Enhanced Contrastive Learning with Multi-view Longitudinal Data for Chest X-ray Report Generation

Kang Liu[1], Zhuoqi Ma[1*], Xiaolu Kang[1], Yunan Li[1], Kun Xie[1], Zhicheng Jiao[2], Qiguang Miao[1]

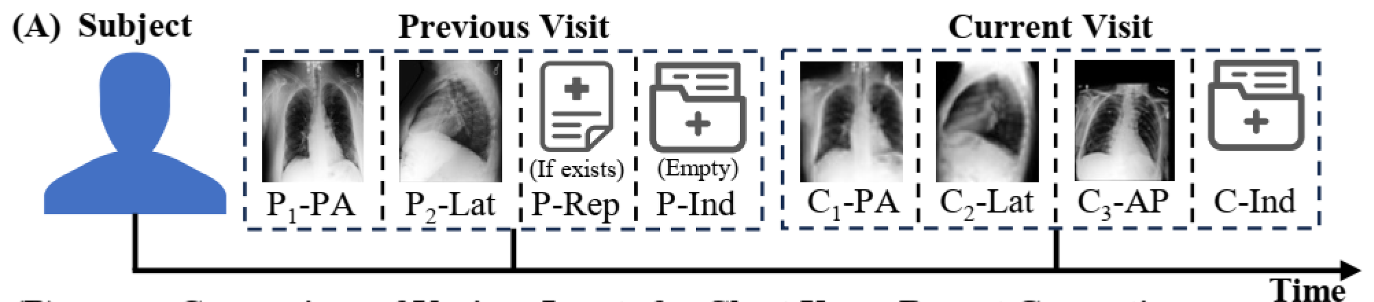
[1]Xidian University, China [2]Brown University, USA

CVPR Nashville JUNE 11-15, 2025

Introduction

Motivation

- In clinical practice, radiologists integrate **patients' symptoms with current multi-view imaging studies** to assess disease status. Simultaneously, they may consider **medical history to evaluate disease progression**, enhancing diagnostic accuracy.
- Most methods rely on **single-view images**, limiting diagnostic accuracy and ignoring disease progression. Even with longitudinal data, current visits are still analyzed using single images.



(B) Comparison of Various Inputs for Chest X-ray Report Generation

Previous Visits			Current Visit				Input
P_1-PA	P_2-Lat	P-Rep	C_1-PA	C_2-Lat	C_3-AP	C-Ind	
✗	✗	✗	✓	✗	✗	✗	Single Images
✗	✗	✗	✓	✓	✗	✗	Dual Views
✓	✗	✗	✓	✗	✗	✗	Longitudinal Data
✓	✗	✓	✓	✓	✓	✓	MVL Data (Ours)

Figure 1. (A) shows medical historical data of a subject (patient) over time. (B) compares inputs for RRG, with *AP* and *PA* as frontal views, and *Lat* and *Rep* as a lateral view and its report. *Ind* and *MVL Data* are "INDICATION" and multi-view longitudinal data.

