

SQV-R2Gen: Structured Query and Quality-Verified Spatial Evidence Distillation for Medical Report Generation (ChinaMM 2026)

Yuan Wang

China West Normal University

Xiaoou Liu

China West Normal University

Liming Xu

`xulimmail@gmail.com`

China West Normal University

Quan Tang

Northwestern Polytechnical University

Xianhua Zeng

Chongqing University of Posts and Telecommunications

Weisheng Li

Chongqing University of Posts and Telecommunications

Research Article

Keywords: Medical report generation, structured queries, spatial evidence distillation, clinical faithfulness

Posted Date: October 15th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10608113/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

SQV-R2Gen: Structured Query and Quality-Verified Spatial Evidence Distillation for Medical Report Generation (ChinaMM 2026)

Yuan Wang¹, Xiaoou Liu¹, Liming Xu^{1, 2*}, Quan Tang²,
Xianhua Zeng³, Weisheng Li³

¹Artificial Intelligence Research Institute, School of Computer Science,
China West Normal University, Nanchong, 637009, China.

²School of Automation, Northwestern Polytechnical University, Xi'an,
710129, China.

³Chongqing Key Laboratory of Image Cognition, College of Computer
Science and Technology, Chongqing University of Posts and
Telecommunications, Chongqing, 400065, China.

*Corresponding author(s). E-mail(s): xulimmail@gmail.com;

Contributing authors: wangyuan912@stu.cwnu.edu.cn;

liuxo@std.cwnu.edu.cn; quant@mail.nwpu.edu.cn;

zengxh@cqupt.edu.cn; liws@cqupt.edu.cn;

Abstract

Medical report generation aims to generate clinically accurate reports from medical images. Despite recent progress, existing methods still suffer from three prominent limitations: they often rely on holistic visual representations while overlooking fine-grained local lesion evidence, weakly supervised spatial cues are noisy and unstable and thus difficult to use as reliable supervision, and current alignment strategies are largely global or coarse-grained, lacking an explicit connection between structured abnormality semantics and trustworthy local visual evidence. To address these limitations, we propose SQV-R2Gen, a structured-query and quality-verified framework for medical report generation. Firstly, free-text reports are parsed into structured abnormality queries to construct query-conditioned supervisory interface for local pathological findings. Then, medical vision-language teacher module is introduced to generate candidate heatmaps, and multi-dimensional quality verification mechanism is designed to filter noisy responses by considering anatomical priors,

laterality constraints, spatial coverage, response concentration and boundary penalties. Besides, the verified spatial evidence is distilled into the report generator and jointly optimized with a disease-level semantic consistency objective, thereby strengthening local abnormality modeling and clinical faithfulness. **Experiments on benchmark datasets demonstrate that SQV-R2Gen achieves competitive language-generation performance and favorable precision-oriented clinical efficacy, while providing training-time traceable weak spatial supervision for abnormality-related findings.**

Keywords: Medical report generation, structured queries, spatial evidence distillation, clinical faithfulness

1 Introduction

Medical report generation is a representative medical image analysis task at the intersection of computer vision, natural language processing and clinical knowledge. It aims to translate pathological evidence in medical images into clinically standardized, semantically precise, and diagnostically meaningful reports [1–3]. Comparing with generic image captioning, it requires accurate recognition of fine-grained lesions, anatomical locations, and abnormal attributes, while maintaining strict clinical factual consistency and conformity to radiological writing conventions. In clinical practice, image interpretation and report drafting still rely heavily on radiologists, resulting in time-consuming and labor-intensive. With continuing growth of medical imaging, automatic report generation has become increasingly important for improving clinical efficiency, reducing physician workload and supporting AI-assisted diagnosis [4].

Most existing studies formulate medical report generation within the encoder-decoder paradigm that maps medical images to diagnostic text. Now, it has evolved from early CNN-RNN [5] architectures to advanced frameworks based on Vision Transformers [6], vision-language pretraining, and large language models. Recent progress mainly focuses on enhancing visual representation [7], improving cross-modal interaction [8, 9] and strengthening linguistic quality. However, three limitations remain prominent. First, existing end-to-end models often overlook fine-grained local evidence, such as lesion position, anatomical detail, and laterality. Then, weak spatial cues are usually noisy and unstable, making them difficult to serve as reliable supervision. Additionally, current alignment strategies are still dominated by global or coarse-grained matching, lacking an explicit mechanism to connect structured abnormality semantics with trustworthy local evidence [10, 11].

As illustrated in Fig. 1, the limitations lead to that diagnostic statements are often generated from holistic representations, while clinically critical local evidence is insufficiently modeled. In chest radiography, diagnostic conclusions usually depend not only on global appearance but also on lesion location, organ-level details, laterality, and abnormal extent [12]. Therefore, an effective report generation model should simultaneously capture global pathological context and establish precise correspondences between local visual evidence and fine-grained textual semantics.

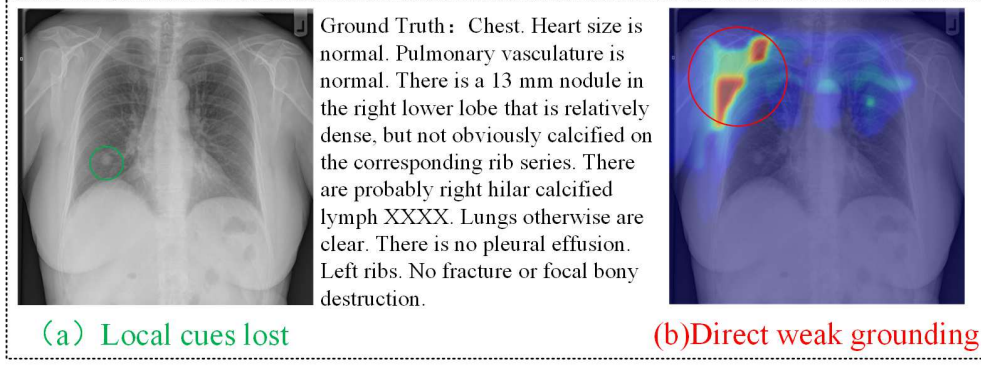


Fig. 1 Motivation of SQV-R2Gen. (a) Holistic visual representations may weaken or miss clinically important local lesion cues. (b) Directly using weakly supervised pseudo-heatmaps often introduces noisy and unreliable spatial responses. This motivates the construction of structured and trustworthy local evidence for medical report generation.

Motivated by the above observations, we propose SQV-R2Gen, a structured-query and quality-verified framework for medical report generation. **The proposed method separates offline structured-query-based evidence construction from evidence-guided model optimization and image-only inference, thereby avoiding report-derived information leakage during deployment. Without requiring manual region-level annotations, SQV-R2Gen provides quality-controlled supervision over local lesion evidence and aims to improve clinical faithfulness.** The main contributions are summarized as follows:

- We propose Structured Query and Quality-Verified R2Gen (SQV-R2Gen), a weakly supervised spatial evidence framework for medical report generation, which converts free-text reports into structured abnormality queries and constructs query-conditioned local supervision without manual box annotations.
- We design a quality verification mechanism for weakly supervised heatmaps, which incorporates anatomical priors and multidimensional scoring to filter noisy responses and retain reliable spatial evidence.
- We introduce a spatial evidence distillation strategy for report generation, which injects verified local evidence into the generation backbone and couples it with disease-level semantic consistency to improve abnormality perception and report faithfulness.
- **Experiments on two public benchmark datasets demonstrate that SQV-R2Gen achieves competitive language-generation performance and favorable precision-oriented clinical efficacy, while providing training-time traceable weak spatial supervision for abnormality-related findings.**

The remainder of this paper is organized as follows. Section 2 reviews related work. Section 3 presents the proposed SQV-R2Gen, including deep framework, structured query-guided spatial evidence construction and report generation with spatial evidence distillation. Section 4 gives and analyzes results, and Section 5 concludes.

2 Related Work

This section reviews two fields of research closely related to our work: radiology report generation and weakly supervised local grounding. The former focuses on generating clinically accurate reports from medical images, the latter aims to establish reliable correspondences between textual abnormality descriptions and image regions.

2.1 Radiology Report Generation

Radiology report generation aims to automatically generate clinically accurate and coherent diagnostic reports from medical images. Early deep learning-based methods mainly followed an encoder–decoder paradigm, where visual representations are translated into free-text descriptions. R2Gen [13] introduces memory-driven Transformer to enhance report-level coherence, and CMN [14] exploits cross-modal memory to retrieve informative visual–textual patterns. To further improve clinical correctness, subsequent studies incorporated disease priors, knowledge bases, and semantic alignment mechanisms. On this line, PGT [15] uses prior knowledge to guide report generation, and M2KT [16] introduces learned knowledge base with multi-modal alignment to better associate visual observations with medical descriptions.

Recent studies have increasingly focused on learning clinically meaningful representations through fine-grained cross-modal correspondence. TSGET [17] proposes two-stage global enhanced Transformer to strengthen interaction between image and report. MRG-SILC [18] introduces sentence-level image language contrastive learning for multi-grained report generation. CAMANet [19] uses class activation maps to guide cross-modal attention. In addition, FMVP [20] formulates report generation as an observation-to-concept process, and GHFC [21] exploits graph-guided hybrid feature consistency for semi-supervised medical report generation. More recently, DMVF [7] investigates dual-modality visual feature flow to improve medical report generation from complementary visual representations.

Beyond these approaches, PPKED [22] combines posterior abnormal-region knowledge with prior information retrieved from reports and a medical knowledge graph, whereas ICT [23] calibrates inter-report and intra-report information before incorporating it into the generation process. Recent large-language-model-based systems follow a complementary direction. R2GenGPT [24] maps visual features into the embedding space of a frozen language model through a lightweight alignment module, while R2-LLM [25] adapts a general-domain language model to radiology reporting through in-domain instance induction and coarse-to-fine refinement.

Other studies address temporal reasoning, causal representation learning, disease-aware representation, sequence efficiency, and multi-view alignment. DDaTR [26] models differences between current and prior examinations through dynamic feature alignment and residual propagation. CMCRL [27] mitigates cross-modal confounding through radiology-specific pretraining and causal intervention, while DCG [28] strengthens disease-related representation learning. R2Gen-Mamba [29] employs a selective state-space model to reduce sequence-modeling complexity, whereas AAMN [30] explicitly aligns frontal and lateral image features before visual–textual fusion.

