Jurnal Sains dan Kesehatan (JUSIKA)
Vo.7, No.1 Juni 2023
e-ISSN: 2962-1569; p-ISSN: 2580-8362, Hal 53-74
DOI: https://doi.org/10.57214/jusika.v7i1.1127

Evidence-Grounded Chest X-ray Report Generation with Retrieval, Citation, and Hallucination Control

Danang[1*], Toni Wijanarko Adi Putra[2]
[1-2]Universitas Sains dan Teknologi Komputer, Indonesia
Email: danang150787@gmail.com [1], toni.wijanarko@stekom.ac.id [2]

**Corresponding Author: danang150787@gmail.com*

Abstract: Chest X-ray report generation has become an important topic in vision-to-language research. However, fully generative models often create fluent clinical reports that may contain unsupported or inaccurate statements, leading to hallucination problems and reducing reliability. This study investigates evidence-grounded report generation using the Open-i (IU X-ray) dataset with two main goals: generating coherent radiology reports from X-ray images and minimizing unsupported clinical entities through evidence retrieval. Four experimental models were evaluated: a baseline image-to-report model (E1), an alignment-enhanced model using the InfoNCE objective (E2), a retrieval-grounded model that incorporates Top-K evidence sentences with citation markers such as [E#] (E3), and a reranking model that selects the most evidence-supported output (E4). Experimental results on the Open-i test dataset show that grounding methods significantly reduce hallucination rates and improve entity overlap performance. The reranking approach achieves the best grounding quality, although stronger grounding slightly lowers text-overlap scores and increases inference time. Overall, retrieval-based grounding with explicit citations and reranking offers an effective approach for improving factual consistency and reducing unsupported information in automated radiology report generation.

Keywords: chest X-ray; evidence grounding; hallucination control; radiology report generation; retrieval-augmented generation.

1. INTRODUCTION

Radiology reporting remains a high-impact documentation workflow: it converts complex visual evidence into structured clinical language that is used for downstream communication, triage, and longitudinal comparison. Among routine modalities, chest X-ray (CXR) is one of the most common exams in many clinical settings, making it a natural target for automation. Consequently, automatic radiology report generation has attracted substantial attention as a multimodal sequence modeling problem, historically aligned with image captioning but requiring far more domain-specific phrasing, negation handling, and clinically meaningful content selection (Chen et al., 2020, 2021; Jing et al., 2018; Li et al., 2018; Liu et al., 2021). Modern architectures typically combine a convolutional image encoder with an autoregressive language decoder, with variations that add memory modules, cross-modal attention, or explicit disease cues to improve coverage of findings and reduce generic outputs (Chen et al., 2020, 2021; He et al., 2016; Liu et al., 2021; Vaswani et al., 2017).

A persistent failure mode in this setting is hallucination: the generator produces clinically plausible findings that are not supported by the image (or contradict the evidence), which is especially problematic in high-stakes domains. Hallucination in natural language generation is widely recognized as a systematic risk, and in clinical reporting it can manifest as (i) invented abnormalities, (ii) incorrect laterality or anatomy, (iii) wrong severity qualifiers,

Received April 30, 2023; Revised Mei 12, 2023; Accepted Juni 01, 2023
** Danang, danang150787@gmail.com*

or (iv) omission of uncertainty/negation cues, all of which can mislead users who over-trust fluent text (Bender et al., 2021; Ji et al., 2023). Importantly, even when standard n-gram metrics look reasonable, the text can remain clinically unreliable if key entities are unsupported; therefore, controlling hallucination requires mechanisms beyond improving surface-level similarity.

