TraceCXR: Verifiable Chest X-ray Reports via Evidence-Addressable Graph Prompting

Hang Dong[1†], Chao Sun[2,3†], Hao Yan[2], Wei Hu[4*], and Bo Du[5*]

[1] Academy of Advanced Interdisciplinary Studies, Wuhan University, Wuhan, China
[2] School of Computer Science, Wuhan University, Wuhan, China
[3] Institute of Artificial Intelligence, School of Computer Science, Wuhan University, Wuhan, China
[4] Department of Ultrasound, Renmin Hospital East Branch of Wuhan University, Wuhan, China
[5] National Engineering Research Center for Multimedia Software, Wuhan University, Wuhan, China
dubo@whu.edu.cn †Equal contribution; *Corresponding authors.

Abstract. Automatic chest X-ray report generation is increasingly used to assist clinical reading, yet vision–language models may hallucinate findings or miss subtle, clinically meaningful cues. It remains difficult to inspect which localized radiographic evidence supports a generated clinical statement, making outputs hard to verify and audit in practice. We propose TraceCXR, an evidence-traceable report generation framework that attaches clinically relevant generated statements to an inspectable evidence trail consisting of a concept-level node trace and localized visual evidence maps. TraceCXR builds ProtoGraph, which turns clinical concepts into evidence-addressable nodes by associating them with disease-aware visual prototypes, enabling retrieved concepts to be linked back to candidate image evidence. Building on ProtoGraph, EviLoop Prompting retrieves patient-specific graph context from the input image and grounds the retrieved cues back to localized visual evidence; confidence-gated residual injection suppresses unreliable graph prompts when retrieval confidence is low. We further introduce Process-Supervised Node Planning, a lightweight training-time regularizer for clinically relevant node selection and concentrated evidence attribution. Experiments on IU X-ray, MIMIC-CXR, and CheXpert Plus demonstrate improved report quality and clinical correctness, while evidence-centric proxy evaluations indicate improved grounding, retrieval, and traceability.

Keywords: Radiology report generation · Vision-language semantic alignment · Knowledge graphs.

1 Introduction

Faithful medical report generation (MRG) aims to translate chest radiographs into radiologist-style narratives, such as *Findings* and *Impression*, supporting scalable and consistent clinical documentation [4]. Unlike natural-image captioning, MRG must capture subtle abnormalities and clinically meaningful changes under long-tailed disease distributions and dataset bias [24]. Most modern systems learn image–text alignment through attention-based encoder–decoder architectures. However, attention weights alone are insufficient as clinical evidence: they may over-focus on visually salient but clinically irrelevant regions, overlook subtle cues, and produce statements that are difficult to inspect or audit.

Existing MRG methods can be broadly grouped into two lines. The first improves vision–language representations and generation fluency through contrastive pretraining and LLM-based decoders [27, 17], but may still generate unsupported findings under unreliable visual–textual alignment. The second introduces clinical structures or grounding signals, such as knowledge graphs and anatomy- or region-aware conditioning, to guide content selection [35, 21]. However, these methods often lack a closed evidence loop that retrieves patient-specific clinical context and grounds it to localized image evidence, especially under noisy reports and long-tailed observations. This gap leads to three deployment challenges: (1) reports may contain unsupported statements when not tied to radiographic evidence; (2) traceability remains limited because attention maps are insufficient as clinical evidence, while detector- or segmenter-based grounding can be brittle and costly [21]; and (3) retrieved concepts may be noisy or image-misaligned, making knowledge usage difficult to audit [9].

To address these issues, we propose **TraceCXR**, an evidence-traceable report generation framework built around a structured retrieval–grounding loop. In TraceCXR, an evidence trail links a generated clinical statement to retrieved graph nodes, associated visual prototypes, and localized image evidence, enabling concept- and evidence-level inspection. TraceCXR builds ProtoGraph to make clinical concepts evidence-addressable through disease-aware visual prototypes; uses EviLoop Prompting to retrieve image-conditioned graph context and ground it to localized evidence with confidence-gated residual injection; and introduces Process-Supervised Node Planning (PSNP) to regularize node selection and evidence concentration. We further use symmetric image–text contrastive learning with negation-aware hard negatives to strengthen radiograph–report alignment. Our contributions are:

- ProtoGraph links canonical clinical concepts with disease-aware visual prototypes, enabling image-conditioned concept retrieval and prototype-mediated grounding.
- EviLoop Prompting performs visual–graph retrieval, prototype-guided grounding, and confidence-gated prompt injection for LLM-based report generation.
- PSNP provides lightweight process supervision for clinically relevant node selection and concentrated evidence attribution.
- Experiments on report generation and evidence-centric transfer show improved report quality, clinical correctness, and traceability-oriented metrics.

