Global-Local Structure-Aware Alignment for Automated LGE MRI Report Generation

Chenyang Zhou[1], Yun Tian[1*], Wenqi Lu[2], Sandeep S. Hothi[3], Jinyu Zheng[4], Shifeng Zhao[1], and Jinming Duan[5*]

[1] School of Artificial Intelligence, Beijing Normal University, Beijing, China
tianyun@bnu.edu.cn
[2] Department of Computing and Mathematics, Manchester Metropolitan University, Manchester, UK
[3] Department of Cardiology, Heart and Lung Centre, Royal Wolverhampton NHS Trust, Wolverhampton, UK
[4] Department of Radiology, Renji Hospital, Shanghai Jiao Tong University, Shanghai, China
[5] Division of Informatics, Imaging and Data Sciences, University of Manchester, Manchester, UK
jinming.duan@manchester.ac.uk

Abstract. Automated Late Gadolinium Enhancement (LGE) MRI report generation is clinically important for supporting structured cardiac assessment and improving diagnostic efficiency. However, existing methods predominantly rely on CLIP-style global vision–language alignment, which captures coarse image-level semantics but fails to model region-level scar enhancement patterns and cross-slice structural continuity essential for reliable LGE interpretation. To address this limitation, we propose Global–Local Structure-Aware Alignment for LGE Report Generation (GLSA-RG). The framework introduces a unified Structural Graph Representation to encode intra-slice regional interactions, cross-slice continuity, and structured textual dependencies. Building upon these representations, a novel Structure-Aware Graph Alignment mechanism performs fine-grained multi-instance contrastive alignment, complementing global volume–report supervision. This global–local alignment strategy enables explicit structure-consistent cross-modal correspondence while preserving overall semantic coherence. Extensive experiments on two clinical datasets demonstrate that GLSA-RG better captures subtle myocardial enhancement patterns and maintains cross-slice structural consistency, resulting in clinically more reliable LGE MRI reports.

Keywords: Late Gadolinium Enhancement · Report Generation · Global-Local Structure-Aware Alignment.

1 Introduction

Late Gadolinium Enhancement (LGE) has become the clinical gold standard for identifying myocardial scar and fibrosis [7,5]. Accurate interpretation of LGE

images is indispensable for treatment planning in patients with cardiovascular diseases. Clinicians perform detailed slice-by-slice assessment while simultaneously integrating information across slices to form a global interpretation, rendering the process time-consuming and operator-dependent [6,11]. Consequently, automated LGE report generation holds substantial potential for improving diagnostic efficiency and supporting standardized clinical workflows.

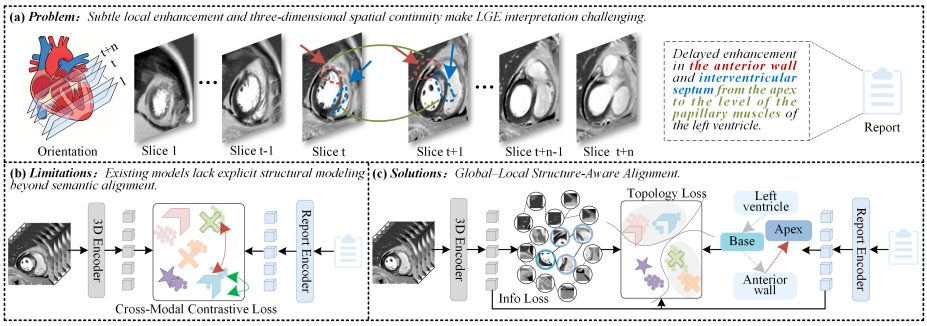

Fig. 1. (a) LGE images present subtle and complex patterns. (b) These characteristics pose challenges for automated report generation. (c) Our method addresses these challenges to produce more reliable reports.