This paper focuses on evidence-grounded report generation on the Open-i (IU X-ray) dataset (Demner-Fushman et al., 2016), a widely used benchmark alongside larger datasets such as ChestX-ray8 (Wang et al., 2017), CheXpert (Irvin et al., 2019), and MIMIC-CXR (Johnson et al., 2019). While larger datasets enable scale, IU X-ray remains valuable for controlled experimentation because it provides paired images and reports and has been widely used to compare modeling choices (Demner-Fushman et al., 2016). Our goal is not merely to generate f luent reports, but to make the model's informational basis explicit so that unsupported content can be detected and discouraged.

discouraged. Instead of relying solely on an autoregressive decoder to internalize clinical knowledge, we explicitly retrieve relevant evidence sentences from training reports and require the generator to cite them in the output. This is conceptually aligned with retrieval-augmented generation: retrieving a small set of textual contexts and then conditioning generation on that context to improve factuality and controllability (Izacard & Grave, 2021; Lewis et al., 2020). In our setting, the retrieved contexts are not general web passages but domain-specific sentences extracted from the training report corpus, which function as an evidence pool. By attaching identifiers [E1]..[EK] to the retrieved sentences and training the generator to emit citations when stating findings, we turn grounding into a concrete constraint that can be verified post-hoc.

Approach overview.: Concretely, our generator is conditioned on (i) visual features from a standard image encoder and (ii) a compact list of retrieved evidence sentences. The architecture follows a practical encoder–decoder pattern (e.g., ResNet-style visual features and Transformer decoding) implemented in PyTorch (He et al., 2016; Paszke et al., 2019; Vaswani et al., 2017). We additionally study a contrastive alignment stage that encourages images and reports to share a common embedding space using an InfoNCE-style objective (Oord et al., 2018). This step is motivated by evidence from multimodal representation learning that better aligned embeddings can improve retrieval quality and downstream generation (Nguyen et al., 2022; Radford et al., 2021). For retrieval itself, we employ dense retrieval in the learned embedding space (e.g., CLIP-style or biomedical VLP embeddings) to obtain top-K evidence candidates per image (Nguyen et al., 2022; Radford et al., 2021).

Hallucination control via reranking.: Citations make grounding checkable, but they do not automatically guarantee correctness: a model can still attach citations incorrectly or omit citations for risky claims. To further reduce unsupported entities, we apply reranking over beam candidates with a groundedness proxy that rewards better alignment between generated clinical entities and the retrieved evidence. This trades extra computation at inference time for lower hallucination risk, enabling a controllable quality–groundedness–latency trade-off.

a. Evaluation and measurement.: We evaluate generation quality with standard text-overlap metrics commonly used in report generation studies, including BLEU, ROUGE, and CIDEr-style proxies (Lin, 2004; Papineni et al., 2002; Vedantam et al., 2015). However, because overlap metrics can miss clinical factuality, we also measure entity overlap and a proxy hallucination rate at the entity level. To operationalize entity-centric evaluation, we draw on established clinical text structuring tools such as CheXbert-style report labeling and RadGraph-style entity/relationship extraction as conceptual anchors for what "clinical entities" mean in radiology text (Boag et al., 2020; Jain et al., 2021). We emphasize that our hallucination measurement is a proxy designed for scalable comparison across experimental variants, rather than a substitute for expert clinical adjudication.
b. Contributions.: We make three contributions:
    a) A four-stage experimental pipeline (E1–E4) that isolates the effect of (i) baseline image-to-text generation, (ii) multimodal contrastive alignment, (iii) retrieval-based evidence grounding with explicit citations, and (iv) reranking-based hallucination control (Izacard & Grave, 2021; Lewis et al., 2020; Oord et al., 2018).
    b) An evidence pool and dense retrieval index built from training reports, enabling top-K evidence retrieval per image and citation-constrained prompting for generation (Nguyen et al., 2022; Radford et al., 2021).
    c) An evaluation suite that combines standard text metrics (BLEU/ROUGE/CIDEr proxy) with entity overlap and an entity-level hallucination-rate proxy, highlighting the quality–groundedness–latency trade-off (Ji et al., 2023; Lin, 2004; Papineni et al., 2002; Vedantam et al., 2015).
c. Scope and non-clinical use.: This is a research prototype trained on public datasets; it is not a clinical device and must not be used for diagnosis or treatment decisions. As discussed broadly in the context of language technologies, misuse and over-reliance can cause harm, particularly when fluent text is mistaken for verified evidence (Bender et al., 2021). We therefore position the contribution as evidence-grounded generation and measurement for research benchmarking, not clinical decision support.