2 Method

TraceCXR generates CXR reports with an inspectable evidence trail linking retrieved clinical concepts to localized image evidence. It contains ProtoGraph,

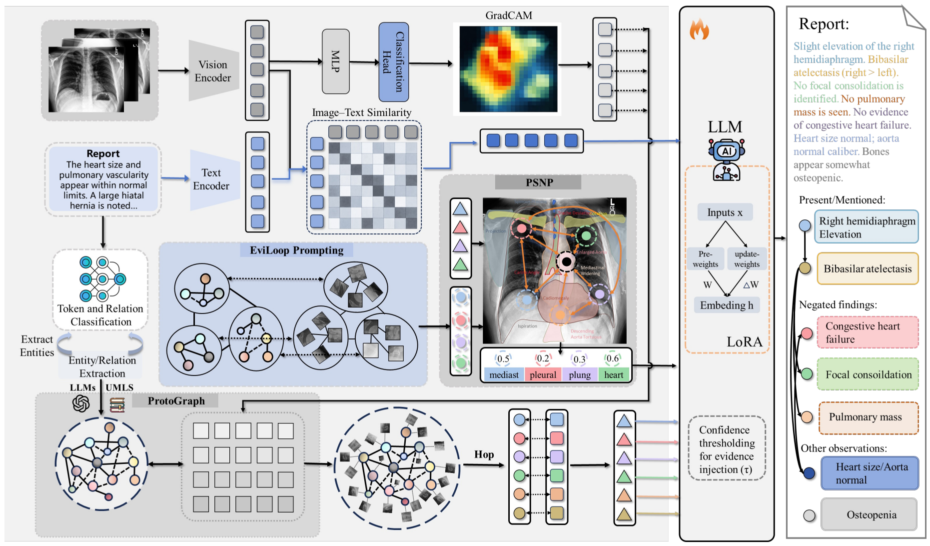

Fig. 1. Overview of TraceCXR. ProtoGraph links clinical concepts to disease-aware prototypes. EviLoop retrieves and grounds graph context through prototypes, then injects gated graph/evidence tokens. PSNP supervises compact node traces.

EviLoop, and PSNP, which respectively build evidence-addressable concepts, perform retrieval–grounding, and regularize node/evidence selection. Let f_I and f_T be the vision and text encoders. For image X, $f_I(X)$ outputs visual tokens $\mathbf{V} \in \mathbb{R}^{n \times d}$ and pooled embedding $\mathbf{z}_I$; for report T, $f_T(T)$ outputs $\mathbf{z}_T$. We use $\tilde{\mathbf{x}} = \mathbf{x}/\|\mathbf{x}\|_2$ for normalization. The clinical graph is $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where nodes are canonical clinical concepts and edges are typed report-derived relations.

2.1 Radiograph–Report Alignment

We learn cross-modal representations using symmetric image–text contrastive learning. Negation-aware hard negatives, i.e., reports or concepts with similar entities but opposite polarity, are included in each mini-batch. For paired samples $\mathcal{B} = \{(X_i, T_i)\}_{i=1}^{B}$, we compute the image-to-text similarity matrix $\mathbf{S}$ and use $\mathbf{S}^\top$ for text-to-image alignment:

$$\mathbf{S}_{ij} = \frac{\tilde{\mathbf{v}}_i^\top \tilde{\mathbf{t}}_j}{\tau}, \quad \mathcal{L}_{\text{align}} = \frac{\mathrm{CE}(\mathbf{S}, \mathbf{y}) + \mathrm{CE}(\mathbf{S}^\top, \mathbf{y})}{2}, \quad \mathbf{y}_i = i. \tag{1}$$

2.2 ProtoGraph: Making Clinical Concepts Evidence-Addressable

ProtoGraph converts report-derived concepts into evidence-addressable nodes by associating each canonical concept with disease-aware visual prototypes.

Clinical graph construction. We parse reports with RadGraph to obtain entities and typed relations. Mentions are normalized to canonical concept nodes that store concept name t_i, entity type, polarity, and anatomical modifier when available; relations such as *located-at* and *modify* are retained as graph edges. With $\mathbf{e}_i = \mathrm{TextEnc}(t_i)$, RGCN gives

$$\mathbf{h}_i = \mathrm{RGCN}_\theta\big(\mathcal{G}, \{\mathbf{e}_j\}_{v_j \in \mathcal{V}}\big)_i \in \mathbb{R}^d. \tag{2}$$

Prototype bank and sparse index. Disease-aware prototypes are mined from weakly supervised high-response patches and clustered into $\mathcal{B} = \{\mathbf{b}_m\}_{m=1}^M$. We denote the normalized projections of node and prototype features as $\tilde{\mathbf{c}}_i = \ell_2(W_g \mathbf{h}_i)$ and $\tilde{\mathbf{b}}_m = \ell_2(W_p \mathbf{b}_m)$. Each node stores a top-K_p prototype index:

$$\mathcal{I}_i^p, \boldsymbol{\omega}_i^p = \mathrm{TopK}_{K_p}\left(\{\tilde{\mathbf{c}}_i^\top \tilde{\mathbf{b}}_m\}_{m=1}^M\right), \quad \mathcal{B}_{sub} = \bigcup_{v_i \in \mathcal{V}_{sub}} \{\mathbf{b}_m \mid m \in \mathcal{I}_i^p\}. \tag{3}$$

Here $\mathcal{B}_{sub}$ is gathered by retrieved nodes during EviLoop, linking graph retrieval to visual grounding.