Despite above advances, they are still relying on global visual features, disease-level labels, suffering from poor consideration of local attention maps. The reliability of intermediate spatial evidence is usually not explicitly verified before being used for supervision or interpretation, which may propagate noisy localization responses or anatomically implausible activations into the report generator.

2.2 Weakly Supervised Local Grounding

Weakly supervised local grounding aims to localize disease-related image regions using image-level labels, textual descriptions, or image-text pairs. It requires no manual pixel-level or bounding-box annotations. In medical vision-language learning, such grounding ability can be important because radiological findings are often associated with specific anatomical regions and local pathological patterns.

By introducing explicit regional priors to guide region-text alignment, EGGCA-Net [31] and EGMA [32] leverage eye-gaze signals of radiologists to guide fine-grained alignment between image regions and textual semantics, which indicates that clinically informed local cues can improve report faithfulness. However, these methods depend on additional gaze annotations or manually acquired priors, which limits their scalability and practical applicability. Then, some investigate weakly supervised visual grounding without pixel-level annotations. DAP [33] suppresses background distraction and enhance disease-relevant localization by analyzing failure modes of medical vision-language models under weak grounding supervision. Anatomy-Grounded Weakly Supervised Prompt Tuning [34] further shows that anatomically constrained prompting produces more stable and clinically plausible grounding than unconstrained text prompting.

All above studies suggest that the effectiveness of weakly supervised localization depends not only on generating heatmaps, but also on improving the reliability and anatomical plausibility of the resulting spatial responses. More recently, local grounding has also been integrated into medical report generation. MAIRA-2 [35] formulates grounded radiology report generation within large-scale multimodal framework. VED [36] shows that grounded local evidence can improve both linguistic quality and clinical accuracy. However, existing methods are still facing challenges in capturing fine-grained evidence. More dramatically, weakly supervised spatial signals often lack systematic quality verification, and most alignment strategies remain global or single-granularity.

Taking these challenges into account, SQV-R2Gen coordinates three complementary operations, each addressing a specific limitation. Structured abnormality queries convert free-text findings into localizable semantic units, reducing the loss of lesion-specific information under report-level supervision. Multidimensional quality verification filters teacher-generated heatmaps using anatomical and spatial criteria, preventing unreliable or anatomically implausible responses from being propagated as training targets. Finally, quality-aware spatial evidence distillation, together with disease-level semantic consistency, explicitly associates abnormality semantics with local visual evidence while preserving an image-only inference pathway.

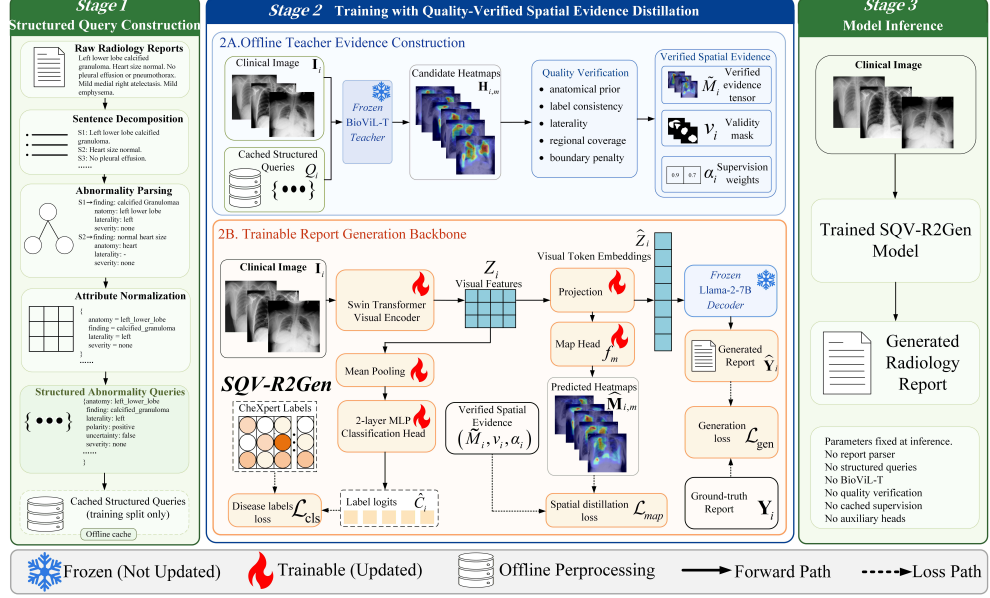


Fig. 2 Overall framework of SQV-R2Gen. In the first stage, structured abnormality queries are constructed from training reports and cached for subsequent supervision construction. In the second stage, the cached queries prompt a frozen BioViL-T teacher to produce candidate heatmaps; quality-verified spatial evidence then supervises the report generator through the generation, disease-level semantic consistency, and spatial distillation objectives. In the third stage, the trained model generates reports from clinical images alone. Snowflake, flame, and database icons denote frozen modules, trainable modules, and offline cached components, respectively. Solid arrows indicate forward computation, whereas dashed arrows indicate supervisory signals. *Best viewed in color.*

3 The Proposed Method

In this section, we present the details of proposed SQV-R2Gen. The overall architecture is illustrated in Fig. 2, which mainly consists of three explicitly separated stages, i.e., offline evidence construction, model optimization with cached supervision, and image-only report inference. Within it, report parsing, CheXpert labeling, structured queries, BioViL-T grounding, quality verification, cached supervision and auxiliary heads are only designed for training not for inference.

3.1 Framework Overview

In offline evidence construction stage, the report is decomposed into abnormality-related sentences and converted into structured queries for given training image-report pair (I_i, Y_i) . Then, BioViL-T [37] produces query-conditioned candidate heatmaps, which will be filtered and weighted through proposed quality verification mechanism. The resulting observation labels, verified heatmaps, validity masks, and supervision

weights are stored as sample-level cache, which can be denoted as

$$\mathcal{C}_i = \mathcal{E}_{\text{off}}(\mathbf{I}_i, \mathbf{Y}_i) = \left\{ \mathbf{c}_i, \widetilde{\mathbf{M}}_i, \mathbf{v}_i, \boldsymbol{\alpha}_i \right\} \quad (1)$$

where $\mathcal{E}_{\text{off}}(\cdot)$ denotes the complete offline preprocessing pipeline, $\mathbf{c}_i$ is the observation-label vector, $\widetilde{\mathbf{M}}_i$ is the fixed-size verified heatmap tensor, $\mathbf{v}_i$ is the validity mask, and $\boldsymbol{\alpha}_i$ contains the corresponding supervision weights.

During model optimization, the cached evidence is loaded directly from disk and is used only to construct auxiliary training objectives. The teacher model, report parser, CheXpert Labeler [38], and quality verification procedure are not executed inside the training loop. The model parameters are optimized as

$$\Theta^* = \arg \min_{\Theta} \sum_{i=1}^{N_{\text{tr}}} \mathcal{L}(\mathbf{I}_i, \mathbf{Y}_i, \mathcal{C}_i; \Theta) \quad (2)$$

where N_{tr} is the number of training samples and $\mathcal{L}$ combines report generation, disease-level semantic consistency, and spatial evidence distillation.

During the inference, i.e., Stage 3, neither ground-truth report nor any report-derived supervision is available or required. Given previously unseen image $\mathbf{I}$, report generation follows the standard image-to-text path, that is

$$\hat{\mathbf{Y}} = \mathcal{F}_{\text{inf}}(\mathbf{I}; \Theta^*) \quad (3)$$

3.2 Structured Query Construction

Structured query-guided spatial evidence construction establishes semantic transition from free-text reports to localized evidence, which enables visual representation to be regularized by lesion-related spatial cues rather than by report-level supervision alone. We firstly process training image-report pairs to recover localized abnormality semantics and convert them into training-only spatial supervision. Let the dataset be $\mathcal{D}_{\text{tr}} = \{(\mathbf{I}_i, \mathbf{Y}_i)\}_{i=1}^{N_{\text{tr}}}$, where $\mathbf{I}_i$ denotes the input image of the i -th training sample, $\mathbf{Y}_i$ denotes its paired report, and N_{tr} is the number of training samples.

For IU X-Ray, $\mathbf{I}_i$ contains both frontal and lateral radiographs for report generation, and teacher-side spatial evidence is constructed on frontal view where coarse thoracic anatomical priors are defined. For MIMIC-CXR, single radiograph is used following official split. We then standardize image-path mapping, sample identifiers, and report normalization before query parsing so that evidence construction and model optimization operate on the same sample-level indexing.

3.2.1 Structured abnormality query parsing

On the basis above split, each report is parsed into a set of structured abnormality queries. For the report $\mathbf{Y}_i$ of the i -th sample, we define

$$\mathcal{P}(\mathbf{Y}_i) = \mathcal{Q}_i = \{q_{i,1}, q_{i,2}, \dots, q_{i,M_i}\} \quad (4)$$

where $\mathcal{P}(\cdot)$ denotes report parsing function, $\mathcal{Q}_i$ denotes structured abnormality query set, and M_i denotes the number of extracted queries in the report. Each query can be represented as

$$q_{i,m} = (f_{i,m}, a_{i,m}, l_{i,m}, \rho_{i,m}, u_{i,m}, \xi_{i,m}, g_{i,m}) \quad (5)$$

where $f_{i,m}$ denotes the standardized abnormality category, $a_{i,m}$ denotes the normalized anatomical region, $l_{i,m}$ denotes the laterality state, $\rho_{i,m}$ denotes the statement polarity, $u_{i,m} \in \{0, 1\}$ denotes the uncertainty flag, $\xi_{i,m}$ denotes the source sentence from which the query is extracted, and $g_{i,m}$ denotes the condensed grounding prompt used to induce spatial evidence.