2. LITERATURE REVIEW

This section provides a state-of-the-art overview and clarifies how our work differs from prior approaches.

Radiology report generation

Automatic radiology report generation is commonly formulated as an encoder–decoder problem that maps an image (or multi-view images) into a long-form narrative report (Jing et al., 2018; Li et al., 2018). Early and classical deep-learning formulations adapt image captioning to clinical language, but radiology reports differ materially from generic captions: they are longer, contain structured rhetorical patterns (e.g., Findings and Impression), require careful handling of negation and uncertainty, and must preserve clinically meaningful entities and relations (Jing et al., 2018; Li et al., 2018). Typical pipelines therefore combine a CNN image encoder (often ResNet-like) and an autoregressive text decoder with attention, with training driven by maximum-likelihood objectives (He et al., 2016; Jing et al., 2018).

Subsequent work improved report coherence and clinical phrasing by strengthening the decoder and introducing explicit memory and cross-modal context. Memory-driven Transformers augment decoding with a learned memory of report patterns and salient tokens, improving the ability to mention common findings in a more structured way (Chen et al., 2020). Cross-modal memory networks extend this idea by building a memory that aligns visual cues with textual patterns, aiming to reduce generic phrasing and improve coverage of clinically relevant findings (Chen et al., 2021). Multi-modal Transformer variants further improve cross-attention between image features and text, providing stronger modeling capacity for long reports and multi-view inputs (Liu et al., 2021).

Despite these advances, two recurring issues remain. First, improvements in standard overlap metrics do not necessarily translate to improved factual correctness; models can still generate fluent but incorrect statements. Second, most report generators are implicitly grounded: they condition on image features but do not expose which textual evidence supports each statement. As a result, it is difficult to verify why a model produced a specific finding, and post-hoc auditing typically relies on manual review or external labelers rather than explicit evidence links. Our work targets this gap by adding retrieval-based evidence sentences and requiring explicit citations, which makes grounding checkable and supports groundedness-aware inference controls.

Grounded generation, alignment, and factuality evaluation

Retrieval-augmented generation (RAG) is a prominent strategy to reduce hallucination by conditioning generation on retrieved text, thereby making relevant information available at

inference time rather than requiring the model to memorize it in parameters (Izacard & Grave, 2021; Lewis et al., 2020). While RAG is often discussed in knowledge intensive NLP, the same principle applies to clinical generation: retrieving task-relevant text can constrain outputs, improve controllability, and enable traceability. In our setting, retrieval is applied to a domain-specific evidence pool built from training reports, allowing the generator to condition on a compact set of candidate evidence sentences and attach citations to them.

In vision–language settings, representation learning and alignment are central to retrieval quality. Contrastive learning objectives such as InfoNCE encourage consistent embeddings across modalities, which can improve both retrieval and downstream generation when paired data are available (Oord et al., 2018). General-purpose visionlanguage pretraining, exemplified by CLIP, demonstrates that large-scale contrastive pretraining yields transferable embeddings suitable for image–text retrieval and zero-shot transfer (Radford et al., 2021). For medical imaging, radiology-specific vision–language pretraining approaches such as BioViL aim to learn embeddings that better reflect biomedical semantics and can strengthen downstream tasks under domain shift (Nguyen et al., 2022). Motivated by these trends, our pipeline includes an explicit alignment stage and uses dense retrieval to obtain top-K evidence sentences per image.

However, introducing retrieval also introduces practical trade-offs. Retrieval increases inference latency (index lookup and additional conditioning tokens), and retrieved contexts can be partially irrelevant or contain distractor phrases, which can degrade fluency or even introduce new errors if the generator copies inappropriate evidence. These considerations motivate grounding-aware controls: rather than trusting retrieval blindly, the system should measure whether generated entities are supported by the retrieved evidence and preferentially select candidates with better support.