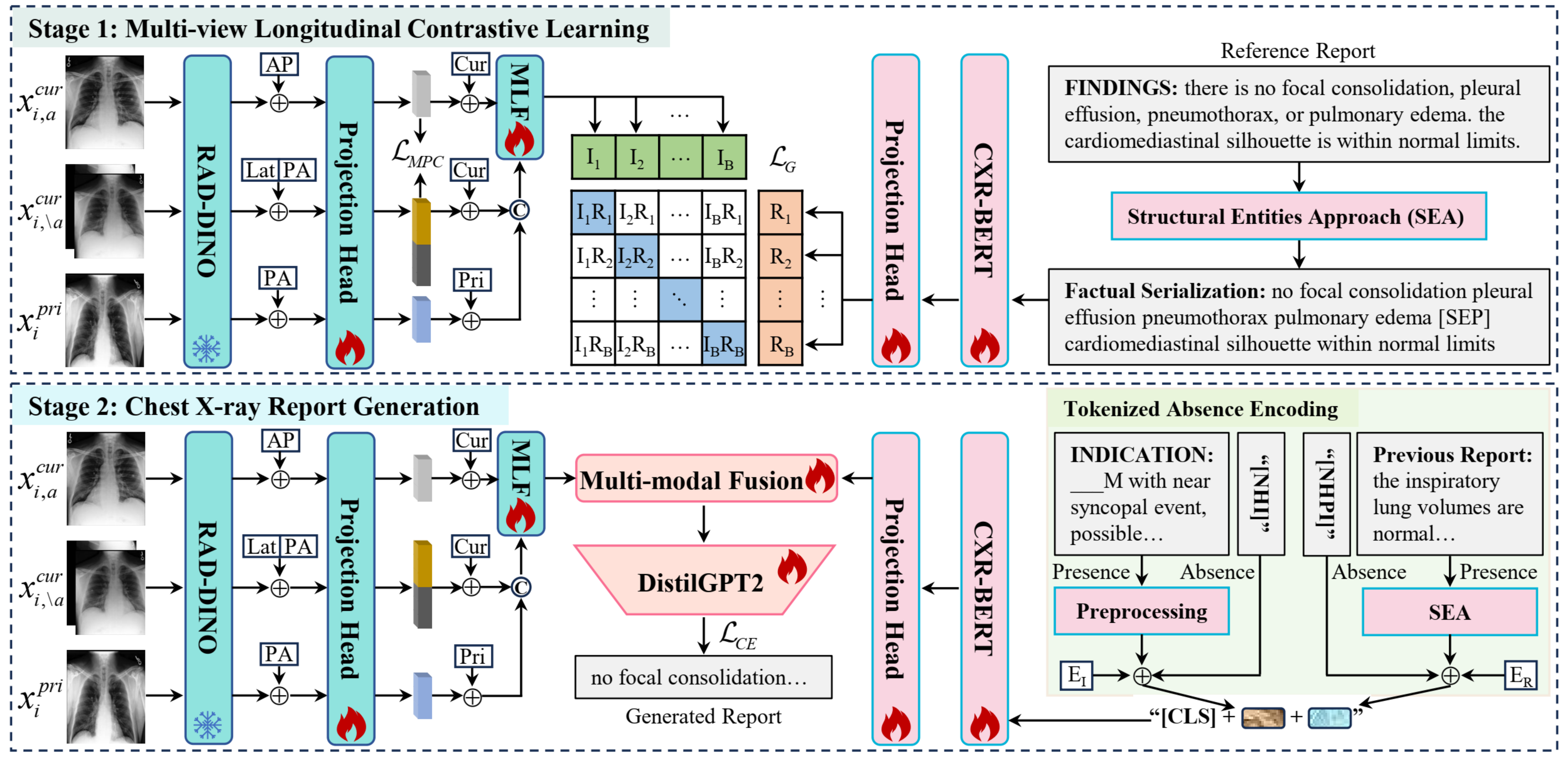
Code

WeChat

Contribution

- We propose a novel multi-view longitudinal contrastive learning method that **flexibly integrates multi-view longitudinal (MVL) data** and leverages inherent spatiotemporal information from reports to supervise the pre-training of visual and textual representations.
- We introduce a **tokenized absence encoding technique** to handle missing patient-specific prior knowledge, enabling flexible adaptation to scenarios with or without such data.

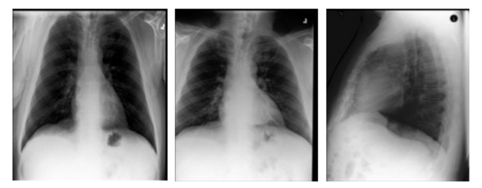
Overview



$x^{cur}_{i,a}$ and $x^{cur}_{i,\backslash a}$: Anchor scan and auxiliary references (if available) from the current visit. x^{pri}_i: Previous image (if exists). 🔥: Trainable parameters.