The structured formulation preserves compositional semantics among abnormality category, anatomy, and laterality, thereby converting localized clinical descriptions originally entangled in long-form reports into independently localizable semantic units. Comparing with whole-report supervision, such a representation is more compatible with establishing explicit correspondences to subsequent local spatial responses. For example, when the report contains the sentence ‘‘There are streaky bibasilar opacities,’’ the corresponding structured query can be written as

```
{
  "finding": "opacity",
  "anatomy": "lung_bases",
  "laterality": "bilateral",
  "polarity": "positive",
  "uncertainty": false,
  "source_sentence": "There are streaky bibasilar opacities",
  "query_prompt": "bilateral opacity"
}
```

Above example illustrates how the proposed query construction process converts free-text abnormality descriptions into anatomically grounded and localizable semantic units for subsequent spatial evidence generation. To make query construction reproducible without introducing additional inference-time dependency, we use fixed normalization vocabulary and rule-based anatomy-prior to map during offline preprocessing. The finding vocabulary is initialized from CheXpert space and extended with common localized radiographic findings after synonym normalization, including atelectasis, cardiomegaly, consolidation, edema, pleural effusion, pneumothorax, opacity or infiltration, pneumonia, nodule or mass, enlarged mediastinum and others.

Negated or normal findings are retained for disease-label consistency but are not used as positive spatial targets. Anatomical terms, such as the left lung, right lung, bilateral lungs, upper/middle/lower lung zones, lung apex, lung base, hilum, cardiac silhouette, mediastinum, pleural space, costophrenic angle, and diaphragm, are mapped to coarse binary prior masks. Specifically, lung-zone terms are obtained by splitting the valid lung field into upper, middle, and lower thirds; cardiac and mediastinal terms are mapped to central thoracic regions; pleural and costophrenic terms are mapped to peripheral or inferior thoracic regions. Laterality terms are normalized into left, right, bilateral, central, or unspecified states and are converted into the laterality prior $\mathbf{R}_{i,m}$.

When anatomical field cannot be reliably normalized, valid thoracic mask is used as conservative prior rather than discarding query. Polarity, uncertainty and severity are preserved as textual attributes in $g_{i,m}$ where uncertain findings are retained with reduced confidence during label construction and severity is not used to define a separate anatomical mask.

The normalized queries, source-sentence indices and corresponding observation labels are cached at sample level. In subsequent quality-verification, each is paired with training image to prompt frozen BioViL-T teacher, and query is used to construct supervision during training and is never provided to deployed model.

3.3 Quality-Verified Spatial Evidence Distillation

This component converts the cached abnormality queries into quality-controlled spatial supervision and distills the retained evidence into the report generator. It comprises query-conditioned grounding, multidimensional heatmap verification, fixed-size evidence construction, offline caching, evidence-guided visual learning, and joint report-generation optimization.

3.3.1 Query-conditioned grounding task construction

When obtaining $\mathcal{Q}_i$, we further reformulate each structured query into executable local grounding task. Specifically, the m -th task of the i -th sample is defined as

$$t_{i,m} = (\mathbf{I}_i, q_{i,m}, \eta_{i,m}) \quad (6)$$

where $\eta_{i,m}$ denotes the task metadata, including sample identity, query index, and source-sentence index. The corresponding task set is written as

$$\mathcal{T}_i = \{t_{i,m}\}_{m=1}^{M_i} \quad (7)$$

Then, localized abnormality descriptions originally embedded in free-text reports are transformed into weakly supervised grounding in form of image-query pairs.

3.3.2 Candidate heatmap generation via teacher grounding

For each grounding task, we employ **BioViL-T** [37] to generate a query-conditioned candidate spatial response. Let $\mathcal{G}$ denote the grounding function implemented by the teacher model, which maps an image-prompt pair to a spatial response map. The candidate heatmap corresponding to the m -th query of the i -th sample is defined as

$$\mathbf{H}_{i,m} = \mathcal{G}(\mathbf{I}_i, g_{i,m}) \quad (8)$$

where $\mathbf{H}_{i,m} \in \mathbb{R}^{h \times w}$ denotes the raw candidate heatmap, and h and w denote its spatial height and width, respectively. Since different prompts encode different abnormality semantics, the teacher can induce distinct localized responses on the same image. Nevertheless, $\mathbf{H}_{i,m}$ is only a weakly supervised candidate response rather than a manually annotated ground-truth map, and it may be corrupted by background activation, boundary overflow, anatomical drift, or laterality mismatch.

3.3.3 Quality verification of candidate heatmaps

Teacher-generated heatmaps are weak supervisory candidates and cannot be treated as ground-truth lesion annotations. We then introduce explicit quality verification to suppress anatomically implausible, laterality-inconsistent, overly diffuse or boundary-dominated responses. Each raw candidate heatmap is min-max normalized as

$$\bar{\mathbf{H}}_{i,m} = \frac{\mathbf{H}_{i,m} - \min(\mathbf{H}_{i,m})}{\max(\mathbf{H}_{i,m}) - \min(\mathbf{H}_{i,m}) + \varepsilon}, \quad (9)$$

where $\varepsilon = 10^{-6}$ is used for numerical stability. Let $\mathbf{T} \in \{0,1\}^{h \times w}$ denote the valid thoracic mask, $\mathbf{A}_{i,m} \in \{0,1\}^{h \times w}$ denote the anatomy-specific prior mask determined by $a_{i,m}$, $\mathbf{R}_{i,m} \in \{0,1\}^{h \times w}$ denote the laterality prior determined by $l_{i,m}$, and $\mathbf{B} \in \{0,1\}^{h \times w}$ denote a boundary-band mask covering the outer image margin. The activated support of the normalized heatmap is defined by

$$\Omega_{i,m} = \{(x, y) \mid \bar{\mathbf{H}}_{i,m}(x, y) \geq \delta_h\}, \quad (10)$$

where δ_h is activation threshold. We compute seven-dimensional descriptor, that is,

$$\phi_{i,m} = \phi(\mathbf{H}_{i,m}) = [\phi_{i,m}^{(1)}, \phi_{i,m}^{(2)}, \phi_{i,m}^{(3)}, \phi_{i,m}^{(4)}, \phi_{i,m}^{(5)}, \phi_{i,m}^{(6)}, \phi_{i,m}^{(7)}]^\top \in [0, 1]^7, \quad (11)$$

where seven components correspond to thoracic-region consistency, anatomy-prior consistency, laterality consistency, activated-area plausibility, boundary reliability, peak response, and activation-mass plausibility, which can be respectively defined as

$$\phi_{i,m}^{(1)} = \frac{\sum_{x,y} \bar{\mathbf{H}}_{i,m}(x, y) \mathbf{T}(x, y)}{\sum_{x,y} \bar{\mathbf{H}}_{i,m}(x, y) + \varepsilon}, \quad (12)$$

$$\phi_{i,m}^{(2)} = \frac{\sum_{x,y} \bar{\mathbf{H}}_{i,m}(x, y) \mathbf{A}_{i,m}(x, y)}{\sum_{x,y} \bar{\mathbf{H}}_{i,m}(x, y) + \varepsilon}, \quad (13)$$

$$\phi_{i,m}^{(3)} = \frac{\sum_{x,y} \bar{\mathbf{H}}_{i,m}(x, y) \mathbf{R}_{i,m}(x, y)}{\sum_{x,y} \bar{\mathbf{H}}_{i,m}(x, y) + \varepsilon}, \quad (14)$$

$$\phi_{i,m}^{(4)} = \mathbb{I} \left[a_{\min} \leq \frac{|\Omega_{i,m}|}{hw} \leq a_{\max} \right], \quad (15)$$

$$\phi_{i,m}^{(5)} = 1 - \frac{\sum_{x,y} \bar{\mathbf{H}}_{i,m}(x, y) \mathbf{B}(x, y)}{\sum_{x,y} \bar{\mathbf{H}}_{i,m}(x, y) + \varepsilon}, \quad (16)$$

$$\phi_{i,m}^{(6)} = \max_{x,y} \bar{\mathbf{H}}_{i,m}(x, y), \quad (17)$$

and

$$\phi_{i,m}^{(7)} = 1 - \left| \frac{1}{hw} \sum_{x,y} \bar{\mathbf{H}}_{i,m}(x, y) - \mu_0 \right|. \quad (18)$$

where $\mathbb{I}[\cdot]$ denotes indicator function, $a_{\min}$ and $a_{\max}$ define plausible activation-area range, and μ_0 is expected normalized activation mass. Large value of each descriptor indicates more reliable spatial evidence, and all descriptor values are clipped to $[0, 1]$ before scoring. Then, the final quality score is computed by an explicit weighted aggregation:

$$\omega_{i,m} = \beta^\top \phi_{i,m} = \sum_{r=1}^7 \beta_r \phi_{i,m}^{(r)}, \quad \beta_r \geq 0, \quad \sum_{r=1}^7 \beta_r = 1. \quad (19)$$

Unless otherwise specified, we use equal descriptor weights, i.e., $\beta_r = 1/7$, to avoid introducing an additional learnable verifier and to keep the scoring rule reproducible across datasets. According to $\omega_{i,m}$, each candidate heatmap is assigned to a quality bucket:

$$b_{i,m} = \begin{cases} \text{keep}, & \omega_{i,m} \geq \tau_k, \\ \text{weak}, & \tau_w \leq \omega_{i,m} < \tau_k, \\ \text{drop}, & \omega_{i,m} < \tau_w, \end{cases} \quad (20)$$

where τ_k and τ_w denote the keep and weak-evidence thresholds, respectively. The supervision weight associated with each candidate is then defined as

$$\alpha_{i,m} = \begin{cases} 1, & b_{i,m} = \text{keep}, \\ \gamma, & b_{i,m} = \text{weak}, \\ 0, & b_{i,m} = \text{drop}, \end{cases} \quad (21)$$

where $\gamma \in (0, 1)$ controls the contribution of weak but still plausible evidence. This rule allows high-quality responses to provide full supervision, weak responses to provide softened supervision, and unreliable responses to be excluded.

The retained evidence set is

$$\tilde{\mathcal{H}}_i = \left\{ \tilde{\mathbf{H}}_{i,m} = \bar{\mathbf{H}}_{i,m} \mid b_{i,m} \in \{\text{keep}, \text{weak}\} \right\}, \quad (22)$$

where $\tilde{\mathbf{H}}_{i,m} \in [0, 1]^{h \times w}$ denotes the verified teacher heatmap retained after filtering. For intuitive illustration, a lightweight retained evidence record can be written as

```
{
  "finding": "opacity",
  "anatomy": "lung_bases",
  "laterality": "bilateral",
  "query_prompt": "bilateral opacity",
  "bucket": "keep",
  "heatmap_weight": 1.0,
  "quality_score": 0.7253
}
```

where `bucket` indicates the post-filtering quality level, `heatmap_weight` is instantiated by $\alpha_{i,m}$, and `quality_score` corresponds to $\omega_{i,m}$.