Evaluating factuality for radiology text remains challenging. Standard text-overlap metrics such as BLEU, ROUGE, and CIDEr quantify surface similarity and are widely reported in report generation (Lin, 2004; Papineni et al., 2002; Vedantam et al., 2015). However, they often correlate weakly with clinical correctness because many semantically distinct reports share similar phrasing, and because n-gram overlap can be inflated by common boilerplate language. As a consequence, a model can score well while still hallucinating clinically important entities.

Entity-based and label-based evaluation proxies provide more targeted signals. CheXbert-style labelers produce structured disease labels from free-text reports, enabling comparisons at the level of clinically meaningful conditions rather than n-grams (Boag et al., 2020). RadGraph provides a complementary approach by extracting entities and relations,

offering a finer-grained representation of clinical content that can be used to measure entity overlap or entity-level inconsistencies (Jain et al., 2021). In our work, we report standard n-gram metrics alongside entity overlap and an entity-level hallucination proxy, enabling a clearer view of the quality–groundedness trade-off introduced by retrieval and citation constraints.

How our work differs.: In contrast to prior encoder–decoder report generators that rely on implicit grounding, our method makes the conditioning information explicit by retrieving evidence sentences and requiring citations in the generated output. Compared to generic RAG, our retrieval targets a curated evidence pool from radiology reports and is paired with groundedness-aware reranking to mitigate retrieval noise and reduce unsupported entities (Izacard & Grave, 2021; Lewis et al., 2020). Compared to purely metric-driven reporting, our evaluation emphasizes entity-centric proxies to better capture factuality and hallucination behaviors (Boag et al., 2020; Jain et al., 2021; Ji et al., 2023).

3. PROPOSED METHOD

Problem setup

Given a chest X-ray study (single frontal view or paired frontal–lateral views), we denote the visual input as I and the target free-text radiology report as $\boldsymbol{Y} = (\boldsymbol{y_1}, \ldots, \boldsymbol{y_T})$, typically containing a Findings section and (optionally) an Impression. The learning objective is to estimate a conditional language model $p\theta(Y \mid I)$ and decode a report that is both fluent and clinically supported by the available evidence. We evaluate four variants (E1–E4) under an identical data split and preprocessing pipeline so that changes in behavior can be attributed to the introduced components rather than data differences.

Our central research question is: how can we reduce unsupported clinical content in generated reports without sacrificing too much fluency or incurring prohibitive inference cost? To answer this, we progressively add (i) a baseline image-to-text generator (E1), (ii) an alignment objective to improve cross-modal representation consistency (E2), (iii) retrieval-based evidence grounding with explicit citations (E3), and (iv) groundedness-aware reranking to suppress unsupported entities (E4).

Data and evidence pool construction

We use the Open-i (IU X-ray) dataset, which includes paired radiology reports and corresponding chest X-ray images (frontal and optionally lateral views) (Demner-Fushman et al., 2016). We follow a fixed train/validation/test split and apply a consistent text normalization

procedure (e.g., whitespace normalization, punctuation standardization, and section parsing into Findings/Impression when available). Because our grounding mechanism operates at the sentence level, we construct an evidence pool from training reports only to avoid test leakage.

a. Evidence extraction.: From each training report, we split Findings and Impression into sentences using rule-based segmentation. We then filter out very short fragments (e.g., headings, isolated tokens, or sentence stubs), normalize punctuation and casing, and deduplicate by normalized text. Each evidence sentence is stored as ej with: (i) a stable identifier (used later as [E#]), and (ii) an empirical frequency count fj indicating how often a sentence (after normalization) appears in the training corpus. The frequency can be used for diagnostic analysis (e.g., to quantify common boilerplate and long-tail rare sentences) and for optional retrieval heuristics if needed.
b. Why sentence-level evidence?: Sentence-level units are a pragmatic compromise: they are short enough to be injected into a prompt context budget, yet often capture complete clinical assertions (e.g., "No pleural effusion."). This enables checkable grounding at inference time: claims in the generated report should be accompanied by at least one citation to a retrieved evidence sentence.