AP, PA, and Lat: View positional embeddings. Cur and Pri: Temporal positional embeddings. E_I and E_R: Segment embeddings. ❄: Frozen parameters.

MLF: Multi-view longitudinal fusion network. [CLS], [SEP], [NHI], and [NHPR]: Special tokens. ⊕: Summation. ©: Concatenation.

 B-4↑: **0.608**/0.000; RG↑: **0.838**/0.323; Ind ✓; PI ✓; PR ×	**Reference report:** the lungs are well expanded and clear. cardiomediastinal and hilar contours are unremarkable. there is no pleural effusion or pneumothorax. sternotomy wires are again noted, with fracture of the two upper wires unchanged from prior exam.	**Generated report by SEI:** frontal and lateral views of the chest were obtained . the patient is status post median sternotomy and cabg. no focal consolidation, pleural effusion, or evidence of pneumothorax is seen. the cardiac and mediastinal silhouettes are unremarkable.	**Generated report by MLRG (Ours):** the lungs are well expanded and clear. cardiomediastinal and hilar contours are unremarkable. there is no pleural effusion or pneumothorax. sternotomy wires are intact.
 B-4↑: **0.804**/0.288; RG↑: **0.769**/0.258; Ind ✓; PI ✓; PR ×	**Reference report:** as compared to the previous radiograph, the patient has received a new right internal jugular vein catheter. the course of the catheter is unremarkable, the tip of the catheter projects over the lower svc. there is no evidence of complications, notably no pneumothorax. otherwise unchanged radiographic appearance.	**Generated report by SEI:** as compared to the previous radiograph, the patient has been extubated and the nasogastric tube has been removed. the right internal jugular vein catheter and the right internal jugular vein catheter are in unchanged position. the lung volumes have slightly decreased . there is no evidence of pneumothorax. the lung volumes remain low . moderate cardiomegaly persists.	**Generated report by MLRG (Ours):** as compared to the previous radiograph, **the patient has received a right internal jugular vein catheter**. the course of the catheter is unremarkable, the tip of the catheter projects over the mid svc. there is no evidence of complications, notably no pneumothorax. otherwise, the radiograph is unchanged.

Summary

- MLRG captures inter-view differences and extracts spatial features from current multi-view images and temporal features from longitudinal data, effectively leveraging report-based spatiotemporal cues for pre-training. We further introduce a tokenized absence encoding technique to handle missing patient-specific prior knowledge.
- MLRG achieves a +2.3% BLEU-4 on MIMIC-CXR, +5.5% F1-CheXbert on MIMIC-ABN, and +2.7% F1-RadGraph on Two-view CXR).
- Email: kangliu@stu.xidian.edu.cn (Kang Liu), zhuoqima@xidian.edu.cn (Zhuoqi Ma)
- Code: https://github.com/mk-runner/MLRG