Recent multimodal large language models (MLLMs) have shown promising performance in medical report generation [12,23,25]. However, reliable LGE report generation remains challenging (Fig. 1a). While these approaches perform well on 2D imaging modalities [22,14], they typically operate on single-image inputs and lack volume-level reasoning capability [13]. As shown in Fig. 1(b), existing CT and MRI report generation frameworks generally adopt volumetric visual encoders and compress volume features into a single global representation, performing CLIP-style global semantic alignment between the volume and the report [2,24]. However, given that 3D medical image–text datasets are substantially smaller than natural image–text corpora, reliance on global compression and global alignment may weaken fine-grained structural representations, causing subtle yet clinically critical pathological patterns to be overshadowed by dominant global semantics [19]. To mitigate this limitation, several approaches introduce 2D-enhanced mechanisms within 3D modeling frameworks, such as Med-2E3 [21] and HSENet [19], which encode volumetric data and individual slices separately and fuse them at the feature level to incorporate both global context and local visual details. Although such designs improve representational capacity, they remain based on feature-level aggregation, leading to insufficient modeling of fine-grained spatial dependencies and cross-slice structural consistency.

To address this limitation, we propose **G**lobal–**L**ocal **S**tructure-**A**ware **A**lignment for LGE Report Generation (GLSA-RG), a framework that shifts from

purely global semantic matching to explicit structure-aware cross-modal alignment. GLSA-RG consists of two key components. As shown in Fig. 2, we introduce a unified **Structural Graph Representation (SGR)** module to explicitly model anatomical structure in both visual and textual modalities. For volumetric images, we model intra-slice regional interactions to preserve local enhancement patterns and inter-slice dependencies to maintain short-axis continuity. For textual reports, we organize descriptions into a standardized myocardial segment graph, providing structured anchors for fine-grained correspondence. However, structural modeling alone does not guarantee that corresponding visual and textual structures are consistently aligned. We therefore propose a **Structure-Constrained Alignment (SCA)** module, which enforces cross-modal alignment between structure-aware representations, jointly optimized with global volume–report alignment. By explicitly coupling structural representation learning with structure-constrained alignment, GLSA-RG enables reliable and clinically consistent LGE report generation. **Our main contributions are:**

1. We propose a global–local structure-aware alignment framework for LGE report generation, which integrates unified structural graph representations with structure-constrained cross-modal alignment to explicitly model region-level interactions and spatial continuity beyond global semantic matching.
2. Extensive experiments demonstrate that our method consistently outperforms existing approaches across multiple evaluation metrics.
3. We construct a multi-center LGE MRI dataset with paired clinical reports, addressing the scarcity of data for automated cardiac report generation.

2 Methodology

As illustrated in Fig. 2, GLSA-RG is designed to address the limitations of global semantic alignment in LGE report generation. The framework comprises two components. The Structural Graph Representation module builds unified visual and textual structures through hierarchical graph modeling (Sec. 2.1). The Global–Local Structure-Aware Alignment module aligns them at both global and structural levels, jointly optimized with report generation (Sec. 2.2).

2.1 Structural graph representation module

Global volume representation. As shown in Fig. 2(a), we extract a global volumetric representation that captures high-level semantic information of the LGE volume and serves as global supervision during joint training. Let the visual embedding from the volume encoder be $\mathbf{z}_{\text{img}}^{\text{Global}} \in \mathbb{R}^d$ and the corresponding textual embedding be $\mathbf{z}_{\text{text}}^{\text{Global}} \in \mathbb{R}^d$.