Baseline image-to-report model (E1)

The baseline follows a standard encoder–decoder design widely used in long-form radiology report generation:

a. Image encoder. A ResNet-50 backbone extracts visual features from the chest X-ray image(s), producing an image feature sequence HI (He et al., 2016). For multi-view studies, features can be concatenated or pooled into a unified representation (we keep the treatment fixed across E1–E4).
b. Text decoder. A Transformer decoder generates tokens autoregressively with cross-attention over HI (Vaswani et al., 2017). The decoder predicts a distribution over the vocabulary at each step and is trained with teacher forcing.

Let $\mathbf{y}\ \boldsymbol{y_{<t}}$ denote previously generated tokens. We minimize the token-level cross-entropy:

$$\boldsymbol{L_C} = -\sum_{\boldsymbol{t=1}}^{\boldsymbol{T}} \boldsymbol{lop_\theta}\,(\boldsymbol{y_{<t}}, \boldsymbol{I}) \tag{1}$$

At inference time, we decode with beam search to produce fluent long-form reports. This baseline provides a reference point for (i) surface-level text quality and (ii) the tendency to introduce clinically plausible but unsupported content.

Multimodal alignment (E2)

While E1 conditions on visual features, the learned image and text representations can still be weakly aligned, which can harm both retrieval quality (later in E3) and content selection during decoding. To encourage consistent image–text representations, we add a contrastive alignment loss over a minibatch of paired samples.

a. Embeddings and projection heads.: Let h_I be a pooled image representation derived from H_I, and let h_Y be a pooled text representation derived from the decoder (or a lightweight text encoder applied to the ground-truth report). We apply projection heads and $\ell 2$ normalization to obtain v_I and v_I. Using temperature τ, the InfoNCE loss is:

$$\boldsymbol{ain} = -\frac{\mathbf{1}}{\boldsymbol{B}}\sum_{\boldsymbol{k=1}}^{\boldsymbol{B}} \quad \boldsymbol{lo}\ \frac{\boldsymbol{ex}\left(\langle \boldsymbol{v_i^{(k)}}, \boldsymbol{v_t^{(k)}}\rangle/\boldsymbol{\tau}\right)}{\sum_{\boldsymbol{m=1}}^{\boldsymbol{B}} \quad \boldsymbol{ex}\left(\langle \boldsymbol{v_i^{(k)}}, \boldsymbol{v_t^{(m)}}\rangle/\boldsymbol{\tau}\right)} \tag{2}$$

This objective follows the widely used contrastive formulation for representation consistency (Oord et al., 2018), and is conceptually related to vision–language pretraining approaches that learn transferable aligned embeddings (Nguyen et al., 2022; Radford et al., 2021).

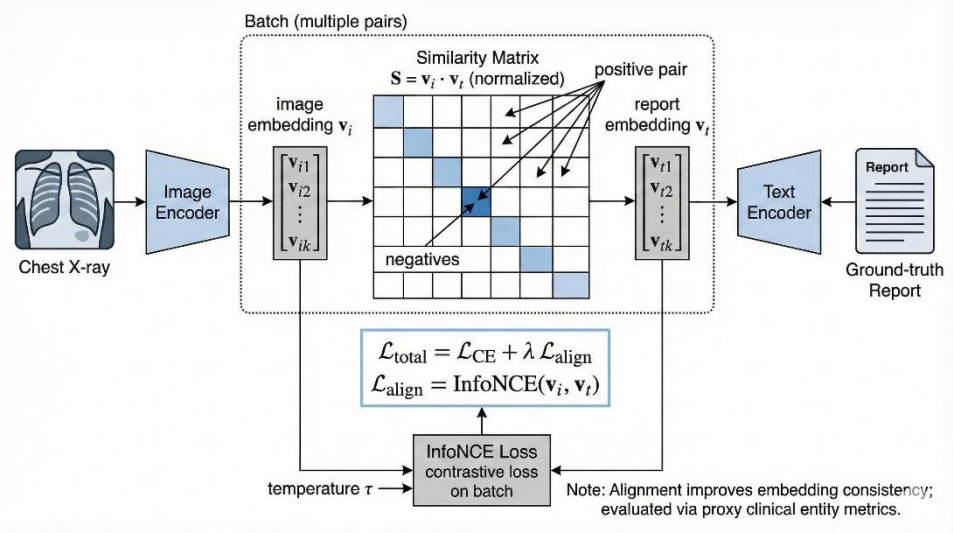

Figure 1. Multimodal alignment objective (E2): InfoNCE over image and report embeddings improves representation.