Experiments

Dataset	Input	Method	Venue	NLG Metrics ↑						CE Metrics ↑			
				B-1	B-2	B-3	B-4	MTR	R-L	RG	P	R	F1
M-CXR	SI	SA [56]	EMNLP'23	-	0.184	-	-	-	-	0.228	-	-	0.394
	SI	MET [53]	CVPR'23	0.386	0.250	0.169	0.124	0.152	0.291	-	0.364	0.309	0.311
	SI	KiUT [17]	CVPR'23	0.393	0.243	0.159	0.113	0.160	0.285	-	0.371	0.318	0.321
	SI	CoFE [26]	ECCV'24	-	-	-	0.125	**0.176**	<u>0.304</u>		0.489	0.370	0.405
	SI	MAN [42]	AAAI'24	0.396	0.244	0.162	0.115	0.151	0.274	-	0.411	0.398	0.389
	SI	B-LLM [28]	AAAI'24	<u>0.402</u>	<u>0.262</u>	<u>0.180</u>	0.128	<u>0.175</u>	0.291	-	0.465	**0.482**	<u>0.473</u>
	SI	DCG [27]	ACMMM'24	0.397	0.258	0.166	0.126	0.162	0.295	-	0.441	0.414	0.404
	SI	Med-LLM [31]	ACMMM'24	-	-	-	0.128	0.161	0.289	-	0.412	0.373	0.395
	SI+Ind	SEI [29]	MICCAI'24	0.382	0.247	0.177	<u>0.135</u>	0.158	0.299	<u>0.249</u>	<u>0.523</u>	0.410	0.460
	MVD	FMVP [32]	TMM'23	0.389	0.236	0.156	0.108	0.150	0.284	-	0.332	0.383	0.336
	Long	HERGen [49]	ECCV'24	0.395	0.248	0.169	0.122	0.156	0.285	-	-	-	-
	MVL	CXRMate [36]	arXiv'23	0.361	0.223	0.150	0.108	0.159	0.263	0.238	0.495	0.367	0.422
	MVL	**MLRG(Ours)**	-	**0.411**	**0.277**	**0.204**	**0.158**	**0.176**	**0.320**	**0.291**	**0.549**	<u>0.468</u>	**0.505**
	-	Δ (%) ↑	-	+0.9	+1.5	+2.4	+2.3	+0.1	+1.6	+4.2	+2.6	-1.4	+3.2
M-ABN	SI	R2Gen♭ [8]	EMNLP'20	0.253	0.144	0.092	0.063	0.106	0.229	0.179	0.444	0.425	0.434
	SI	CMN♭ [9]	ACL'21	0.256	0.147	0.095	0.066	0.110	0.230	0.183	<u>0.466</u>	<u>0.454</u>	<u>0.460</u>
	SI+Ind	SEI♭ [29]	MICCAI'24	<u>0.267</u>	<u>0.157</u>	<u>0.104</u>	<u>0.073</u>	<u>0.114</u>	<u>0.231</u>	<u>0.191</u>	<u>0.466</u>	0.408	0.435
	MVL	**MLRG(Ours)**	-	**0.332**	**0.199**	**0.132**	**0.094**	**0.136**	**0.248**	**0.219**	**0.513**	**0.517**	**0.515**
	-	Δ (%) ↑	-	+6.5	+4.2	+2.8	+2.1	+2.2	+1.7	+2.8	+4.7	+6.3	+5.5
T-CXR	DV	R2Gen♭ [8]	EMNLP'20	0.346	0.219	0.153	0.113	0.141	0.302	0.267	0.478	0.329	0.390
	DV	CMN♭ [9]	ACL'21	0.387	0.241	0.166	0.122	0.151	0.310	0.268	0.496	0.336	0.401
	DV+Ind	SEI♭ [29]	MICCAI'24	<u>0.409</u>	<u>0.263</u>	<u>0.186</u>	<u>0.140</u>	<u>0.168</u>	<u>0.320</u>	<u>0.301</u>	<u>0.522</u>	<u>0.447</u>	<u>0.481</u>
	MVL	**MLRG(Ours)**	-	**0.417**	**0.276**	**0.200**	**0.154**	**0.178**	**0.331**	**0.328**	**0.532**	**0.474**	**0.501**
	-	Δ (%) ↑	-	+0.8	+1.3	+1.4	+1.4	+1.0	+1.1	+2.7	+1.0	+2.7	+2.0