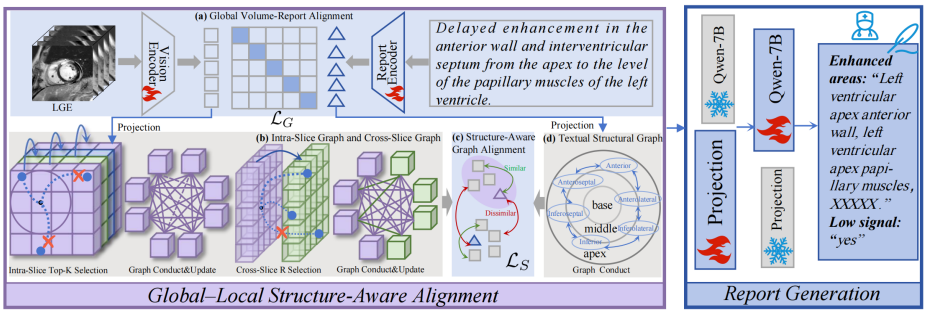

Fig. 2. Overview of GLSA-RG. (b) and (d) correspond to structural feature learning: (b) learns hierarchical visual structural representations through intra-slice and cross-slice graph modeling, while (d) constructs structured textual representations via a standardized segment-level graph. (a) and (c) perform cross-modal alignment: (a) captures global volume–report semantic alignment, and (c) establishes structure-aware graph alignment at the node level. We further adopt a three-stage training paradigm.

Intra-slice graph. The intra-slice graph (ISG) models spatially coherent regional interactions within each short-axis slice by jointly enforcing geometric locality and semantic consistency. Given patch-level features $\tilde{\mu}_d = \{\tilde{\mu}_d^j\}_{j=1}^{N_d}$ from slice d, we construct a node set $V_d = \{v_d^j\}_{j=1}^{N_d}$, where each node v_d^j corresponds to patch feature $\tilde{\mu}_d^j$ with in-plane center $p_d^j \in \mathbb{R}^2$. We project the features into a shared embedding space using a linear mapping $\hat{\mu}_d^j = \tilde{\mu}_d^j W$. To avoid dense and noisy connections, candidate neighbors of v_d^j are restricted to a local geometric neighborhood $\mathcal{N}_{\text{geo}}(v_d^j) = \{v_d^k \mid \|p_d^j - p_d^k\|_2 \le r_1\}$, where r_1 controls local spatial continuity within each slice. Within this set, we retain the $TopK$ most semantically similar patches to form the intra-slice graph, as shown in Eq. (1):

$$\mathcal{N}_{\text{ISG}}(v_d^j) = \text{TopK}\Big(\text{sim}(\hat{\mu}_d^j, \hat{\mu}_d^k) \;\Big|\; v_d^k \in \mathcal{N}_{\text{geo}}(v_d^j)\Big). \tag{1}$$

The resulting sparse graph $G_d^{\text{ISG}} = (V_d, E_d^{\text{ISG}})$ connects only spatially adjacent patches with consistent enhancement patterns. We then apply a multi-head graph attention network to aggregate information from $\mathcal{N}_{\text{ISG}}(v_d^j)$ and obtain structure-enhanced node representations $\mu_d^{\text{ISG},j}$, which preserve intra-slice topology while improving regional coherence.

Cross-slice graph. To capture structural continuity along the short-axis direction, we construct a cross-slice graph (CSG) between adjacent slices by modeling geometric alignment and semantic correspondence. Given structure-enhanced node features $\mu_d^{\text{ISG},j}$ and $\mu_{d+1}^{\text{ISG},k}$ from slices d and $d{+}1$, with in-plane patch centers $p_d^j, p_{d+1}^k \in \mathbb{R}^2$, we linearly project the features into a shared matching space before cross-slice attention. Cross-slice connections are restricted to spatially

aligned patches, as shown in Eq. (2):

$$\mathcal{N}_{\mathrm{CSG}}(v_d^j) = \{v_{d+1}^k \mid \|p_d^j - p_{d+1}^k\|_2 \le r_2\}, \tag{2}$$

where r_2 controls inter-slice spatial tolerance. A distance-aware prior is incorporated into attention to softly favor better-aligned patches. Graph attention is then applied over $\mathcal{N}_{\mathrm{CSG}}(v_d^j)$ to aggregate cross-slice information. By connecting only adjacent slices, CSG enforces short-axis continuity, while stacking multiple layers enables progressive information propagation across the volume.