2.3 EviLoop Prompting: Bidirectional Evidence Loop for LLMs

EviLoop performs image-conditioned graph retrieval, graph prompt summarization, prototype-mediated grounding, and confidence-gated injection.

Visual→Graph retrieval. Given normalized image embedding $\tilde{\mathbf{v}}$ and node embeddings $\tilde{\mathbf{h}}_i = \mathbf{h}_i / \|\mathbf{h}_i\|_2$, we retrieve and expand nodes by

$$s_i = \frac{\tilde{\mathbf{v}}^\top \tilde{\mathbf{h}}_i}{\tau_g}, \quad \mathcal{V}_{top} = \mathrm{TopK}_{K_0}(\{s_i\}_{v_i \in \mathcal{V}}), \quad \mathcal{V}_{sub} = \mathrm{Expand}(\mathcal{V}_{top}, \mathcal{E}; h), \tag{4}$$

$$\alpha_{k,i} = \mathrm{softmax}_{v_i \in \mathcal{V}_{sub}}(\mathbf{q}_k^\top \mathbf{h}_i / \sqrt{d}), \quad \mathbf{G}_{ctx}[k] = \sum_{v_i \in \mathcal{V}_{sub}} \alpha_{k,i} \mathbf{h}_i. \tag{5}$$

Here K_0 is the initial node budget, h is the graph expansion hop number, and K_g learnable queries $\{\mathbf{q}_k\}$ summarize the retrieved context.

Graph→Visual grounding via prototypes. Let $\mathrm{Attn}(\cdot)$ denote scaled dot-product attention. Graph prompts first attend to the selected prototypes and are then grounded to image tokens:

$$\mathbf{E}_{proto} = \mathrm{Attn}(\mathbf{G}_{ctx} \mathbf{W}_Q^g, \mathbf{B}_{sub} \mathbf{W}_K^p, \mathbf{B}_{sub} \mathbf{W}_V^p), \tag{6}$$

$$\mathbf{A} = \mathrm{softmax}_n\left(\frac{(\mathbf{E}_{proto} \mathbf{W}_Q^e)(\mathbf{V}\mathbf{W}_K^v)^\top}{\sqrt{d}}\right) \in \mathbb{R}^{K_g \times n}. \tag{7}$$

The image-grounded evidence tokens are computed as $\mathbf{E}_{img} = \mathbf{A}(\mathbf{V}\mathbf{W}_V^v)$, and each row of $\mathbf{A}$ is a localized evidence map for one graph prompt token.

Confidence-gated residual injection. A scalar gate suppresses unreliable graph cues using the top-K_0 retrieval distribution:

$$\mathbf{p}^{top} = \mathrm{softmax}(\{s_i \mid v_i \in \mathcal{V}_{top}\}), \quad c_{ent} = 1 - H(\mathbf{p}^{top}) / \log K_0,$$

$$g = \mathbb{I}[\max(\mathbf{p}^{top}) \geq \tau_{abs}] \cdot \sigma(a c_{ent} + b), \tag{8}$$

$$\mathbf{X}_{LLM} = [\mathbf{V};\ g\mathbf{G}_{ctx};\ g\mathbf{E}_{img}]. \tag{9}$$

where $H(\mathbf{p}^{top}) = -\sum_i p_i^{top} \log(p_i^{top} + \epsilon)$ and τ_{abs} is an absolute confidence floor.

2.4 Process-Supervised Node Planning

PSNP is a training-time regularizer. For each training sample, the pseudo target node set $\mathcal{P}$ is built from affirmed RadGraph entities in the reference report after concept normalization. Negated and uncertain entities are retained as polarity attributes but are not positive retrieval targets, and reference reports are not used at inference time.
Node-level retrieval regularization. Let $\mathbf{s} \in \mathbb{R}^{|\mathcal{V}|}$ be the visual-to-graph retrieval scores in Eq. (4). Hard negatives $\mathcal{N}_h$ are high-scoring nodes outside $\mathcal{P}$:

$$\mathcal{L}_{graph} = \frac{1}{|\mathcal{P}|} \sum_{v_i \in \mathcal{P}} \log(1 + \exp(-s_i)) + \frac{1}{|\mathcal{N}_h|} \sum_{v_j \in \mathcal{N}_h} \log(1 + \exp(s_j)). \tag{10}$$

Evidence concentration. We penalize diffuse grounding by minimizing the normalized row-wise entropy of $\mathbf{A}$:

$$\mathcal{L}_{evi} = \frac{1}{K_g} \sum_{k=1}^{K_g} \left[-\frac{1}{\log n} \sum_{r=1}^{n} A_{k,r} \log(A_{k,r} + \epsilon) \right]. \tag{11}$$

2.5 Generation Fine-tuning and Objective

We fine-tune the report generator on $\mathbf{X}_{LLM}$ with

$$\begin{gathered} \mathcal{L}_{LM} = -\sum_{\ell=1}^{T} \log p_\phi(y_\ell \mid y_{<\ell}, \mathbf{X}_{LLM}), \\ \mathcal{L}_{proc} = w(r)(\lambda_{graph}\mathcal{L}_{graph} + \lambda_{evi}\mathcal{L}_{evi}), \quad \mathcal{L}_{gen} = \mathcal{L}_{LM} + \mathcal{L}_{proc}. \end{gathered} \tag{12}$$

where $w(r) = \min(1, r/R_{warm})$. The alignment and generation stages use $\mathcal{L}_{align}$ and $\mathcal{L}_{gen}$, respectively.