b. Total training objective.: We optimize the combined objective:

$$\boldsymbol{ttl} = \boldsymbol{L_C} + \boldsymbol{\lambda ain} \tag{3}$$

Intuitively, LCE trains the generator to produce fluent reports, while Lalign shapes a cross-modal embedding space that later supports dense retrieval and improves representation robustness. As illustrated in Figure 1, the alignment stage pulls matched

image–report pairs together in embedding space while pushing mismatched pairs apart, improving retrieval consistency and supporting better downstream groundedness proxies.

Evidence retrieval and citation-constrained prompting (E3)

E3 introduces retrieval-based grounding by conditioning generation on a compact set of evidence sentences extracted from training reports, in the spirit of retrieval-augmented generation (Izacard & Grave, 2021; Lewis et al., 2020). The key design choice is to make the conditioning information explicit and auditable: retrieved sentences are presented to the generator with identifiers, and the generator is instructed to cite them when stating findings.

a) Dense indexing of evidence sentences.: We embed each evidence sentence ej with a sentence encoder g(·) and build a dense index over the evidence pool. For a given input image I, we compute a query embedding q(I) derived from the image encoder (and optionally the aligned embedding head from E2). We retrieve the top-K evidence sentences by cosine similarity:

$$\boldsymbol{score}\left(\boldsymbol{I}, \boldsymbol{e_j}\right) = \frac{\langle \boldsymbol{q}(\boldsymbol{I}), \boldsymbol{g}\left(\boldsymbol{e_j}\right)\rangle}{|\boldsymbol{q}(\boldsymbol{I})||\boldsymbol{g}\left(\boldsymbol{e_j}\right)|} \quad (4)$$

and select ***and select*** A $\boldsymbol{E}(\boldsymbol{I}) = \{\boldsymbol{e}_{(\mathbf{1})}, \ldots, \boldsymbol{e}_{(\boldsymbol{K})}\}$

Prompt builder and citation constraint.: A prompt builder serializes E(I) into a short context list where each sentence is prefixed with an identifier [E1]..[EK]. The decoder is instructed to include citations [E#] when stating f indings (e.g., “No pleural effusion [E3].”). This converts grounding into a verifiable interface: we can check whether risky claims are accompanied by citations and whether generated entities are present in the cited evidence.

As shown in Figure 2, the E3 pipeline retrieves top-K evidence sentences from the training-derived evidence pool, injects them into the generation context, and enforces citation-aware generation to make support explicit.