Textual structural graph. As shown in Fig. 2c, We construct a Textual Structural Graph (TSG) based on the standardized AHA 17-segment myocardial model [4]. Each node corresponds to a predefined myocardial segment, serving as a clinically standardized indexing scheme to organize report descriptions in a consistent structural space across samples. The graph topology follows the adjacency relations defined in the AHA model, connecting segments within the same circumferential ring or adjacent rings to reflect spatial proximity in a standardized cardiac layout. Additionally, corpus-level co-occurrence statistics from the training reports are incorporated as fixed auxiliary edges to reinforce frequently co-mentioned segment pairs. Let $H^{\text{text}} \in \mathbb{R}^{N_t \times d}$ denote the initial segment-level embeddings. A multi-layer graph attention network propagates contextual information over the predefined graph:

$$H^{(l+1)} = \mathrm{GAT}^{(l)}(H^{(l)}, A), \tag{3}$$

where A encodes the predefined structural adjacency. The resulting embeddings capture structured textual dependencies that facilitate fine-grained alignment.

2.2 Global-local structure-aware alignment

Global volume-report alignment. Let $\mathbf{z}_{\text{img}}^{G,i}$ and $\mathbf{z}_{\text{text}}^{G,i}$ denote the global image and text embeddings of the i-th sample in a mini-batch of size B. Following common paradigms [18], we optimize a symmetric InfoNCE objective [16]:

$$\mathcal{L}_{\mathrm{GVA}} = -\frac{1}{2B}\sum_{i=1}^{B}\left[\log\frac{\exp(\mathrm{sim}(\mathbf{z}_{\text{img}}^{G,i}, \mathbf{z}_{\text{text}}^{G,i})/\tau)}{\sum_{j=1}^{B}\exp(\mathrm{sim}(\mathbf{z}_{\text{img}}^{G,i}, \mathbf{z}_{\text{text}}^{G,j})/\tau)} + \log\frac{\exp(\mathrm{sim}(\mathbf{z}_{\text{text}}^{G,i}, \mathbf{z}_{\text{img}}^{G,i})/\tau)}{\sum_{j=1}^{B}\exp(\mathrm{sim}(\mathbf{z}_{\text{text}}^{G,i}, \mathbf{z}_{\text{img}}^{G,j})/\tau)}\right]. \tag{4}$$

Structure-aware graph alignment. Let $\{\mathbf{z}_{\text{img}}^{L,i,n}\}_{n=1}^{N_v}$ and $\{\mathbf{h}_{\text{text}}^{i,m}\}_{m=1}^{N_t}$ denote structure-aware visual node features and textual structural node features, respectively. Inspired by attention-based multiple instance learning [10], we compute a soft correspondence for each textual node $w_{i,m,n} = \mathrm{softmax}_n\Big(\mathrm{sim}(\mathbf{h}_{\text{text}}^{i,m}, \mathbf{z}_{\text{img}}^{L,i,n})/\tau\Big)$. Fine-grained structure-aware alignment is achieved via the multi-instance objective, as shown in Eq. (5).

$$\mathcal{L}_{\mathrm{SGA}} = -\frac{1}{BN_t}\sum_{i=1}^{B}\sum_{m=1}^{N_t}\log\frac{\sum_{n=1}^{N_v} w_{i,m,n}\,\exp(\mathrm{sim}(\mathbf{h}_{\text{text}}^{i,m}, \mathbf{z}_{\text{img}}^{L,i,n})/\tau)}{\sum_{j=1}^{B}\sum_{n=1}^{N_v}\exp(\mathrm{sim}(\mathbf{h}_{\text{text}}^{i,m}, \mathbf{z}_{\text{img}}^{L,j,n})/\tau)}. \tag{5}$$

Overall objective. We supervise report generation with an autoregressive cross-entropy loss conditioned on both global and structure-aware visual representations:

$$\mathcal{L}_{\mathrm{RG}} = -\frac{1}{B}\sum_{i=1}^{B}\sum_{t=1}^{T_i}\log P\Big(y_{i,t} \mid y_{i,<t}, \{\mathbf{z}_{\mathrm{img}}^{G,i}, \mathbf{z}_{\mathrm{img}}^{L,i}\}\Big), \tag{6}$$

where $\mathbf{z}_{\mathrm{img}}^{G,i}$ and $\mathbf{z}_{\mathrm{img}}^{L,i}$ denote the global and structure-aware visual embeddings of the i-th sample. The final objective jointly optimizes global alignment Eq. (4), structure-aware alignment Eq. (5), and report generation Eq. (6), as shown in Eq. (7):

$$\mathcal{L}_{\mathrm{total}} = \lambda_1\mathcal{L}_{\mathrm{GVA}} + \lambda_2\mathcal{L}_{\mathrm{SGA}} + \mathcal{L}_{\mathrm{RG}}. \tag{7}$$

