices:** CINE, LGE, FAT CINE, WATER CINE, T1 MAP PRE, T1 MAP POST, T2 MAP, T2 STAR, DTI, HASTE, PERFUSION, FLOW THROUGH PLANE, FLOW IN PLANE, FLOW THROUGH PLANE PHASE, FLOW IN PLANE PHASE, LOCALIZER, TI SCOUT, AXIAL STACK, MRA, unknown. **Output Rules:** - Do not infer or assume missing information. - Output exactly one label from the list. - If the sequence cannot be determined, output "unknown".	**Task**: Extract the cardiac view from the Series Description. **Choices**: 2CH, 3CH, 4CH, RV3CH, LVOT, RVOT, RVIN, SAX, AO, MPA, MV, TV, MVTV, MULTI, AXIAL, SAGITTAL, CORONAL, unknown. **Output Rules**: - Matching is case-insensitive. - Use only explicit information. - If multiple labels match, choose the first in the list. - Output exactly one label; if none apply, output "unknown".
Domain-informed Prompt	**Task:** [same as above] **Choices:** [same list as above] **Mapping rules:** - "MOLLI" → T1 MAP PRE (default) or POST if specified. - "DME", "DE", "SSHOT", "LGE", "PSIR" → LGE. - "loc", "localizer" → LOCALIZER. - "MRA", "CEMRA" → MRA. - Flow: - Contains "flow", "throughplane", "inplane" → Flow. - "PHASE" or suffix "_P" → Phase flow. - "MAG", "MAGNITUDE" or no "PHASE" → Magnitude flow. - "AO", "AORTA", "MV", "MPA", "TV" → Through-plane; otherwise In-plane. - Cine: - Both "fat" and "water" → choose last occurrence. - Only "water" → WATER CINE. - Only "fat" → FAT CINE. - Else → CINE. **Output rules**: [same as above]	**Task**: [same as above] **Choices**: [same list as above] **Mapping rules**: - "mv/tv", "mv-tv", "mv tv" → MVTV - 2CH or VLA → 2CH - 4CH or HLA → 4CH - "3ch", "three chamber" → 3CH - "rvin", "rvinf" → RVIN- MV → indicators of mitral valve. - "sax", "sa loc", "sa_loc", "multi_sa", "multi sa", "short axis" → SAX - "aortic", "ao", "aorta" → AO - "mpa", "main pulmonary artery" → MPA - "mitral valve", " mv " → MV - "tricuspid valve", " tv " → TV Output: - Use only explicit information + rules above. - Return exactly one label; if none match, output "unknown". **Output rules**: [same as above]

Figure 5: Prompt design used for automated extraction of sequence and view labels from CMR Series Descriptions. Each column corresponds to a prediction task, and each row shows the system instruction, general prompt, and domain-informed prompt used to guide label generation. For clarity and space considerations, the prompts shown here are condensed versions of the full instructions used in the study.