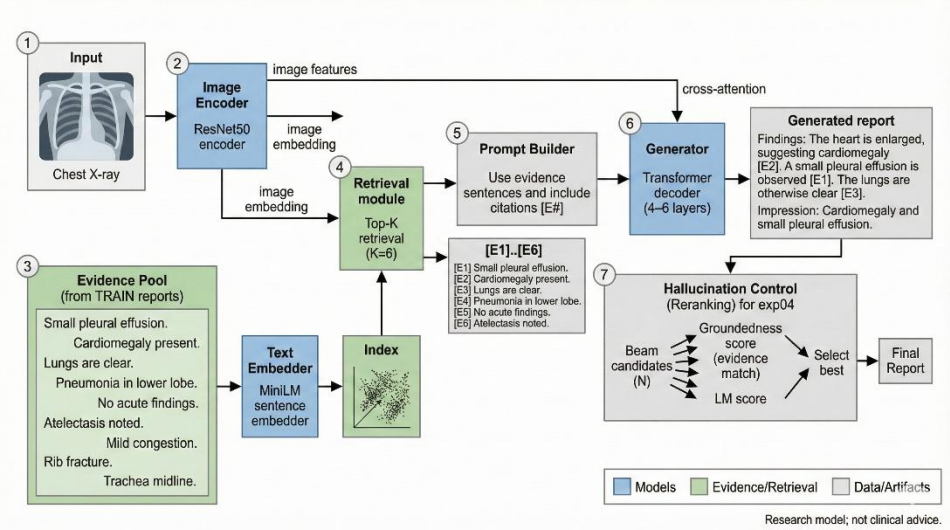

Figure 2. Evidence-grounded report generation: retrieve top-K evidence sentences from training reports, inject them into.

c)Training vs. inference for retrieval.: To train citation behavior robustly, it is beneficial to ensure that the evidence context is relevant to the target report. In practice, this can be implemented by using high-relevance evidence contexts during training (e.g., selecting evidence that best matches the ground-truth text) while using image-based retrieval at inference time. We keep the mechanism fixed across experiments and report the resulting trade-offs in both text quality and groundedness.

Hallucination control via reranking (E4)

Evidence grounding improves auditability, but hallucination can still occur if the decoder ignores evidence, over generalizes from common patterns, or attaches citations incorrectly. We therefore add an explicit reranking stage to suppress unsupported entities by preferring candidates that are better supported by the retrieved evidence.

a. Candidate generation.: We decode N candidates $\{\hat{Y}^{(n)}\}$ using beam search (or diverse beam variants). Each candidate has a language-model score $S_{LM}(\hat{Y})$ computed from token log-probabilities under the decoder.
b. Groundedness proxy.: We define a groundedness score $S_{grd}(\hat{Y},E(I))$ that rewards alignment between clinical entities mentioned in $\hat{Y}$ and entities present in the retrieved evidence set. Entity extraction is implemented via lightweight clinical NLP proxies such as CheXbert-style labelers and/or RadGraph-style entity extraction (Boag et al., 2020; Jain et al., 2021). We can additionally penalize (i) entities that appear in $\hat{Y}$ but not in evidence, and (ii) high-risk sentences lacking any citation, yielding a pragmatic hallucination-reduction heuristic.

c. Final scoring.: We combine fluency and groundedness:

$$S(\hat{Y}) = \alpha S_{LM}(\hat{Y}) + (1-\alpha) S_{grd}\left(\hat{Y}, E(I)\right) \quad (5)$$

The highest-scoring candidate is returned as the final report. This design exposes an explicit control knob α: smaller α prioritizes groundedness (lower hallucination proxy) at the cost of potentially reduced fluency, while larger α favors natural language likelihood.

d. Implementation note.: All components are implemented in PyTorch to ensure reproducibility and consistent training/inference behavior across E1–E4 (Paszke et al., 2019).

4. RESULT AND DISCUSSION

Experimental setup

Hardware/Software. All experiments were executed in a controlled Colab environment to ensure consistent reproducibility across variants. Training and inference were run on Google Colab with a Tesla T4 GPU (CUDA) for model optimization and GPU latency measurements, and we additionally report a CPU baseline to approximate lower-resource deployment conditions. The implementation is based on PyTorch (Paszke et al., 2019) and standard Transformer components (Vaswani et al., 2017), with fixed random seeds where applicable to reduce run-to-run variance. We keep the same preprocessing, tokenization, maximum sequence length, decoding configuration (beam size), and evaluation scripts across E1–E4 unless explicitly stated, so differences in results can be attributed to architectural or inference-time changes rather than inconsistent experimental conditions.

Dataset and splits. We use the Open-i (IU X-ray) dataset (Demner-Fushman et al., 2016), which provides paired chest X-ray images and radiology reports. Following common practice, we filter samples with missing images, invalid image files, or malformed report text. After preprocessing and filtering, the processed split sizes are: train 3044, validation 375, and test 407. All evidence sentences used for retrieval are constructed exclusively from the training reports to avoid test leakage.