3 Experiments

Datasets: We retrospectively collected cardiac MRI data from 2,038 patients across two clinical centers, each including a short-axis LGE (SAX LGE) sequence and its corresponding diagnostic report. The development cohort (n = 1,870, split into training, validation, and test sets with a ratio of 8:1:1) includes normal, myocardial infarction, and dilated cardiomyopathy (DCM) cases. The external cohort (n = 168) primarily comprises DCM cases to evaluate generalization under clinical domain shift. Variations in report terminology, granularity, and imaging vendors (Siemens vs. Philips) introduce substantial heterogeneity across centers, providing a challenging setting for robustness evaluation. All images were converted to NIfTI format with spatial metadata preserved, and volumes were cropped to 256×256 pixels during training to optimize computational efficiency.

Implementation details and evaluation metrics: We adopt DCFormer [1] as the visual encoder and ClinicalBERT [9] as the report encoder, training the model for 50 epochs with a weight decay of 0.1. For report generation, we use Qwen2.5-7B-Instruct [26] as the language backbone. During instruction tuning, the visual encoder and language model are frozen, and only the multimodal MLP adapter is trained for 3 epochs with a batch size of 16 and a learning rate of 1×10^{-4}. In the final fine-tuning stage, we apply LoRA [8] with rank 16, scaling factor $\alpha = 32$, and dropout 0.05, updating only the LoRA modules and projection layers. All experiments are implemented using PyTorch and trained on two NVIDIA A800 GPUs with DeepSpeed ZeRO-2 and BF16 mixed precision. By default, we set $\lambda_1 = 0.5$ and $\lambda_2 = 0.3$ to balance the different loss terms. Performance is measured via BLEU (BL) [17], ROUGE (RG) [15], METEOR (MTR) [3], and BERTScore (BERT) [27]. The code is available at `https://github.com/luxiaozhou89-netizen/GLSA-RG`. In addition, we used GPT-5 with structured prompts to automatically evaluate the clinical accuracy and consistency of the reports, reported as LLMScore (LLM), providing a complementary assessment to traditional NLG metrics [2].

Table 1. Comparison with state-of-the-art methods on two datasets. † denotes models retrained using Dataset A (training set).

Model	BL-1	BL-2	BL-3	BL-4	RG-1	RG-L	MTR	BERT	LLM
Dataset A (Internal Test)									
M3D† [2]	13.20	10.60	8.39	6.27	23.97	21.54	29.11	70.82	21.87
MedM-VL† [20]	18.29	13.32	10.71	5.42	21.70	20.95	27.18	82.14	29.44
Med-2E3†[21]	20.21	17.30	14.36	10.86	33.34	27.30	29.21	84.21	30.11
Med3DVLM† [24]	36.18	30.28	24.47	19.93	55.12	50.33	47.92	86.78	44.10
HSENet† [19]	28.89	22.36	18.16	14.14	48.93	44.02	41.26	86.15	40.49
GLSA-RG (Ours)	**41.13**	**34.43**	**27.83**	**22.83**	**58.70**	**54.31**	**53.28**	**92.63**	**48.49**
Dataset B (External Validation)									
M3D [2]	8.62	4.17	1.21	0.19	10.76	13.21	19.66	66.51	15.23
MedM-VL [20]	14.29	10.32	6.71	1.42	18.70	17.95	25.18	79.04	26.17
Med-2E3 [21]	15.58	10.57	7.64	2.09	21.52	20.12	27.33	81.11	29.71
Med3DVLM [24]	27.61	22.55	17.16	12.22	38.68	37.83	47.46	83.45	40.11
HSENet [19]	20.46	16.36	10.93	6.40	30.55	28.18	34.96	81.96	38.21
GLSA-RG (Ours)	**35.61**	**28.01**	**23.35**	**17.77**	**45.38**	**49.46**	**47.06**	**88.69**	**43.93**