3.3.4 Fixed-size evidence tensor construction

To support mini-batch training, we retain at most K heatmaps for each sample. For the i -th sample, the number of retained heatmaps is defined as

$$L_i = \min(|\tilde{\mathcal{H}}_i|, K) \quad (23)$$

where $\tilde{\mathcal{H}}_i$ denotes the set of verified heatmaps for the i -th sample, and $|\tilde{\mathcal{H}}_i|$ is its cardinality. The retained heatmaps are sorted in descending order according to their quality scores and are denoted by $\tilde{\mathbf{H}}_{i,(1)}, \dots, \tilde{\mathbf{H}}_{i,(L_i)}$, where $\omega_{i,(1)} \geq \dots \geq \omega_{i,(L_i)}$.

The ordered heatmaps are stacked along a new first dimension to form

$$\mathbf{S}_i = \text{Stack}(\tilde{\mathbf{H}}_{i,(1)}, \dots, \tilde{\mathbf{H}}_{i,(L_i)}) \in \mathbb{R}^{L_i \times h \times w} \quad (24)$$

where $\text{Stack}(\cdot)$ concatenates retained two-dimensional heatmaps along heatmap dimension while preserving spatial size $h \times w$. When $L_i < K$, zero padding is applied along the same dimension to obtain fixed-size evidence tensor, i.e.,

$$\tilde{\mathbf{M}}_i = \text{Pad}_K(\mathbf{S}_i) \in \mathbb{R}^{K \times h \times w} \quad (25)$$

where $\text{Pad}_K(\cdot)$ appends exactly $K - L_i$ all-zero maps of size $h \times w$ to the end of $\mathbf{S}_i$.

The first L_i slices of $\tilde{\mathbf{M}}_i$ contain retained heatmaps, and remaining $K - L_i$ slices correspond to padded positions. If $L_i = 0$, $\tilde{\mathbf{M}}_i$ is directly initialized as all-zero tensor. Then, binary vector is introduced to distinguish retained heatmaps from padded positions, i.e.,

$$\mathbf{v}_i = (\underbrace{1, \dots, 1}_{L_i}, \underbrace{0, \dots, 0}_{K-L_i}) \in \{0, 1\}^K \quad (26)$$

where $v_{i,m} = 1$ indicates that the m -th slice contains retained heatmap, and $v_{i,m} = 0$ indicates padded position. Corresponding supervision-weight vector is defined as

$$\boldsymbol{\alpha}_i = (\alpha_{i,1}, \dots, \alpha_{i,L_i}, \underbrace{0, \dots, 0}_{K-L_i}) \in [0, 1]^K \quad (27)$$

For each valid position $m \leq L_i$, $\alpha_{i,m}$ is determined by quality-aware weighting rule and all padded positions are assigned zero weight. Consequently, $\mathbf{v}_i$ determines whether tensor position participates in spatial distillation, and $\boldsymbol{\alpha}_i$ controls relative contribution of each retained heatmap.

3.3.5 Offline evidence caching

The structured-query parsing and teacher-grounding pipeline is executed once before report-generator training. Specifically, for each sample, we extract observation labels and abnormality attributes from paired report, construct corresponding structured queries, generate candidate heatmaps with frozen BioViL-T teacher, and apply quality verification. The verified heatmap tensor $\tilde{\mathbf{M}}_i$, validity mask $\mathbf{v}_i$, supervision weights $\boldsymbol{\alpha}_i$, and observation-label vector $\mathbf{c}_i$ are then serialized and kept.

The offline design separates evidence construction from model optimization. During training, the data loader directly retrieves $(\mathbf{I}_i, \mathbf{Y}_i, \mathbf{c}_i, \widetilde{\mathbf{M}}_i, \mathbf{v}_i, \boldsymbol{\alpha}_i)$ without invoking the report parser, CheXpert Labeler [38], BioViL-T [37], or quality verification module. Consequently, no gradients are propagated through the teacher-side pipeline, and the teacher model does not occupy GPU memory during report-generator optimization.

The cached evidence acts only as an auxiliary target for disease-level semantic supervision and spatial evidence distillation. It is not concatenated with the image as an inference-time input and is not required by the language decoder. Moreover, no ground-truth report from the test sample is accessed by the model during report generation, thereby preventing report-derived information leakage at inference.

Using the cached verified spatial evidence as training-only supervision, we construct a report generation model equipped with evidence distillation. During optimization, the model contains the visual encoder, the visual-to-language projection module, the frozen language decoder, and two auxiliary training branches for disease-level semantic consistency and spatial evidence distillation. During inference, the auxiliary classification and map-prediction branches are removed, and the deployed model retains only the visual encoder, projection module, and language decoder.

3.3.6 Visual encoding

For the input image $\mathbf{I}_i$, we first employ a Swin Transformer [39] to extract patch-level visual features:

$$\mathbf{Z}_i = f_v(\mathbf{I}_i) \in \mathbb{R}^{N_p \times d_v} \quad (28)$$

where $f_v(\cdot)$ denotes the visual encoder, N_p denotes the number of visual tokens, and d_v denotes the visual feature dimension. To make the visual features compatible with the language decoder, we further project them into the language embedding space:

$$\widehat{\mathbf{Z}}_i = f_p(\mathbf{Z}_i) = \text{LN}(\mathbf{Z}_i \mathbf{W}_p + \mathbf{b}_p) \quad (29)$$

where $f_p(\cdot)$ denotes the visual-to-language projection module, $\mathbf{W}_p \in \mathbb{R}^{d_v \times d_l}$ and $\mathbf{b}_p \in \mathbb{R}^{d_l}$ are learnable parameters, d_l denotes the language embedding dimension, and $\text{LN}(\cdot)$ denotes layer normalization.

3.3.7 Disease-level semantic consistency

Beyond the report generation objective, we impose an auxiliary disease-level semantic constraint on the visual representation. The 14 thoracic observations are automatically extracted from the original reports using **CheXpert Labeler** [38]¹, yielding a standardized binary observation vector, which can be denoted as

$$\mathbf{c}_i \in \{0, 1\}^{14} \quad (30)$$

Let $f_c(\cdot)$ denote the classification head. The predicted observation logits of the i -th sample are given by

$$\widehat{\mathbf{c}}_i = f_c(\text{Pool}(\widehat{\mathbf{Z}}_i)) \in \mathbb{R}^{14} \quad (31)$$

¹<https://github.com/stanfordmlgroup/chexpert-labeler>

where $\text{Pool}(\cdot)$ denotes the global pooling operator. The corresponding multi-label classification loss is defined as

$$\mathcal{L}_{\text{cls}} = -\frac{1}{14B} \sum_{i=1}^B \sum_{j=1}^{14} [c_{i,j} \log \sigma(\hat{c}_{i,j}) + (1 - c_{i,j}) \log (1 - \sigma(\hat{c}_{i,j}))] \quad (32)$$

where B denotes the mini-batch size, $c_{i,j}$ and $\hat{c}_{i,j}$ denote the target and predicted logit of the j -th observation for the i -th sample, respectively, and $\sigma(\cdot)$ denotes the sigmoid function.

3.3.8 Spatial evidence distillation

To ensure that the verified local evidence explicitly regularizes visual representation learning, we introduce a spatial evidence distillation branch. Let

$$f_m : \mathbb{R}^{N_p \times d_l} \rightarrow \mathbb{R}^{K \times h \times w} \quad (33)$$

denote the map prediction head. The student-predicted response maps are defined as

$$\widehat{\mathbf{M}}_i = \sigma(f_m(\widehat{\mathbf{Z}}_i)) \in [0, 1]^{K \times h \times w} \quad (34)$$

where $\widehat{\mathbf{M}}_{i,m} \in [0, 1]^{h \times w}$ denotes the m -th student-predicted response map after sigmoid normalization. Each $\widehat{\mathbf{M}}_{i,m}$ is aligned with the corresponding verified teacher target $\widetilde{\mathbf{M}}_{i,m}$.

For the m -th valid slice, the pixel-wise distillation term is defined as

$$\ell_{\text{map}}^{(i,m)} = \frac{1}{hw} \sum_{x=1}^h \sum_{y=1}^w \varphi(\widehat{\mathbf{M}}_{i,m}(x, y), \widetilde{\mathbf{M}}_{i,m}(x, y)) \quad (35)$$

where $\varphi(a, b) = (a-b)^2$ is instantiated as the mean squared error between the predicted response and the verified teacher heatmap. By incorporating both the validity mask and the supervision weights, the overall spatial distillation loss is formulated as

$$\mathcal{L}_{\text{map}} = \frac{1}{B} \sum_{i=1}^B \frac{1}{\sum_{m=1}^K v_{i,m} + \varepsilon} \sum_{m=1}^K v_{i,m} \alpha_{i,m} \ell_{\text{map}}^{(i,m)} \quad (36)$$

where $v_{i,m}$ denotes the validity indicator, $\alpha_{i,m}$ denotes the supervision weight of the m -th verified heatmap, and ε is a small constant to avoid division by zero. In this way, high-quality evidence contributes stronger spatial supervision, weak-quality evidence contributes softer supervision, and discarded candidates do not participate in training.