3 Experiments

Datasets and evaluation. We evaluate TraceCXR on report quality, entity-level clinical correctness, and evidence-centric grounding or transfer. Report generation is evaluated on IU X-ray [6], MIMIC-CXR [11], and CheXpert Plus [2]. Evidence-centric transfer uses MS-CXR [1], CheXpert 5×200 [8], SIIM [12], and NIH/RSNA [25, 20]. We report NLG metrics, RadGraph entity-level precision/recall/F1, and task-specific transfer metrics; "–" denotes unreported entries.
Clinical correctness and traceability. We extract RadGraph entities from generated and reference reports, keep affirmed non-negated entities, normalize them to the same concept vocabulary, and compute entity-level precision, recall, and F1. Trace Coverage (TC) is the fraction of affirmed generated entities whose normalized concepts appear in the retrieved node trace.
Implementation. TraceCXR is trained in two stages: radiograph–report contrastive alignment and LoRA fine-tuning of the Llama2-7B generator with EviLoop

Table 1. Performance on IU X-ray (IU), CheXpert Plus (CXP), and MIMIC-CXR (MIMIC). Best/second-best are shown in **bold**/<u>underlined</u>; "–" denotes unreported.

Dataset	Methods	Publication	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-L	METEOR	CIDEr
IU	SentSAT+KG [35]	AAAI 2020	0.441	0.291	0.203	0.147	0.367	–	–
	R2Gen [4]	EMNLP 2020	0.470	0.304	0.219	0.165	0.371	0.187	–
	AlignTrans [34]	MICCAI 2021	0.484	0.313	0.225	0.173	0.379	0.204	–
	CMCL [16]	ACL 2021	0.473	0.305	0.217	0.162	0.378	0.186	–
	PPKED [17]	CVPR 2021	0.483	0.315	0.224	0.168	0.376	0.187	0.351
	R2GenCMN [3]	ACL-IJCNLP 2021	0.475	0.309	0.222	0.170	0.375	0.191	–
	MSAT [28]	MICCAI 2022	0.481	0.316	0.226	0.171	0.372	0.190	0.394
	DCL [13]	CVPR 2023	–	–	–	0.163	**0.383**	0.193	<u>0.586</u>
	METransformer [26]	ACM MM 2023	0.483	0.322	0.228	0.172	<u>0.380</u>	0.192	0.435
	R2GenGPT [27]	Meta Radiology 2023	0.465	0.299	0.214	0.161	0.376	0.219	0.542
	PromptMRG [10]	AAAI 2024	0.401	–	–	0.098	0.160	**0.281**	–
	SILC [15]	IEEE TMI 2024	0.472	0.321	0.234	0.175	0.379	0.192	0.368
	CXPMRG-Bench [24]	CVPR 2025	<u>0.491</u>	<u>0.330</u>	**0.241**	<u>0.185</u>	0.371	0.216	0.524
	DuCo-Net [19]	IEEE Access 2025	**0.500**	<u>0.330</u>	0.220	0.160	0.260	<u>0.240</u>	–
	TraceCXR	Ours	<u>0.491</u>	**0.334**	<u>0.239</u>	**0.187**	**0.383**	0.221	**0.634**
CXP	R2Gen [4]	EMNLP 2020	0.301	0.179	0.118	0.081	0.246	0.113	0.077
	R2GenCMN [3]	ACL-IJCNLP 2021	0.321	0.195	0.128	0.087	0.256	0.127	0.102
	MSAT [28]	MICCAI 2022	–	–	–	0.036	0.156	0.066	0.018
	XProNet [22]	ECCV 2022	0.364	0.225	0.148	0.100	0.265	0.146	0.121
	ORGan [7]	ACL 2023	0.320	0.196	0.128	0.086	0.261	0.135	0.107
	R2GenGPT [27]	Meta Radiology 2023	0.361	0.224	0.149	0.101	<u>0.266</u>	0.145	0.123
	ASGMD [30]	ESWA 2024	0.267	0.149	0.094	0.063	0.220	0.094	0.044
	PromptMRG [10]	AAAI 2024	0.326	0.174	–	0.095	0.222	0.121	0.044
	R2GenCSR [23]	arXiv 2024	0.364	0.225	0.148	0.100	0.265	0.146	0.121
	Token-Mixer [33]	IEEE TMI 2024	**0.378**	<u>0.231</u>	<u>0.153</u>	0.091	0.262	0.135	0.098
	CXPMRG-Bench [24]	CVPR 2025	–	–	–	<u>0.112</u>	**0.276**	**0.157**	<u>0.139</u>
	MCA-RG [29]	MICCAI 2025	<u>0.367</u>	0.218	0.149	0.102	<u>0.266</u>	<u>0.147</u>	–
	TraceCXR	Ours	**0.378**	**0.243**	**0.162**	**0.114**	0.265	**0.157**	**0.144**
MIMIC	M2Transformer [5]	CVPR 2020	0.332	0.210	0.142	0.101	0.264	0.134	0.142
	R2Gen [4]	EMNLP 2020	0.353	0.218	0.145	0.103	0.277	0.142	–
	PPKED [17]	CVPR 2021	0.360	0.224	0.149	0.106	0.284	0.149	0.237
	R2GenCMN [3]	ACL-IJCNLP 2021	0.353	0.218	0.148	0.106	0.278	0.142	–
	WCL [31]	Findings EMNLP 2021	–	–	–	0.067	0.241	0.100	–
	GSK [32]	MedIA 2022	0.363	0.228	0.156	0.115	0.284	–	0.203
	MSAT [28]	MICCAI 2022	0.373	0.235	0.162	0.120	0.282	0.143	0.299
	METransformer [26]	CVPR 2023	0.386	0.250	0.169	0.124	0.291	0.152	**0.362**
	R2GenGPT (Deep) [27]	Meta-Radiology 2023	0.411	0.267	<u>0.186</u>	<u>0.134</u>	<u>0.297</u>	0.160	0.269
	RGRG [21]	CVPR 2023	–	–	–	0.121	0.283	0.160	–
	DCG [14]	ACM MM 2024	–	–	–	0.126	0.295	0.162	–
	CXPMRG-Bench [24]	CVPR 2025	**0.422**	<u>0.268</u>	0.184	0.133	0.289	**0.167**	0.241
	MCA-RG [29]	MICCAI 2025	0.411	0.256	0.175	0.128	**0.300**	<u>0.163</u>	–
	McDiM [18]	arXiv 2025	0.328	0.185	0.109	–	0.265	–	0.297
	TraceCXR	Ours	<u>0.412</u>	**0.269**	**0.187**	**0.136**	<u>0.297</u>	**0.167**	<u>0.324</u>