Comparison with state-of-the-art methods: As shown in Table 1, GLSA-RG consistently outperforms recent 3D medical MLLMs on both datasets. Our method achieves the best performance across all metrics, improving BL-4 from 19.93 (Med3DVLM) to 22.83 and ROUGE-L to 54.31. It also attains the highest BERTScore (92.63) and LLM-based score (48.49), indicating superior semantic fidelity and clinical coherence. Under cross-center domain shift, GLSA-RG maintains clear margins, raising BL-4 from 12.22 to 17.77 and BERTScore from 83.45 to 88.69 compared to the strongest baseline. The stable improvements in MTR and LLM-based evaluation further indicate that structure-aware alignment generalizes well beyond the training distribution. Overall, the results on both datasets demonstrate that explicitly modeling structural representations and performing structure-constrained alignment substantially improves clinical description accuracy and cross-domain robustness. Fig. 3 presents a representative qualitative example, where our method better captures subtle enhancement patterns (e.g., apical inferior wall enhancement), reduces localization errors, and preserves cross-slice continuity, reflecting improved modeling of region-level and cross-slice enhancement. Nevertheless, over-interpretation may still occur in cases with severe motion artifacts. Furthermore, an independent assessment conducted by two cardiologists with over 10 years of clinical experience indicates that the method achieves better performance in terms of enhancement detection accuracy and report usability.

Ablation study on proposed sign graph modules: Table 2(a) evaluates the contribution of the three structural modules. Removing all graph components leads to a clear drop across metrics, confirming the necessity of structure-aware modeling. Without ISG, the CSG+TSG variant yields larger gains on embedding-based and clinical similarity metrics (MTR, BERT, LLM) but smaller

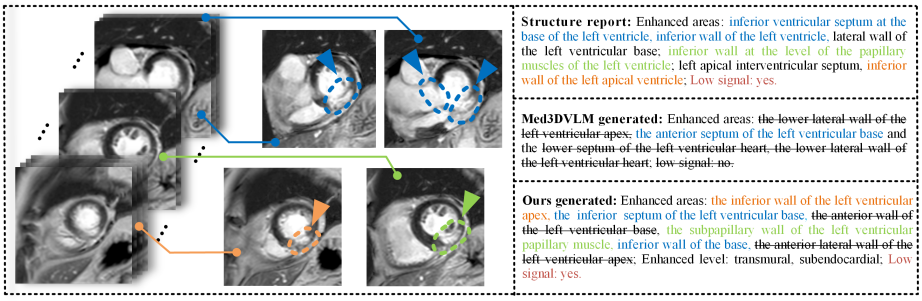

Fig. 3. Comparison between the ground truth and generated LGE reports. Color-coded text corresponds to spatially highlighted enhancement regions. Our method captures subtle local enhancement patterns and preserves spatial continuity across adjacent slices, producing more precise and structurally consistent descriptions compared to existing approaches.

Table 2. Ablation studies of the proposed sign graph framework on Dataset A.