3.3.9 LLM-based report decoding

Based on the visual conditional representation, we adopt Llama-2-7B-Chat [40] as the report decoder for autoregressive text generation. Let the ground-truth report be

denoted as

$$\mathbf{Y}_i = (y_{i,1}, \dots, y_{i,T_i}) \quad (37)$$

where T_i denotes the report length. We further denote the ground-truth token prefix by $\mathbf{y}_{i,<t} = (y_{i,1}, \dots, y_{i,t-1})$. During training, the conditional probability of the t -th token is written as

$$P(y_{i,t} | \mathbf{y}_{i,<t}, \hat{\mathbf{Z}}_i) = f_g(\mathbf{y}_{i,<t}, \hat{\mathbf{Z}}_i) \quad (38)$$

where $f_g(\cdot)$ denotes the conditional token distribution parameterized by the frozen Llama-2-7B-Chat decoder and $\mathbf{y}_{i,<t}$ denotes the ground-truth prefix before step t . The corresponding generation loss is defined as

$$\mathcal{L}_{\text{gen}} = -\frac{1}{B} \sum_{i=1}^B \sum_{t=1}^{T_i} \log P(y_{i,t} | \mathbf{y}_{i,<t}, \hat{\mathbf{Z}}_i) \quad (39)$$

During training, the Llama-2-7B-Chat decoder is kept frozen, and gradients from the generation loss update the visual encoder and the visual-to-language projection module. The verified spatial evidence regularizes visual representation learning only during optimization. During inference, the decoder generates the report autoregressively from the projected visual representation and previously generated tokens; BioViL-T and the report-derived supervision pipeline are not executed.

3.3.10 Joint optimization

The overall optimization objective of SQV-R2Gen combines the report generation loss, the disease-level semantic consistency loss, and the spatial evidence distillation loss:

$$\mathcal{L} = \mathcal{L}_{\text{gen}} + \lambda_{\text{cls}} \mathcal{L}_{\text{cls}} + \lambda_{\text{map}} \mathcal{L}_{\text{map}}. \quad (40)$$

where λ_{cls} and λ_{map} denote the weighting coefficients for disease-level semantic supervision and spatial evidence distillation, respectively. Through this joint optimization, the model encourages local abnormality perception while preserving report-generation quality, thereby strengthening the correspondence between generated findings and the underlying visual evidence.

3.4 Image-Only Report Inference

The report-derived supervision branches are used exclusively during optimization. Once training is completed, the disease classification head and map prediction head are not required for report generation and can be removed from the inference graph. For a test image $\mathbf{I}$, the deployed model performs

$$\mathbf{I} \xrightarrow{f_v} \mathbf{Z} \xrightarrow{f_p} \hat{\mathbf{Z}} \xrightarrow{f_g} \hat{\mathbf{Y}} \quad (41)$$

where f_v , f_p , and f_g denote the visual encoder, visual-to-language projection module, and autoregressive language decoder, respectively.

Neither the paired ground-truth report nor the structured queries, CheXpert labels, teacher-generated heatmaps, or quality scores are available to the model during inference. In particular, BioViL-T is not loaded or executed at test time. The additional teacher-side computation is therefore confined to one-time offline preprocessing of the training data. The online deployment path remains an image-to-report generator and introduces no teacher-model latency or teacher-model memory consumption.

4 Experiments

To verify the effectiveness of proposed SQV-R2Gen, we conduct extensive experiments on IU X-Ray and MIMIC-CXR datasets through quantitative comparison, and qualitative analysis. Then, clinical efficacy and ablation studies are provided to analyze the key modules. The implementations can be accessed at <https://github.com/WangYuan912/SQV-R2Gen>.

4.1 Datasets and Evaluation Metrics

Experiments are conducted on two public benchmark datasets, namely IU X-Ray [41] and MIMIC-CXR [42]. The IU X-Ray dataset, released by Indiana University, contains chest radiographs paired with corresponding radiology reports. MIMIC-CXR, collected at Beth Israel Deaconess Medical Center, is a large-scale chest X-ray dataset with paired radiology reports and is widely used as a representative benchmark for automatic radiology report generation. In all experiments, samples with missing reports are excluded, and the same training, validation, and test splits as prior report-generation studies are adopted. The detailed numbers of images and reports used in our experiments are summarized in Table 1.

Table 1 Statistics and partition of IU X-Ray and MIMIC-CXR.

Dataset	Type	Train	Val	Test	Total
IU X-Ray	Image	5,226	748	1,496	7,470
	Report	2,770	395	790	3,955
MIMIC-CXR	Image	368,960	2,991	5,159	377,110
	Report	222,758	1,808	3,269	227,835

To comprehensively evaluate the proposed method, we adopt both natural language generation and clinical efficacy metrics, following prior work [16, 43, 44]. Specifically, report quality is evaluated using BLEU-1 through BLEU-4 [45], METEOR [46], and ROUGE-L [47]. To assess clinical faithfulness, we further employ the CheXpert [38] tool to map both generated reports and ground-truth reports into 14 predefined CheXpert observations, based on which Precision, Recall, and F1-score are computed. These metrics together provide a comprehensive evaluation of both linguistic quality and clinical correctness.

4.2 Implementation Details

SQV-R2Gen is implemented in PyTorch based on the R2GenGPT framework [24]. Following recent LLM-based report generations, we use Swin Transformer [39] as visual encoder and Llama-2-7B-Chat [40] as report decoder. The language decoder is kept frozen, and optimize visual encoder, visual-to-language projection module. For IU X-Ray, frontal and lateral views are jointly used for report generation, and teacher-side spatial evidence is constructed on the frontal view. For MIMIC-CXR, a single frontal image is used following previous works.

For quality-verified evidence distillation, at most $K = 3$ verified heatmaps are retained for each sample. The keep and weak thresholds are fixed to $\tau_k = 0.60$ and $\tau_w = 0.40$, respectively, and seven quality descriptors are assigned equal weights. The supervision weight for weak evidence is set to $\gamma = 0.50$. The disease-level branch predicts 14 thoracic observations extracted by the CheXpert Labeler [38]. The model is trained using AdamW with an initial learning rate of 1×10^{-4} and a weight decay of 1×10^{-2} . The loss weights are set to $\lambda_{\text{cls}} = 0.3$ and $\lambda_{\text{map}} = 0.15$.

During structured-query construction, CheXpert labeling, BioViL-T grounding and quality verification are performed. The resulted labels, verified heatmaps, validity masks, and quality-aware weights will be cached and loaded during training. During inference, SQV-R2Gen retains visual encoder, visual-to-language projection and frozen Llama-2-7B-Chat decoder. Report parser, CheXpert Labeler, BioViL-T teacher, quality-verification module and auxiliary heads are not used during inference.

4.3 Quantitative Analysis

Table 2 reports natural language generation (NLG) results on IU X-Ray and MIMIC-CXR dataset, and NLG evaluator include BLEU@N, METEOR, and ROUGE-L. Since the values are collected from corresponding publications, they characterize comparative performance under the reported settings rather than a unified re-evaluation with identical encoders, decoders, and pretraining configurations. From Table 2, we can generally see that our proposed SQV-R2Gen achieve satisfactory generation comparing with other methods.

On IU X-Ray, SQV-R2Gen achieves the best BLEU-1, BLEU-2 and BLEU-4 scores, and it also gains second-best BLEU-3 and METEOR scores. TSGET remains strongest on BLEU-3, METEOR, and ROUGE-L, indicating the effectiveness of global contextual enhancement for maximizing several text-overlap metrics. AAMN achieves the second-best BLEU-1 score, showing the value of explicit interaction between frontal and lateral image features, while M2KT and SQV-R2Gen share the second-best ROUGE-L score of 0.399. PPKED, M2KT, and ICT progressively improve upon the early R2Gen baseline by introducing posterior/prior knowledge, learned medical knowledge, and calibrated contextual information. The LLM-based R2GenGPT and R2-LLM provide strong linguistic priors, whereas MRG-SILC improves sentence-level topic representation through fine-grained contrastive learning and hierarchical topic-to-word generation. Nevertheless, its gains are more evident on the larger MIMIC-CXR dataset than on IU X-Ray, where the original study did not apply reinforcement-learning fine-tuning.

Table 2 Comparison of natural language generation performance on the IU X-Ray and MIMIC-CXR datasets. The best results are shown in bold, the second-best results are underlined, and “–” indicates that the corresponding metric was not reported. For PPKE on IU X-Ray, the original paper reports ROUGE-L and CIDEr rather than METEOR. Thus, METEOR score is marked as “–”, ROUGE-L score is 0.376, and reported CIDEr score of 0.351 is not included in this table.

Dataset	Method	Year	Venue	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L
IU X-Ray	R2Gen [13]	2020	EMNLP	0.470	0.304	0.219	0.165	0.187	0.371
	PPKE [22]	2021	CVPR	0.483	0.315	0.224	0.168	–	0.376
	M2KT [16]	2023	MIA	0.497	0.319	0.230	0.174	–	<u>0.399</u>
	ICT [23]	2023	JBHI	0.503	0.341	0.246	0.186	0.208	0.390
	R2GenGPT [24]	2023	Meta-Radiology	0.488	0.316	0.228	0.173	0.211	0.377
	TSGET [17]	2024	JBHI	0.500	<u>0.349</u>	0.256	<u>0.194</u>	0.218	0.402
	R2-LLM [25]	2024	AAAI	0.499	0.323	0.238	0.184	0.208	0.390
	MRG-SILC [18]	2024	TMI	0.472	0.321	0.234	0.175	0.192	0.379
	DDaTR [26]	2025	TMI	0.421	–	–	0.103	0.160	0.307
	CMCRL [27]	2025	TIP	0.505	0.334	0.245	0.190	0.210	0.394
	R2Gen-Mamba [29]	2025	ISBI	0.482	0.315	0.228	0.176	0.208	0.382
	AAMN [30]	2026	Multimed. Syst.	<u>0.506</u>	0.340	0.245	0.177	0.216	0.368
	SQV-R2Gen	2026	This work	0.520	0.354	<u>0.253</u>	0.201	<u>0.217</u>	<u>0.399</u>
MIMIC-CXR	R2Gen [13]	2020	EMNLP	0.353	0.218	0.145	0.103	0.142	0.277
	PPKE [22]	2021	CVPR	0.360	0.224	0.149	0.106	0.149	0.284
	M2KT [16]	2023	MIA	0.386	0.237	0.157	0.111	–	0.274
	ICT [23]	2023	JBHI	0.376	0.233	0.157	0.113	0.144	0.276
	R2GenGPT [24]	2023	Meta-Radiology	<u>0.411</u>	<u>0.267</u>	<u>0.186</u>	0.134	0.160	<u>0.297</u>
	TSGET [17]	2024	JBHI	0.398	0.248	0.169	0.121	0.149	0.281
	R2-LLM [25]	2024	AAAI	0.402	0.262	0.180	0.128	0.175	0.291
	MRG-SILC [18]	2024	TMI	0.406	<u>0.267</u>	0.190	<u>0.141</u>	0.163	0.309
	DDaTR [26]	2025	TMI	0.401	–	–	0.113	0.162	0.272
	CMCRL [27]	2025	TIP	0.400	0.245	0.165	0.119	0.150	0.280
	R2Gen-Mamba [29]	2025	ISBI	0.352	0.222	0.152	0.110	0.141	0.284
	AAMN [30]	2026	Multimed. Syst.	0.395	0.245	0.161	0.120	0.158	0.281
	SQV-R2Gen	2026	This work	0.413	0.269	0.174	0.142	<u>0.172</u>	0.286

The strong performance of SQV-R2Gen on IU X-ray suggests that structured abnormality semantics and verified local evidence are effective for selecting clinically relevant words and constructing accurate abnormality-related phrases even under a relatively small training set. Comparing with that primarily optimize global representations or report-level semantics, SQV-R2Gen introduces reliability-control stage before weak spatial responses are used as supervision, which is intended to reduce propagation of anatomically inconsistent, laterality-mismatched, excessively diffuse, or boundary-dominated responses. But, ROUGE-L does not exceed that of TSGET, indicating that the proposed spatial supervision does not uniformly maximize every form of long-range lexical overlap.