Table 2. Ablation study on CheXpert Plus. The variants evaluate how ProtoGraph, PSNP, and EviLoop affect report quality and RadGraph entity-level clinical correctness. Best and second-best results are highlighted in **bold** and <u>underlined</u>.

Dataset	ProtoGraph	PSNP	EviLoop	BLEU-1	BLEU-2	BLEU-3	BLEU-4	RG-L	METEOR	CIDEr	P	R	F1
CXP	–	–	–	0.371	0.233	0.156	0.110	0.276	0.151	0.142	0.377	0.319	0.335
	✓	–	–	0.379	0.236	0.159	0.111	0.278	0.154	<u>0.147</u>	0.385	0.322	0.337
	–	✓	–	0.376	0.239	0.160	0.114	**0.280**	0.152	**0.152**	0.382	0.325	**0.339**
	–	–	✓	<u>0.385</u>	<u>0.242</u>	<u>0.162</u>	<u>0.115</u>	<u>0.279</u>	<u>0.155</u>	0.140	<u>0.391</u>	<u>0.328</u>	<u>0.338</u>
	✓	✓	✓	**0.387**	**0.244**	**0.163**	**0.116**	**0.280**	**0.158**	<u>0.147</u>	**0.393**	**0.330**	**0.339**

prompts. We use a Mamba-based vision encoder, BioClinicalBERT text encoder, and 3-layer RGCN graph encoder. Graph prompting retrieves $K_0 = 32$ nodes, expands them by one hop to at most 64 nodes, uses $K_g = 8$ graph context tokens, and keeps $K_p = 4$ prototypes per concept.

Results. Table 1 reports generation results. TraceCXR is competitive overall, improves BLEU-4 and CIDEr on IU X-ray and CheXpert Plus, and improves BLEU scores on MIMIC-CXR. Its MIMIC-CXR CIDEr remains below

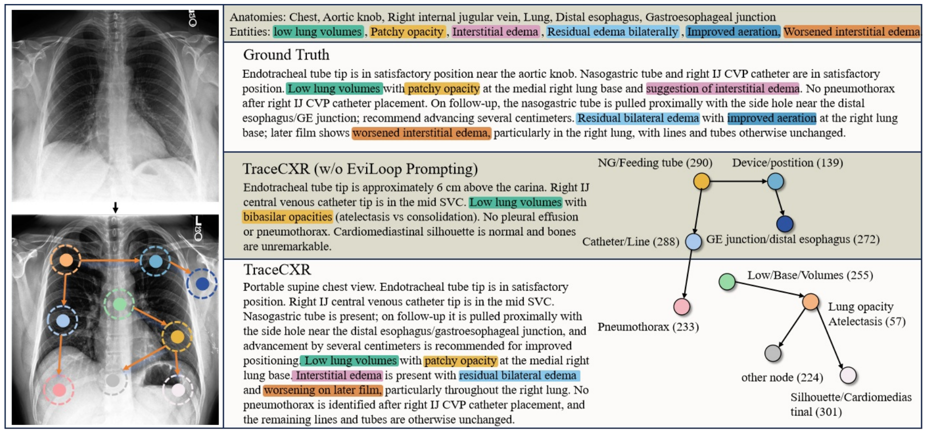

Fig. 2. Qualitative evidence tracing. Generated statements are linked to retrieved graph nodes and localized evidence maps; EviLoop improves concept–region alignment.