Model	BL-1	BL-2	BL-3	BL-4	RG-1	RG-L	MTR	BERT	LLM
(a) Ablation study on sign graph modules									
ISG× CSG× TSG×	31.68	26.05	19.47	14.17	50.32	46.44	42.92	86.70	40.03
ISG× CSG✓ TSG✓	34.04	28.14	22.12	17.80	53.29	48.18	45.18	90.69	42.70
ISG✓ CSG× TSG✓	33.22	26.75	20.46	15.95	52.96	48.14	43.65	91.07	41.21
ISG✓ CSG✓ TSG✓	**41.13**	**34.43**	**27.83**	**22.83**	**58.70**	**54.31**	**53.28**	**92.63**	**48.49**
(b) Ablation study on graph hyperparameters									
$r_1 = 0, K = 6, r_2 = 6$	33.26	27.52	21.79	17.41	53.12	48.92	45.43	91.50	43.98
$r_1 = 2, K = 4, r_2 = 4$	36.47	30.03	23.60	19.01	54.34	50.83	47.58	91.61	45.15
$r_1 = 2, K = 4, r_2 = 6$	37.23	30.85	24.75	20.09	56.33	51.85	49.32	92.46	46.08
$r_1 = 2, K = 6, r_2 = 6$	**41.13**	**34.43**	**27.83**	**22.83**	**58.70**	**54.31**	**53.28**	**92.63**	**48.49**
(c) Ablation study on graph model order									
CSG → ISG	39.83	32.97	25.40	21.55	55.14	47.66	45.06	90.47	46.17
ISG → CSG	**41.13**	**34.43**	**27.83**	**22.83**	**58.70**	**54.31**	**53.28**	**92.63**	**48.49**

improvements on n-gram metrics, suggesting limited sensitivity to subtle intra-slice patterns. Conversely, ISG+TSG improves local coherence yet lacks explicit cross-slice propagation. Combining ISG, CSG, and TSG consistently performs best, indicating that intra-slice and cross-slice modeling are complementary.

Ablation study on graph hyperparameters: Table 2(b) studies the hyperparameters of ISG and CSG. For ISG, setting $r_1 = 0$ disables geometric locality and allows $TopK$ neighbors to be selected from the entire slice, effectively introducing long-range connections. This leads to performance degradation, indicating that spatial constraints are necessary to preserve meaningful local structure. Increasing the intra-slice connectivity (K from 4 to 6) consistently improves results, suggesting that moderately denser local aggregation better captures regional interactions. For CSG, reducing the cross-slice radius (r_2) slightly weakens

performance, implying that sufficient spatial tolerance is required for stable volumetric propagation. Overall, r_1=2, K=6, r_2=6 provides the best balance in our setting.

Ablation study on graph model order: We also investigate the effect of module ordering. As shown in Table 2(c), reversing the default ISG→CSG sequence leads to substantial degradation across all metrics. This empirically suggests that establishing reliable intra-slice structure before aggregating cross-slice information provides a more stable basis for volumetric reasoning, while the inverse ordering injects noise into subsequent fusion. These results support ISG→CSG as the preferred configuration.

4 Conclusion

We propose GLSA-RG, a global–local structure-aware alignment framework for LGE report generation. By coupling structural graph representations with multi-instance contrastive alignment, the method enforces fine-grained structural consistency while preserving global semantic coherence. Experiments on two datasets validate its effectiveness and robustness.

Acknowledgments. This work was supported by the National Key R&D Program of China (No. 2025YFC3508704), and the Beijing Natural Science Foundation Joint Fund Key Project (L251020, L256012). **Disclosure of Interest.** The authors have no competing interests.

References

1. Gorkem Can Ates, Yu Xin, Kuang Gong, and Wei Shao. Dcformer: Efficient 3d vision-language modeling with decomposed convolutions. *arXiv preprint arXiv:2502.05091*, 2025.
2. 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.
3. Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72, 2005.
4. Manuel D. Cerqueira, Neil Weissman, Vasken Dilsizian, Alice K. Jacobs, Sanjiv Kaul, Warren K. Laskey, Dudley John Pennell, John A. Rumberger, Thomas J. Ryan, and Mario S. Verani. Standardized myocardial segmentation and nomenclature for tomographic imaging of the heart. a statement for healthcare professionals from the cardiac imaging committee of the council on clinical cardiology of the american heart association. *Circulation*, 105 4:539–42, 2002.
5. Yifan Gao, Shaohao Rui, Haoyang Su, Jinyi Xiang, Lianming Wu, and Xiaosong Wang. A composite alignment-aware framework for myocardial lesion segmentation in multi-sequence cmr images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 3–13. Springer, 2025.