On MIMIC-CXR, different strengths of comparisons become more apparent. MRG-SILC achieves the highest BLEU-3 and ROUGE-L, and the second-best BLEU-4 score. These results are consistent with explicit modeling of sentence topics and two-stage topic-to-word decoding strategy, which is designed to generate multi-sentence reports with more accurate topic organization and sentence content. R2GenGPT achieves strong BLEU-3 and ROUGE-L performance through linguistic prior of frozen LLM, while R2-LLM obtains the highest METEOR score through domain adaptation and coarse-to-fine refinement. TSGET, CMCRL, DDaTR, R2Gen-Mamba, and AAMN provide complementary improvements through global enhancement, causal learning, temporal modeling, efficient encoding, and cross-modal alignment, respectively.

Comparing with R2GenGPT backbone, SQV-R2Gen improves BLEU-1 from 0.411 to 0.413, BLEU-2 from 0.267 to 0.269, BLEU-4 from 0.134 to 0.142, and METEOR from 0.160 to 0.172. Compared with MRG-SILC, SQV-R2Gen performs better on BLEU-1, BLEU-2, BLEU-4, and METEOR, whereas MRG-SILC remains stronger on BLEU-3 and ROUGE-L. This complementary pattern suggests that sentence-level topic modeling is particularly effective for preserving multi-sentence lexical structure, while quality-verified spatial evidence provides stronger support for abnormality-related word selection and phrase-level matching.

Overall, no single method dominates all NLG metrics on both datasets because the compared approaches optimize different aspects of report generation, including recurrent report patterns, prior knowledge, global context, LLM adaptation, sentence-topic organization, temporal progression, and computational efficiency. The main advantage of SQV-R2Gen is therefore not an across-the-board increase in conventional text-overlap scores. Instead, it achieves consistently competitive linguistic performance while introducing an explicit mechanism to verify the reliability of weakly supervised local evidence before distillation. Together with the clinical efficacy results reported below, these findings support the central objective of SQV-R2Gen: improving evidence-grounded and precision-oriented radiology report generation without sacrificing overall language quality.

4.3.1 Clinical Efficacy

Table 3 reports clinical efficacy results on MIMIC-CXR. The comparison is based on Precision, Recall, and F1-score, and potential differences in label aggregation and evaluation implementations are considered when interpreting cross-paper results.

SQV-R2Gen achieves the highest Precision and Recall of 0.476 and 0.418, respectively, together with the second-best F1-score of 0.411. Compared with R2GenGPT, SQV-R2Gen improves Precision, Recall, and F1-score by 21.4%, 8.0%, and 5.7%, respectively. Compared with DCG, it improves Precision by 7.9%, Recall by 1.0%, and F1-score by 1.7%. It also exceeds DMVF by 5.5% in Precision and 8.3% in Recall, although its F1-score is marginally lower by 0.002.

The strong Precision of SQV-R2Gen indicates that the generated reports contain fewer falsely predicted abnormal findings, while highest Recall shows that improvement is not achieved simply by adopting an overly conservative generation strategy. These gains mainly arise from the proposed supervision mechanism, which transforms free-text reports into structured abnormality queries and filters weakly supervised spatial responses according to their anatomical and spatial reliability before distillation. Consequently, the report generator receives more trustworthy local supervision regarding both the presence of abnormalities and their corresponding anatomical regions.

DMVF reports marginally higher F1-score, whereas SQV-R2Gen achieves higher reported Precision and Recall. Since the values in Table 3 are collected from corresponding publications, differences in label aggregation and evaluation implementations may affect direct cross-paper comparison. It should be interpreted as literature-reported evidence rather than as fully controlled re-evaluation. Under reported settings, SQV-R2Gen exhibits favorable balance between precision-oriented faithfulness, abnormality coverage and F1 performance.

Table 3 Comparison of clinical efficacy on the MIMIC-CXR dataset. The best results are shown in bold, and the second-best results are underlined.

Dataset	Method	Precision	Recall	F1-Score
MIMIC-CXR	Transformer [48]	0.331	0.224	0.228
	R2Gen [13]	0.333	0.273	0.276
	CMN [14]	0.334	0.275	0.278
	R2GenGPT [24]	0.392	0.387	0.389
	M2KT [16]	0.420	0.339	0.352
	TRINet [49]	0.393	0.351	0.371
	DCG [28]	0.441	<u>0.414</u>	0.404
	DMVF [7]	<u>0.451</u>	0.386	0.413
	SQV-R2Gen	0.476	0.418	<u>0.411</u>

4.4 Qualitative Analysis

Fig. 3 presents a representative qualitative comparison among R2Gen, M2KT, R2GenGPT, and SQV-R2Gen. The figure shows the ground-truth report, its sentence decomposition, the structured abnormality queries extracted from the positive sentences, and the reports generated by different methods. In this case, the reference

	Ground Truth: Heart is mildly enlarged stable. Mediastinal contour is normal. Pulmonary vascularity is normal. Lungs are hyperexpanded but clear. No pleural effusions or pneumothoraces.	Sid 0: Heart is mildly enlarged stable. Sid 1: Mediastinal contour is normal. Sid 2: Pulmonary vascularity is normal. Sid 3: Lungs are hyperexpanded but clear. Sid 4: No pleural effusions or pneumothoraces.
Sid 0: Structured abnormality query: <pre>{ anatomy: heart finding: cardiomegaly polarity: positive uncertainty: false severity: mild laterality: none intent: verify source_sentence: Heart is mildly enlarged stable. }</pre>	Sid 3: Structured abnormality query: <pre>{ anatomy: lungs finding: emphysema polarity: positive uncertainty: false severity: none laterality: none intent: verify source_sentence: Lungs are hyperexpanded but clear. }</pre>	
R2Gen: The soft tissues are normal. Both lungs are clear and expanded. Heart size and pulmonary vascularity appear within normal limits. Negative for pneumoperitoneum. No pneumothorax, pleural effusion, or focal airspace disease.		
M2KT: Visualized osseous structures appear intact. Low lung volumes. Normal cardiac contour. Mediastinum is unremarkable. Bibasilar atelectasis versus scarring. No large pleural effusions.		
R2GenGPT: Cardiomedastinal silhouette is within normal limits. Pulmonary vascularity is within normal limits. Lungs are clear bilaterally. No pneumothorax or pleural effusion. No acute osseous findings.		
Ours: Bony structures are intact. Stable mild heart enlargement. Mediastinal contours are maintained. Pulmonary vasculature is normal in caliber. Hyperinflated lungs. The pleural spaces are clear. Clear hyperexpanded lungs bilaterally with no pneumothorax or pleural effusion.		

Fig. 3 Qualitative comparison of report generation results. For a representative case, we present the ground-truth report, its sentence decomposition, and the structured abnormality queries derived from the positive sentences, together with the reports generated by R2Gen, M2KT, R2GenGPT, and SQV-R2Gen.

report contains two key positive findings, namely mild cardiomegaly and hyperexpanded lungs, which provide a suitable example for evaluating whether a model can preserve fine-grained abnormality semantics beyond generic normal descriptions.

R2Gen generates a largely normal-template report and fails to recover both mild heart enlargement and pulmonary hyperinflation, indicating limited sensitivity to subtle but clinically meaningful abnormalities. M2KT partially captures thoracic semantics, but its prediction drifts toward unrelated findings such as low lung volumes and bibasilar atelectatic or scarring patterns, revealing insufficient case-specific abnormality discrimination. R2GenGPT produces fluent and well-formed text, yet still tends to favor dominant normal expressions and misses the two positive findings, suggesting that strong language priors alone do not guarantee faithful abnormality grounding.

In contrast, SQV-R2Gen produces a report that is substantially more consistent with the reference. It successfully identifies stable mild heart enlargement and hyperinflated lungs, while also preserving normal mediastinal contour, normal pulmonary vascularity, and negative pleural findings. This advantage mainly comes from the proposed structured abnormality query construction and query-conditioned spatial evidence learning. By explicitly encoding abnormality attributes such as anatomy, finding category, polarity, and severity, and by injecting quality-verified local evidence into generation, SQV-R2Gen improves abnormality fidelity and yields clinically more faithful reports.

4.5 Limitations and Discussions

As shown in Table 4, the complete SQV-R2Gen achieves the best results on six of the eight reported metrics. Relative to the base model, it improves BLEU-1, BLEU-4, METEOR, and ROUGE-L from 0.458, 0.161, 0.193, and 0.363 to 0.520, 0.201, 0.217, and 0.399 on IU X-Ray, and from 0.352, 0.121, 0.146, and 0.274 to 0.413, 0.142, 0.172, and 0.286 on MIMIC-CXR. Structured abnormality queries improve all four metrics on both datasets, indicating that abnormality-level semantic units provide a more targeted supervisory interface than report-level generation alone. Direct distillation of raw teacher heatmaps yields further gains, confirming that query-conditioned responses contain useful local evidence.