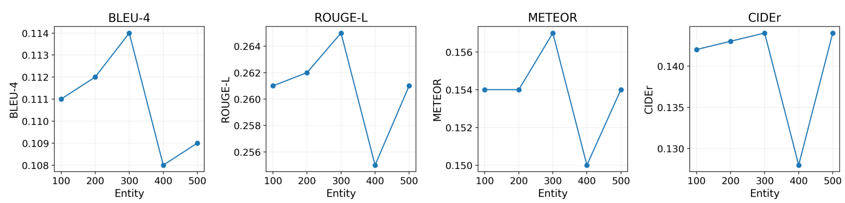

Fig. 3. Sensitivity to entity budget. Moderate budgets improve graph context, while large budgets introduce noisy nodes and reduce metrics.

the strongest captioning-oriented baseline, indicating room for better sequence-level matching.

Ablation. Table 2 shows that EviLoop gives the largest gains on most NLG and precision-related metrics. ProtoGraph and PSNP add complementary improvements by stabilizing concept retrieval and encouraging compact evidence attribution; the full model gives the best or near-best balance. Figure 3 shows that small budgets may miss concepts, whereas overly large budgets introduce low-confidence nodes and degrade metrics. Figures 4–5 compare language–clinical trade-offs and entity-level correctness.

Qualitative analysis. Figure 2 illustrates how generated statements are linked to retrieved graph nodes and localized evidence maps. The confidence gate reduces uncertain graph cues in low-confidence cases.

Case-level trace coverage (TC). TC measures whether affirmed entities in the generated report are represented in the retrieved node trace. In the shown case, 4–5 out of 5 affirmed entities are covered (TC≈0.80–1.00), and the overall non-negated TC across the test set is around 85%, indicating that most generated clinical concepts are accompanied by node-level traces.

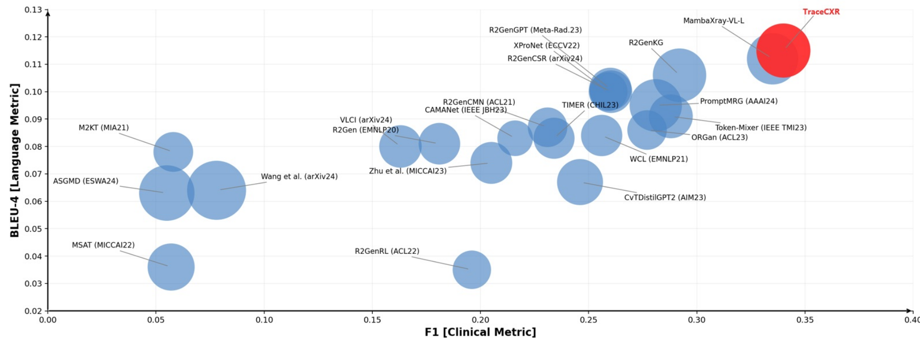

Fig. 4. Language–clinical trade-off on CheXpert Plus. Representative methods are compared by BLEU-4 and RadGraph entity-level F1.

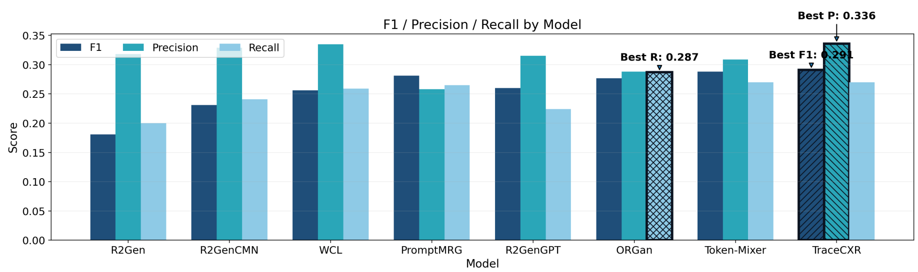

Fig. 5. Entity-level clinical correctness on CheXpert Plus. Comparison between TraceCXR and representative baselines in precision, recall, and F1.

Table 3. Evidence-centric transfer evaluation on MS-CXR phrase grounding, CheXpert 5×200 image-to-text retrieval, SIIM segmentation, and NIH/RSNA classification. Z/S denotes zero-shot evaluation, and 1%/10%/100% denote annotation ratios.

