Sparse-Gated Alignment in Vision–Language Pretraining Improves Zero-Shot Classification and Localization of Chest Findings on Radiographs and CT

Seongeun Lee, Junhyun Park, Hannah Yun, Hyunwoong Kim, Jonggwon Park

- DEEPNOID Inc., Korea, Republic of

jgpark@deepnoid.com

PURPOSE :

Trustworthy clinical use of AI for chest imaging requires not only accurate finding detection but also region-level evidence indicating the image region supporting each finding — a capability that has been difficult to achieve from report-level supervision without manual pixel- or voxel-level annotation. To address this, we propose a vision–language pretraining framework with sparse-gated alignment that jointly enables zero-shot classification and finding-specific spatial localization of chest findings on radiographs and CT, learned from radiology reports.

METERIALS AND METHOD :

We developed a vision–language pretraining framework with sparse-gated alignment, comprising (i) a sparse patch-level gate that concentrates image–text alignment on regions relevant to each report sentence and (ii) a feature-level regularization that anchors intermediate patch representations to a frozen self-supervised encoder, preserving fine-grained spatial features and suppressing representational drift in irrelevant regions. Pretraining used 353,530 chest radiographs (MIMIC-CXR) and 43,383 chest CT volumes (CT-RATE), each paired with the report; reports were decomposed into finding sentences by a large language model. No pixel- or voxel-level annotations or localization labels were used. Zero-shot evaluation covered 8 chest radiograph tasks (4 classification, 2 grounding, 2 segmentation) and 3 chest CT tasks (2 classification, 1 free-text volumetric segmentation), compared with prior models.

RESULTS :

The proposed model outperformed prior radiology vision–language models on every comparable task. On chest radiographs, classification macro AUROC reached 0.825–0.917 across four datasets, exceeding prior best on each. Grounding accuracy reached 0.892 on MS-CXR (prior best, 0.844) and 0.723 on ChestX-Det10 (prior best, 0.622); segmentation Dice reached 0.317 on SIIM pneumothorax (prior best, 0.171) and 0.434 on ChestX-Det (prior best, 0.332). On chest CT, classification AUROC reached 0.850 on CT-RATE (prior best, 0.790) and 0.789 on Rad-ChestCT (prior best, 0.730). Free-text descriptions

additionally produced anatomically specific volumetric maps without mask supervision (Dice 0.128 on ReXGroundingCT), the first reported zero-shot result on this
task.

Table. Zero-shot performance on chest radiograph and chest CT tasks. Best per-task zero-shot results are in **bold**.

(A) Chest radiograph

Method	Classification				Grounding		Segmentation	
	PadChest	PadChest20†	ChestXray14	Open-I	ChestX-Det10	MS-CXR	ChestX-Det	SIIM
	Macro AUROC				Acc.	Acc.	Macro Dice	Dice
GLoRIA	0.565	0.558	0.610	0.589	0.367	—	—	—
KAD	0.750	0.735	0.789	0.807	0.391	—	—	—
MedKLIP	0.629	0.688	0.726	0.759	0.481	0.407	0.249	0.044
CARZero	0.810	0.837	0.811	0.838	0.543	0.749	0.304	0.081
MedSigLIP	0.736	0.747	0.742	0.763	—	—	—	—
RadZero	0.841	0.871	0.804	0.847	0.622	0.844	0.332	0.171
Ours	**0.863**	**0.917**	**0.825**	**0.881**	**0.723**	**0.892**	**0.434**	**0.317**

† PadChest20: tail-20 rare-class subset of PadChest.

(B) Chest CT

Method	Classification	
	CT-RATE	Rad-ChestCT
	Macro AUROC	
CT-CLIP	0.731	0.629
Merlin	0.728	0.644
fVLM	0.778	0.680
ViSD-Boost	0.790	0.694
RadZero3D	0.762	0.690
COLIPRI	0.785	0.730
Ours	**0.850**	**0.789**

Method	Setting	Free-text Segmentation
		ReXGroundingCT
		Dice
SAT	Supervised	0.131
VoxTell	Supervised	0.282
Ours	Zero-shot	**0.128**

CONCLUSION :

Sparse-gated alignment in vision–language pretraining enabled joint zero-shot classification and finding-specific localization on chest radiographs and CT from radiology reports alone, providing region-level evidence for each finding without dense manual annotation.