Replacing raw responses with quality-verified evidence provides additional improvements; for example, BLEU-4 and METEOR on MIMIC-CXR increase from 0.136 and 0.163 to 0.139 and 0.168. This result indicates that unfiltered teacher responses contain noisy or spatially implausible activations and that anatomical and spatial verification produces more reliable supervision. Quality-aware weighting further improves performance by assigning supervisory strength according to estimated reliability rather than treating all retained heatmaps equally.

Adding disease-level semantic consistency improves BLEU-1, BLEU-4, and METEOR on both datasets but slightly reduces ROUGE-L from 0.400 to 0.399 on IU X-Ray and from 0.288 to 0.286 on MIMIC-CXR. This small trade-off is consistent with the different objectives: DSC regularizes the pooled representation toward the presence or absence of clinically salient observations, whereas ROUGE-L measures longest-subsequence agreement and is sensitive to lexical order and report structure. Because the decreases are only 0.001 and 0.002 while the remaining NLG metrics improve, the contribution of DSC should be interpreted jointly with the other language-generation and clinical-efficacy results.

Table 4 Ablation studies of SQV-R2Gen on IU X-Ray and MIMIC-CXR. SQ denotes structured abnormality query modeling. Raw denotes direct spatial evidence distillation using raw teacher-generated heatmaps. Ver. denotes quality verification. Wgt. denotes quality-aware weighting of the retained evidence. DSC denotes disease-level semantic consistency. The best results are shown in bold and the second-best results are underlined.

Variant	SQ	Raw	Ver.	Wgt.	DSC	IU X-Ray				MIMIC-CXR			
						BLEU-1	BLEU-4	METEOR	ROUGE-L	BLEU-1	BLEU-4	METEOR	ROUGE-L
Base						0.458	0.161	0.193	0.363	0.352	0.121	0.146	0.274
+SQ	✓					0.486	0.176	0.204	0.382	0.373	0.127	0.153	0.279
+SQ+Raw-SED	✓	✓				0.500	0.189	0.211	0.392	0.391	0.136	0.163	0.283
+SQ+Ver.-SED	✓		✓			0.506	0.193	0.214	0.396	0.402	0.139	0.168	0.287
+SQ+Ver.+Wgt.-SED	✓		✓	✓		<u>0.515</u>	<u>0.199</u>	<u>0.216</u>	0.400	<u>0.410</u>	<u>0.141</u>	<u>0.171</u>	0.288
Ours	✓		✓	✓	✓	0.520	0.201	0.217	<u>0.399</u>	0.413	0.142	0.172	<u>0.286</u>

Limitations. The proposed framework has several limitations. First, although BioViL-T is used only during one-time offline preprocessing and is not required during model optimization or inference, the quality of the cached supervision remains dependent on the teacher’s grounding capability. The quality verification mechanism can reject anatomically implausible or low-confidence responses, but it cannot recover a true lesion that is completely missed by the teacher. Second, the current implementation retains at most $K = 3$ heatmaps per sample. This fixed capacity simplifies batching but may omit secondary findings in examinations containing multiple abnormalities. Third, the evaluation is restricted to two chest X-ray datasets, and the generalization of the framework to other imaging modalities and anatomical regions remains to be established. Finally, report parsing, CheXpert-based label extraction, and quality scoring are automatic procedures and may introduce additional supervision noise. Direct lesion-level localization evaluation and radiologist assessment would provide stronger evidence for the clinical reliability and interpretability of the verified heatmaps.

4.6 Efficiency and Resource Analysis

To evaluate the practical deployment cost, we further compare SQV-R2Gen with the baseline model in terms of inference speed, GPU memory consumption, deployed model size, and report-generation quality. The comparison is conducted on the IU X-Ray test set under the same decoding setting. The proposed structured query construction, BioViL-T grounding, quality verification, disease classification branch, and map prediction branch are used only for offline supervision construction or training-time regularization. They are not involved in the deployed inference graph. Therefore, at test time, SQV-R2Gen follows the same image-to-report pathway as the baseline model, consisting of the visual encoder, the visual-to-language projection module, and the frozen Llama-2-7B decoder.

Table 5 Inference efficiency and resource comparison on the IU X-Ray test set. Deployed Params denotes the number of parameters retained in the test-time image-to-report model. Mem. denotes the peak allocated GPU memory during inference. Latency is the average inference time per generated report, and throughput is the corresponding number of reports processed per second.

Method	Deployed Params (B)	Mem. (GB)	Latency (s/report)	Throughput (reports/s)	Tokens/s	BLEU-4	ROUGE-L
Baseline model	6.83	20.64	0.826	1.211	75.65	0.146	0.362
SQV-R2Gen	6.78	18.35	0.695	1.439	77.93	0.201	0.399

Note: Structured query construction, BioViL-T grounding, quality verification, and auxiliary training heads are used only during training and are removed during inference. The deployed inference graph is identical in structure to the baseline model.

As shown in Table 5, SQV-R2Gen improves BLEU-4 and ROUGE-L from 0.146 and 0.362 to 0.201 and 0.399, respectively, without introducing teacher-side computation at test time. Compared with the baseline model, SQV-R2Gen reduces the peak

GPU memory from 20.64 GB to 18.35 GB and decreases the per-report latency from 0.826 s to 0.695 s. The corresponding throughput increases from 1.211 to 1.439 report-s/s. These results show that the proposed quality-verified spatial evidence distillation improves the training supervision rather than enlarging the deployed inference graph. Therefore, the performance gain of SQV-R2Gen is obtained with an online deployment cost comparable to the baseline image-to-report generator.

5 Conclusion

This paper presented SQV-R2Gen, a weakly supervised radiology report generation framework that improves report faithfulness through structured abnormality query modeling and quality-verified spatial evidence distillation. By converting free-text reports into structured abnormality queries and filtering noisy teacher-generated heatmaps with anatomical prior-guided quality verification, SQV-R2Gen provides more reliable local supervision and stronger disease-aware semantic constraints for training. Experiments on two chest X-ray benchmarks show that the proposed method achieves competitive language-generation performance and favorable clinical efficacy, highlighting the value of improving supervision reliability rather than solely increasing model capacity. The current results are limited to chest radiography, and future work will investigate adaptive evidence capacity, multi-teacher verification, and direct lesion-level clinical evaluation.

Declarations

Funding. This work was supported by the National Natural Science Foundation of China (No. 62206624, 62576288), the Postdoctoral Science Foundation of China (No. 2023M732428), and the Sichuan Science and Technology Program (No. 2025NSFSC0456).

Conflict of interest. The authors declare no competing interests.

Ethics approval and consent to participate. This study used only publicly available benchmark datasets. The authors did not collect new data from human participants or animals. Therefore, additional ethics approval and consent to participate were not applicable.

Data availability. The datasets used in this study are publicly available benchmark datasets. The IU X-Ray dataset and MIMIC-CXR dataset can be obtained from their original data sources subject to the corresponding access and data use agreements. No new dataset was generated in this study.

Code availability. The code is available at <https://github.com/WangYuan912/SQV-R2Gen>.

Author contributions. Yuan Wang contributed to the conceptualization, methodology, implementation, experiments, data analysis, and writing of the original draft. Xiaou Liu contributed to data processing, experimental validation, result analysis, and manuscript revision. Liming Xu contributed to conceptualization, supervision,

project administration, funding acquisition, and manuscript review and editing. Quan Tang contributed to methodological discussion and manuscript revision. Xianhua Zeng contributed to technical discussion and manuscript revision. Weisheng Li contributed to supervision, funding support, and manuscript review and editing.

References

- [1] Tanno, R., Barrett, D.G.T., Sellergren, A., Ghaisas, S., Dathathri, S., See, A., Welbl, J., Lau, C., Tu, T., Azizi, S.: Collaboration between clinicians and vision-language models in radiology report generation. *Nat. Med.* **31**(2), 599–608 (2025) <https://doi.org/10.1038/s41591-024-03302-1>
- [2] Wenderott, K., Krups, J., Zaruchas, F., Weigl, M.: Effects of artificial intelligence implementation on efficiency in medical imaging—a systematic literature review and meta-analysis. *npj Digit. Med.* **7**(1), 265 (2024) <https://doi.org/10.1038/s41746-024-01248-9>
- [3] Li, Y., Yang, B., Cheng, X., Zhu, Z., Li, H., Zou, Y.: Unify, align and refine: Multi-level semantic alignment for radiology report generation. In: *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, pp. 2863–2874 (2023). <https://doi.org/10.1109/ICCV51070.2023.00268>
- [4] Najdenkoska, I., Zhen, X., Worring, M., Shao, L.: Uncertainty-aware report generation for chest x-rays by variational topic inference. *Med. Image Anal.* **82**, 102603–102616 (2022) <https://doi.org/10.1016/j.media.2022.102603>
- [5] Xue, Y., Xu, T., Rodney Long, L., Xue, Z., Antani, S., Thoma, G.R., Huang, X.: Multimodal recurrent model with attention for automated radiology report generation. In: *Int. Conf. Med. Image Comput. Comput.-Assist. Interv.*, pp. 457–466 (2018)
- [6] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: *Int. Conf. Learn. Represent.* (2021)
- [7] Tang, Q., Xu, L., Wang, Y., Zheng, B., Lv, J., Zeng, X., Li, W.: Dual-modality visual feature flow for medical report generation. *Med. Image Anal.* **101**, 103413–103424 (2025) <https://doi.org/10.1016/j.media.2024.103413>
- [8] Liu, T., Wang, J., Hu, Y., Li, M., Yi, J., Chang, X., Gao, J., Yin, B.: Hc-llm: Historical-constrained large language models for radiology report generation. In: *Proc. AAAI Conf. Artif. Intell.*, vol. 39, pp. 5595–5603 (2025)
- [9] Zhang, J., Shen, X., Wan, S., Goudos, S.K., Wu, J., Cheng, M., Zhang, W.: A novel deep learning model for medical report generation by inter-intra information calibration. *IEEE J. Biomed. Health Inform.* **27**(10), 5110–5121 (2023)