6. Yifan Gao, Wei Xia, Dingdu Hu, Wenkui Wang, and Xin Gao. Desam: Decoupled segment anything model for generalizable medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 509–519. Springer, 2024.
7. Fumin Guo, Philippa RP Krahn, Terenz Escartin, Idan Roifman, and Graham Wright. Cine and late gadolinium enhancement mri registration and automated myocardial infarct heterogeneity quantification. *Magnetic Resonance in Medicine*, 85(5):2842–2855, 2021.
8. Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022.
9. Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. Clinicalbert: Modeling clinical notes and predicting hospital readmission. *arXiv preprint arXiv:1904.05342*, 2019.
10. Maximilian Ilse, Jakub Tomczak, and Max Welling. Attention-based deep multiple instance learning. In *International conference on machine learning*, pages 2127–2136. PMLR, 2018.
11. Jiayu Lei, Xiaoman Zhang, Chaoyi Wu, Lisong Dai, Ya Zhang, Yanyong Zhang, Yanfeng Wang, Weidi Xie, and Yuehua Li. Autorg-brain: Grounded report generation for brain mri. *arXiv preprint arXiv:2407.16684*, 2024.
12. Christoph Leiter, Piyawat Lertvittayakumjorn, Marina Fomicheva, Wei Zhao, Yang Gao, and Steffen Eger. Towards explainable evaluation metrics for natural language generation. *arXiv preprint arXiv:2203.11131*, 2022.
13. Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36:28541–28564, 2023.
14. Yuci Liang, Xinheng Lyu, Wenting Chen, Meidan Ding, Jipeng Zhang, Xiangjian He, Song Wu, Xiaohan Xing, Sen Yang, Xiyue Wang, et al. Wsi-llava: A multimodal large language model for whole slide image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22718–22727, 2025.
15. Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004.
16. Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.
17. Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002.
18. Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.
19. Yanzhao Shi, Xiaodan Zhang, Junzhong Ji, Haoning Jiang, Chengxin Zheng, Yinong Wang, and Liangqiong Qu. Hsenet: Hybrid spatial encoding network for 3d medical vision-language understanding. *arXiv preprint arXiv:2506.09634*, 2025.
20. Yiming Shi, Shaoshuai Yang, Xun Zhu, Haoyu Wang, Xiangling Fu, Miao Li, and Ji Wu. Medm-vl: What makes a good medical lvlm? In *International Workshop on Agentic AI for Medicine*, pages 290–299. Springer, 2025.
21. Yiming Shi, Xun Zhu, Kaiwen Wang, Ying Hu, Chenyi Guo, Miao Li, and Ji Wu. Med-2e3: A 2d-enhanced 3d medical multimodal large language model. In *2025*

IEEE International Conference on Bioinformatics and Biomedicine (BIBM), pages 2754–2759. IEEE, 2025.

22. Omkar Chakradhar Thawakar, Abdelrahman M Shaker, Sahal Shaji Mullappilly, Hisham Cholakkal, Rao Muhammad Anwer, Salman Khan, Jorma Laaksonen, and Fahad Khan. Xraygpt: Chest radiographs summarization using large medical vision-language models. In *Proceedings of the 23rd workshop on biomedical natural language processing*, pages 440–448, 2024.
23. Zifeng Wang, Zhenbang Wu, Dinesh Agarwal, and Jimeng Sun. Medclip: Contrastive learning from unpaired medical images and text. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, volume 2022, page 3876, 2022.
24. Yu Xin, Gorkem Can Ates, Kuang Gong, and Wei Shao. Med3dvlm: An efficient vision-language model for 3d medical image analysis. *IEEE Journal of Biomedical and Health Informatics*, 2025.
25. Qilong Xing, Zikai Song, Youjia Zhang, Na Feng, Junqing Yu, and Wei Yang. Mcarg: Enhancing llms with medical concept alignment for radiology report generation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 380–390. Springer, 2025.
26. Qwen An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxin Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yi-Chao Zhang, Yunyang Wan, Yuqi Liu, Zeyu Cui, Zhenru Zhang, Zihan Qiu, Shanghaoran Quan, and Zekun Wang. Qwen2.5 technical report. *ArXiv*, abs/2412.15115, 2024.
27. Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*, 2019.