Model	M/S	F/R	PI	Stage 1		Stage 2		NLG metrics ↑						CE metrics ↑	
				$\mathcal{L}_G$	$\mathcal{L}_{MPC}$	Ind	PR	B-1	B-2	B-3	B-4	MTR	R-L	RG	F1
(a)	M	F	✓	✗	✗	✓	✓	0.346	0.235	0.173	0.136	0.154	0.305	0.258	0.373
(b)	M	F	✓	✓	✓	✗	✗	0.385	0.239	0.162	0.118	0.155	0.283	0.238	0.479
(c)	M	F	✗	✓	✗	✓	✓	0.384	0.257	0.188	0.146	0.165	0.310	0.267	0.455
(d)	M	F	✗	✓	✓	✓	✓	0.395	0.257	0.183	0.138	0.167	0.302	0.278	0.503
(e)	M	F	✓	✓	✓	✓	✗	0.392	0.265	0.195	0.153	0.171	0.316	0.281	0.476
(f)	M	F	✓	✓	✓	✗	✓	0.387	0.240	0.163	0.118	0.156	0.281	0.243	0.484
(g)	M	R	✓	✓	✓	✓	✓	0.403	0.267	0.193	0.148	0.172	0.309	0.287	**0.510**
(h)	S	F	✓	✓	✗	✓	✓	0.400	0.269	0.196	0.151	0.173	0.314	0.289	0.498
MLRG	M	F	✓	✓	✓	✓	✓	**0.411**	**0.277**	**0.204**	**0.158**	**0.176**	**0.320**	**0.291**	0.505

Table 3. Ablation study on the MIMIC-CXR dataset. "M/S" refers to methods that utilize current **M**ulti-view images or **S**ingle images as input. "F/R" indicates alignment based on either **F**actual serialization or **R**eport. "PI", "PR", and "Ind" represent **P**revious **I**mages, **P**revious **R**eports, and "INDICATION", respectively. The best values are emphasized in **bold**.

Setting	%	B-2 ↑	B-4 ↑	MTR ↑	R-L ↑	RG ↑
w/ Ind	57.8	**0.295**	**0.174**	**0.184**	**0.332**	**0.318**
w/o Ind	42.2	0.253	0.137	0.166	0.302	0.254
w/ MV	70.7	**0.282**	**0.161**	**0.179**	**0.323**	**0.301**
w/o MV	29.3	0.264	0.150	0.171	0.310	0.266
w/ MVL	61.4	**0.282**	**0.160**	**0.178**	**0.322**	**0.300**
w/o MVL	38.6	0.270	0.155	0.174	0.316	0.276

Table 4. Breakdown of MLRG's metrics on the MIMIC-CXR test set, categorized by (a) inclusion of indications (Ind), (b) inclusion of current multi-view images (MV), (c) inclusion of multi-view longitudinal data (MVL).

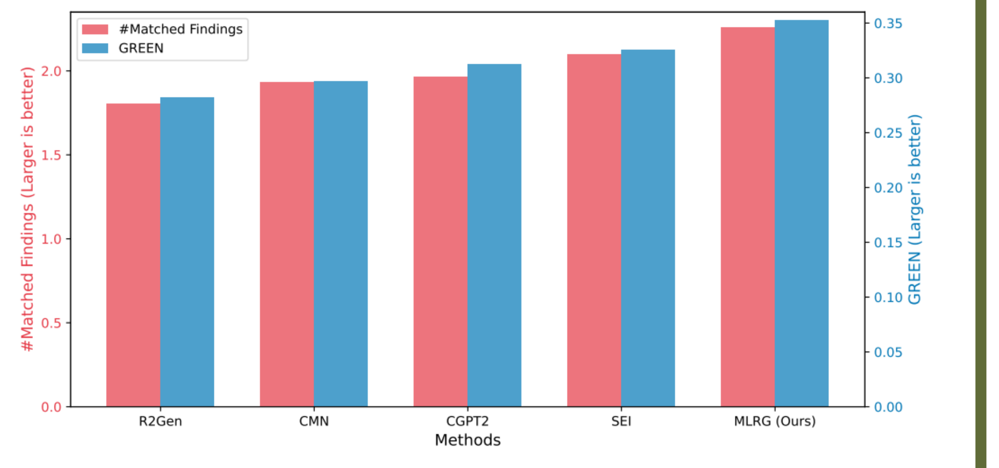


Figure 5. Comparison with baselines on MIMIC-CXR using LLMs. "#Matched Findings" denotes the number of matched findings between generated and reference reports.