- [10] Mei, X., Yang, L., Gao, D., Cai, X., Han, J., Liu, T.: Adaptive medical topic learning for enhanced fine-grained cross-modal alignment in medical report generation. *IEEE Trans. Multimedia* **27**, 5050–5061 (2025)
- [11] Huang, X., Chen, W., Liu, J., Lu, Q., Luo, X., Shen, L.: Damper: A dual-stage medical report generation framework with coarse-grained mesh alignment and fine-grained hypergraph matching. In: *Proc. AAAI Conf. Artif. Intell.*, vol. 39, pp. 3769–3778 (2025)
- [12] Sloan, P., Clatworthy, P., Simpson, E., Mirmehdi, M.: Automated radiology report generation: A review of recent advances. *IEEE Rev. Biomed. Eng.* **18**, 368–387 (2024) <https://doi.org/10.1109/RBME.2024.3408456>
- [13] Chen, Z., Song, Y., Chang, T.-H., Wan, X.: Generating Radiology Reports via Memory-driven Transformer. In: *Proc. Conf. Empirical Methods Nat. Lang. Process.*, pp. 1439–1449 (2020). <https://doi.org/10.18653/v1/2020.emnlp-main.112>
- [14] Chen, Z., Shen, Y., Song, Y., Wan, X.: Cross-modal Memory Networks for Radiology Report Generation. In: *Proc. Annu. Meeting Assoc. Comput. Linguist.*, pp. 5904–5914 (2021)
- [15] Yan, B., Pei, M., Zhao, M., Shan, C., Tian, Z.: Prior guided transformer for accurate radiology reports generation. *IEEE J. Biomed. Health Inform.* **26**(11), 5631–5640 (2022) <https://doi.org/10.1109/JBHI.2022.3197162>
- [16] Yang, S., Wu, X., Ge, S., Zheng, Z., Zhou, S.K., Xiao, L.: Radiology report generation with a learned knowledge base and multi-modal alignment. *Med. Image Anal.* **86**, 102798–102808 (2023) <https://doi.org/10.1016/j.media.2023.102798>
- [17] Yi, X., Fu, Y., Liu, R., Zhang, H., Hua, R.: Tsget: two-stage global enhanced transformer for automatic radiology report generation. *IEEE J. Biomed. Health Inform.* **28**(4), 2152–2162 (2024) <https://doi.org/10.1109/JBHI.2024.3350077>
- [18] 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) <https://doi.org/10.1109/TMI.2024.3372638>
- [19] Wang, J., Bhalerao, A., Yin, T., See, S., He, Y.: CAMANet: class activation map guided attention network for radiology report generation. *IEEE J. Biomed. Health Inform.* **28**(4), 2199–2210 (2024)
- [20] Liu, Z., Zhu, Z., Zheng, S., Zhao, Y., He, K., Zhao, Y.: From observation to concept: A flexible multi-view paradigm for medical report generation. *IEEE Transactions on Multimedia* **26**, 5987–5995 (2024) <https://doi.org/10.1109/TMM.2023.3342691>

- [21] Zhang, K., Jiang, H., Zhang, J., Huang, Q., Fan, J., Yu, J., Han, W.: Semi-supervised medical report generation via graph-guided hybrid feature consistency. *IEEE Transactions on Multimedia* **26**, 904–915 (2024) <https://doi.org/10.1109/TMM.2023.3273390>
- [22] Liu, F., Wu, X., Ge, S., Fan, W., Zou, Y.: Exploring and distilling posterior and prior knowledge for radiology report generation. In: *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, pp. 13753–13762 (2021). <https://doi.org/10.1109/CVPR46437.2021.01354>
- [23] Zhang, J., Shen, X., Wan, S., Goudos, S.K., Wu, J., Cheng, M., Zhang, W.: A novel deep learning model for medical report generation by inter-intra information calibration. *IEEE J. Biomed. Health Inform.* **27**(10), 5110–5121 (2023) <https://doi.org/10.1109/JBHI.2023.3236661>
- [24] Wang, Z., Liu, L., Wang, L., Zhou, L.: R2GenGPT: Radiology report generation with frozen LLMs. *Meta-Radiol.* **1**(3), 100033–100040 (2023) <https://doi.org/10.1016/j.metrad.2023.100033>
- [25] Liu, C., Tian, Y., Chen, W., Song, Y., Zhang, Y.: Bootstrapping large language models for radiology report generation. In: *Proc. AAAI Conf. Artif. Intell.*, vol. 38, pp. 18635–18643 (2024)
- [26] Song, S., Tang, H., Yang, H., Li, X.: Ddatr: Dynamic difference-aware temporal residual network for longitudinal radiology report generation. *IEEE Trans. Med. Imaging* (2025) <https://doi.org/10.1109/TMI.2025.3591364>
- [27] Chen, W., Liu, Y., Wang, C., Zhu, J., Li, G., Liu, C.-L., Lin, L.: Cross-modal causal representation learning for radiology report generation. *IEEE Trans. Image Process.* (2025) <https://doi.org/10.1109/TIP.2025.3568746>
- [28] 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: *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 4967–4975 (2024). <https://doi.org/10.1145/3664647.3681201>
- [29] Sun, Y., Lee, Y.Z., Woodard, G.A., Zhu, H., Lian, C., Liu, M.: R2Gen-Mamba: A selective state space model for radiology report generation. In: *Proc. 2025 IEEE 22nd Int. Symp. Biomed. Imaging (ISBI)*, pp. 1–4 (2025). <https://doi.org/10.1109/ISBI60581.2025.10980814>
- [30] Liang, Z., Hou, X., Chen, F., Zhang, Y.: AAMN: Cross-modal fusion network with association alignment matrix for radiological report generation. *Multimed. Syst.* **32**, 253 (2026) <https://doi.org/10.1007/s00530-026-02337-3>
- [31] Peng, P., Fan, W., Shen, Y., Liu, W., Yang, X., Zhang, Q., Wei, X., Zhou, D.: Eye gaze guided cross-modal alignment network for radiology report generation.

- IEEE J. Biomed. Health Inform. **28**(12), 7406–7419 (2024) <https://doi.org/10.1109/JBHI.2024.3422168>
- [32] Ma, C., Jiang, H., Chen, W., Li, Y., Wu, Z., Yu, X., Liu, Z., Guo, L., Zhu, D., Zhang, T., *et al.*: Eye-gaze guided multi-modal alignment for medical representation learning. *Adv. Neural Inf. Process. Syst.* **37**, 6126–6153 (2024)
 - [33] Huy, T.D., Huynh, D.A., Xie, Y., Qi, Y., Chen, Q., Le Nguyen, P., Tran, S.K., Phung, S.L., Hengel, A., Liao, Z., *et al.*: Seeing the trees for the forest: rethinking weakly-supervised medical visual grounding. In: *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, pp. 24445–24455 (2025)
 - [34] Vilouras, K., Stogiannidis, I., Yan, J., O’Neil, A.Q., Tsaftaris, S.A.: Anatomy-grounded weakly supervised prompt tuning for chest x-ray latent diffusion models. *arXiv preprint arXiv:2506.10633* (2025)
 - [35] Bannur, S., Bouzid, K., Castro, D.C., Schwaighofer, A., Thieme, A., Bond-Taylor, S., Ilse, M., Pérez-García, F., Salvatelli, V., Sharma, H., *et al.*: MAIRA-2: Grounded radiology report generation. *arXiv preprint arXiv:2406.04449* (2024)
 - [36] Aas-Alas, M., Albiol, A., Paredes, R.: Does grounding improve radiology report generation? an empirical study on padchest-gr. In: *Med. Imaging Deep Learn.* (2026)
 - [37] Bannur, S., Hyland, S., Liu, Q., Pérez-García, F., Ilse, M., Castro, D.C., Boecking, B., Sharma, H., Bouzid, K., Thieme, A., Schwaighofer, A., Wetscherek, M., Lungren, M.P., Nori, A., Alvarez-Valle, J., Oktay, O.: Learning to exploit temporal structure for biomedical vision-language processing. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15016–15027 (2023). <https://doi.org/10.1109/CVPR52729.2023.01442>
 - [38] Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., Marklund, H., Haghighi, B., Ball, R., Shpanskaya, K., *et al.*: Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In: *Proc. AAAI Conf. Artif. Intell.*, vol. 33, pp. 590–597 (2019). <https://doi.org/10.1609/aaai.v33i01.3301590>
 - [39] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10012–10022 (2021). <https://doi.org/10.1109/ICCV48922.2021.00986>
 - [40] Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., *et al.*: Llama 2: open foundation and fine-tuned chat models. *arxiv. arXiv preprint arXiv:2307.09288* **10** (2023)
 - [41] 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) <https://doi.org/10.1093/jamia/ocv080>
- [42] Johnson, A.E., 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) <https://doi.org/10.1038/s41597-019-0322-0>
- [43] Qin, H., Song, Y.: Reinforced cross-modal alignment for radiology report generation. In: *Findings Assoc. Comput. Linguist.: ACL 2022*, pp. 448–458 (2022)
- [44] Yan, B., Pei, M., Zhao, M., Shan, C., Tian, Z.: Prior guided transformer for accurate radiology reports generation. *IEEE J. Biomed. Health Inform.* **26**(11), 5631–5640 (2022) <https://doi.org/10.1109/JBHI.2022.3197162>
- [45] Papineni, K., Roukos, S., Ward, T., Zhu, W.-J.: Bleu: a method for automatic evaluation of machine translation. In: *Proc. 40th Annu. Meet. Assoc. Comput. Linguist.*, pp. 311–318 (2002)
- [46] Denkowski, M., Lavie, A.: Meteor 1.3: Automatic metric for reliable optimization and evaluation of machine translation systems. In: *Proc. Sixth Workshop Stat. Mach. Transl.*, pp. 85–91 (2011)
- [47] Lin, C.-Y.: Rouge: A package for automatic evaluation of summaries. In: *Text Summarization Branches Out*, pp. 74–81 (2004)
- [48] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. *Advances in neural information processing systems* **30** (2017)
- [49] Zheng, Z., Zhang, Y., Liang, E., Weng, Z., Chai, J., Li, J.: Trinet: Team role interaction network for automatic radiology report generation. *Comput. Biol. Med.* **183**, 109275 (2024) <https://doi.org/10.1016/j.compbiomed.2024.109275>