Models	Phrase Grounding		Image-to-Text Retrieval		Seg.	NIH (AUC)				RSNA (AUC)			
	CNR	mIoU	Prec@5	Prec@10	Dice	Z/S	1%	10%	100%	Z/S	1%	10%	100%
GLoRIA	0.93	**0.246**	32.6	33.4	63.4	61.0	67.1	76.6	81.3	80.4	86.1	88.0	88.6
ConVIRT	0.818	<u>0.238</u>	30.8	28.2	59.9	66.1	66.2	76.6	81.3	71.5	77.4	80.1	81.3
BioViL	<u>1.140</u>	0.215	–	–	70.0	69.1	69.5	75.3	82.5	82.8	88.1	88.4	89.1
MLIP	–	–	39.0	39.4	–	–	–	–	–	–	–	–	–
REFERS	–	–	–	–	–	–	76.7	80.9	84.7	–	89.4	91.6	**92.7**
MedKLIP	–	–	–	–	–	**76.8**	77.2	78.9	83.2	<u>86.9</u>	87.3	88.0	89.3
MRM	–	–	–	–	–	–	77.3	**83.5**	<u>85.5</u>	–	**91.0**	<u>91.8</u>	91.9
MaCo	1.066	0.228	57.2	54.3	<u>89.4</u>	<u>74.4</u>	<u>78.1</u>	<u>83.4</u>	85.4	**87.5**	<u>89.7</u>	<u>91.8</u>	92.0
SS-ACL	**1.153**	0.229	<u>58.3</u>	<u>55.0</u>	**89.9**	–	–	–	–	–	–	–	–
TraceCXR	1.116	0.220	**63.3**	**59.1**	**89.9**	73.2	**79.1**	<u>83.4</u>	**85.6**	84.8	87.6	**91.9**	<u>92.5</u>

Evidence grounding and transfer. Table 3 reports evidence-centric transfer. TraceCXR achieves the best Prec@5/10 on CheXpert 5×200 retrieval and competitive grounding, segmentation, and classification results, suggesting that evidence-addressable representations transfer beyond report generation.

4 Conclusion

We presented TraceCXR, an evidence-addressable framework for chest X-ray report generation. ProtoGraph links clinical concepts to disease-aware prototypes, EviLoop retrieves and grounds image-conditioned graph cues, and PSNP regularizes node selection and evidence concentration. Experiments show improved report quality, entity-level clinical correctness, and evidence-centric transfer. Trace Coverage provides a proxy for evidence linkage, while future reader studies will assess its clinician-facing utility.

Acknowledgments. This work was supported by the National Key Research and Development Program of China (2023YFC2705702), the LIESMARS Special Research Funding, and the WHU-Kingsoft Joint Lab. Computations were performed at the Supercomputing Center of Wuhan University.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Boecking, B., et al.: Making the most of text semantics to improve biomedical vision–language processing. In: ECCV. pp. 1–21 (2022)
2. Chambon, P., Delbrouck, J.B., Sounack, T., Huang, S.C., Chen, Z., Varma, M., Truong, S.Q., Chuong, C.T., Langlotz, C.P.: CheXpert Plus: Augmenting a large chest x-ray dataset with text radiology reports, patient demographics and additional image formats (2024)
3. Chen, Z., Shen, Y., Song, Y., Wan, X.: Cross-modal memory networks for radiology report generation. In: ACL-IJCNLP. pp. 5904–5914 (2021)
4. Chen, Z., Song, Y., Chang, T.H., Wan, X.: Generating radiology reports via memory-driven transformer. In: EMNLP. pp. 1439–1449 (2020)
5. Cornia, M., Stefanini, M., Baraldi, L., Cucchiara, R.: Meshed-memory transformer for image captioning. In: CVPR. pp. 10575–10584 (2020)
6. Demner-Fushman, D., Kohli, M.D., Rosenman, M.B., Shooshan, S.E., Rodriguez, L., Antani, S., Thoma, G.R., McDonald, C.J.: Preparing a collection of radiology examinations for distribution and retrieval. J. Am. Med. Inform. Assoc. **23**(2), 304–310 (2016)
7. Hou, W., Xu, K., Cheng, Y., Li, W., Liu, J.: ORGAN: Observation-guided radiology report generation via tree reasoning. In: ACL. pp. 8108–8122 (2023)
8. Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., Marklund, H., Haghgoo, B., Ball, R., Shpanskaya, K., Seekins, J., Mong, D.A., Halabi, S.S., Sandberg, J.K., Jones, R., Larson, D.B., Langlotz, C.P., Patel, B.N., Lungren, M.P., Ng, A.Y.: CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In: AAAI. pp. 590–597 (2019)
9. Jain, S., Agrawal, A., Saporta, A., Truong, S.Q., Duong, D.N., Bui, T., Chambon, P., Zhang, Y., Lungren, M.P., Ng, A.Y., Langlotz, C.P., Rajpurkar, P.: RadGraph: Extracting clinical entities and relations from radiology reports. arXiv preprint arXiv:2106.14463 (2021)
10. Jin, H., Che, H., Lin, Y., Chen, H.: Promptmrg: Diagnosis-driven prompts for medical report generation. In: AAAI. pp. 2607–2615 (2024)

11. Johnson, A.E.W., Pollard, T.J., Berkowitz, S.J., Greenbaum, N.R., Lungren, M.P., Deng, C.Y., Mark, R.G., Horng, S.: Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. Sci. Data **6**(1), 317 (2019)
12. Kaggle: SIIM-ACR pneumothorax segmentation. Kaggle Competition (2019)
13. Li, M., Lin, B., Chen, Z., Lin, H., Liang, X., Chang, X.: Dynamic graph enhanced contrastive learning for chest x-ray report generation. In: CVPR. pp. 3334–3343 (2023)
14. Liang, X., Zhang, Y., Wang, D., Zhong, H., Li, R., Wang, Q.: Divide and conquer: Isolating normal-abnormal attributes in knowledge graph-enhanced radiology report generation. In: ACM MM. pp. 4967–4975 (2024)
15. Liu, A., Guo, Y., Yong, J.H., Xu, F.: Multi-grained radiology report generation with sentence-level image-language contrastive learning. IEEE Trans. Med. Imaging **43**(7), 2657–2669 (2024)
16. Liu, F., Ge, S., Wu, X.: Competence-based multimodal curriculum learning for medical report generation. In: ACL-IJCNLP. pp. 3001–3012 (2021)
17. Liu, F., Wu, X., Ge, S., Fan, W., Zou, Y.: Exploring and distilling posterior and prior knowledge for radiology report generation. In: CVPR. pp. 13753–13762 (2021)
18. Mao, J., Wang, Y., Chen, L., Zhao, C., Tang, Y., Yang, D., Qu, L., Xu, D., Zhou, Y.: Discrete diffusion models with MLLMs for unified medical multimodal generation (2025)
19. Rahman, Z.U., Lee, J.H., Vu, D.T., Murtza, I., Kim, J.Y.: Duco-net: Dual-contrastive learning network for medical report retrieval leveraging enhanced encoders and augmentations. IEEE Access **13**, 27462–27476 (2025)
20. Shih, G., et al.: Augmenting the national institutes of health chest radiograph dataset with expert annotations of possible pneumonia. Radiol. Artif. Intell. **1**, e180041 (2019)
21. Tanida, T., Müller, P., Kaissis, G., Rueckert, D.: Interactive and explainable region-guided radiology report generation. In: CVPR. pp. 7433–7442 (2023)
22. Wang, J., Bhalerao, A., He, Y.: Cross-modal prototype driven network for radiology report generation. In: ECCV. pp. 563–579 (2022)
23. Wang, X., Li, Y., Wang, F., Wang, S., Li, C., Jiang, B.: R2GenCSR: Mining contextual and residual information for LLMs-based radiology report generation (2024), accepted by IEEE JBHI
24. Wang, X., Wang, F., Li, Y., Ma, Q., Wang, S., Jiang, B., Li, C., Tang, J.: CXPMRG-Bench: Pre-training and benchmarking for X-ray medical report generation on CheXpert plus dataset. In: CVPR. pp. 5123–5133 (2025)
25. Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., Summers, R.M.: ChestX-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In: CVPR. pp. 3462–3471 (2017)
26. Wang, Z., Liu, L., Wang, L., Zhou, L.: METransformer: Radiology report generation by transformer with multiple learnable expert tokens. In: CVPR. pp. 11558–11567 (2023)
27. Wang, Z., Liu, L., Wang, L., Zhou, L.: R2GenGPT: Radiology report generation with frozen LLMs. Meta-Radiology **1**(3), 100033 (2023)
28. Wang, Z., Tang, M., Wang, L., Li, X., Zhou, L.: A medical semantic-assisted transformer for radiographic report generation. In: MICCAI. pp. 655–664 (2022)
29. Xing, Q., Song, Z., Zhang, Y., Feng, N., Yu, J., Yang, W.: MCA-RG: Enhancing LLMs with medical concept alignment for radiology report generation. In: MICCAI. pp. 380–390 (2025)

30. Xue, Y., Tan, Y., Tan, L., Qin, J., Xiang, X.: Generating radiology reports via auxiliary signal guidance and a memory-driven network. Expert Syst. Appl. **237**, 121260 (2024)
31. Yan, A., He, Z., Lu, X., Du, J., Chang, E., Gentili, A., McAuley, J., Hsu, C.N.: Weakly supervised contrastive learning for chest X-ray report generation. In: Findings of EMNLP. pp. 4009–4015 (2021)
32. Yang, S., Wu, X., Ge, S., Zhou, S.K., Xiao, L.: Knowledge matters: Chest radiology report generation with general and specific knowledge. Med. Image Anal. **80**, 102510 (2022)
33. Yang, Y., Yu, J., Fu, Z., Zhang, K., Yu, T., Wang, X., Jiang, H., Lv, J., Huang, Q., Han, W.: Token-mixer: Bind image and text in one embedding space for medical image reporting. IEEE Trans. Med. Imaging **43**(11), 4017–4028 (2024)
34. You, D., Liu, F., Ge, S., Xie, X., Zhang, J., Wu, X.: Aligntransformer: Hierarchical alignment of visual regions and disease tags for medical report generation. In: MICCAI. pp. 72–82 (2021)
35. Zhang, Y., Wang, X., Xu, Z., Yu, Q., Yuille, A.L., Xu, D.: When radiology report generation meets knowledge graph. In: AAAI. pp. 12910–12917 (2020)