Evidence retrieval configuration. The evidence pool contains 6,750 unique sentences extracted from training reports after sentence segmentation, normalization, and deduplication. Each evidence sentence is embedded using a sentence encoder and stored in a dense index for cosine-similarity retrieval. At inference, each test example retrieves K=6 evidence sentences, which are serialized into a compact context list and injected into the model input. We select a small K to balance two competing goals: (i) providing sufficient supporting content to ground

key findings, and (ii) minimizing prompt length to reduce latency and avoid introducing irrelevant or contradictory phrases.

Training and decoding protocol. E1 and E2 are trained end-to-end using teacher forcing under the same optimizer schedule and early-stopping criterion on validation loss. E3 uses the same generator backbone but conditions decoding on the retrieved evidence list and applies citation-constrained prompting. E4 does not retrain the generator; instead, it generates multiple candidates via beam search and applies groundedness-aware reranking at inference. To ensure fair comparison, we keep the beam size consistent across variants and report latency for the full end-to-end pipeline (retrieval + decoding + optional reranking), as shown in Figure 5**.**

Metrics

We report widely used text-overlap metrics for long-form generation, including BLEU-4 (Papineni et al., 2002), ROUGE-L (Lin, 2004), and a CIDEr-style proxy similarity score (Vedantam et al., 2015). These metrics quantify surface overlap with the reference report and are useful for comparing fluency and phrasing similarity, but they do not directly measure clinical factuality. In radiology reporting, many clinically distinct cases can share common boilerplate language, so overlap metrics can be inflated by generic phrasing and may fail to penalize unsupported findings.

To better approximate factuality, we also report two entity-centric proxy measures:

a. EntityOverlap F1 (proxy). We extract clinical entities from the generated and reference reports and compute F1 overlap. This aims to capture whether the model mentions a similar set of clinical concepts (e.g., findings, anatomy, conditions) as the reference, providing a more content-sensitive signal than n-gram overlap. Entity extraction is implemented with lightweight clinical NLP proxies inspired by labelers and graph-based entity extractors (Boag et al., 2020; Jain et al., 2021).
b. Hallucination rate (proxy). We compute the fraction of generated entities that are not supported by an evidence set. For E3/E4, the evidence set is the retrieved sentence list E(I); for E1/E2 (which have no retrieved evidence), we use a reference-derived entity set as a proxy support set. This definition operationalizes hallucination at the entity level: if a generated entity is absent from the evidence support set, it is treated as unsupported.

We stress that these are proxy metrics designed for scalable comparison across variants; they do not replace expert adjudication, but they provide interpretable signals that align with the main objective of reducing unsupported content.

Main results and discussion

Table I summarizes the results across E1–E4, and Figure 3 visualizes the trend in text-overlap metrics across variants. As shown in Table I, adding grounding and reranking introduces a clear trade-off: overlap-based text metrics may decrease while factuality-oriented proxies improve.

a. Text quality vs. grounding.: E1 achieves the highest ROUGE-L and CIDEr proxy, which is consistent with generation that matches common phrasing and dataset-specific templates. This behavior can increase n-gram overlap even when clinical support is imperfect. Adding multimodal alignment (E2) improves entity overlap while slightly reducing n-gram metrics, suggesting that representation consistency helps the model select more appropriate clinical content without necessarily increasing surface-form similarity. The trend in Figure 3 illustrates this subtle shift: E2 can trade a small loss in overlap for improved content alignment.

Table 1. Main results on the Open-i test split (higher is better for BLEU/ROUGE/CIDEr/Entity-F1; lower is better for Halluc%). CIDEr* is a proxy similarity score. Entity-F1* and Halluc%* are entity-based proxies.

Model	BLEU-4	ROUGE-L	CIDEr*	Entity-F1	Halluc%*	Latency (T4 ms)
E1 Baseline (Img→Rpt)	0.050	0.274	0.070	0.307	61.6	199.3
E2 + Alignment	0.039	0.246	0.062	0.431	58.3	197.1
E3 + Grounding	0.046	0.195	0.028	0.502	22.2	619.4
E4 + Hallucination Ctrl	0.044	0.190	0.028	0.522	8.2	1136.6

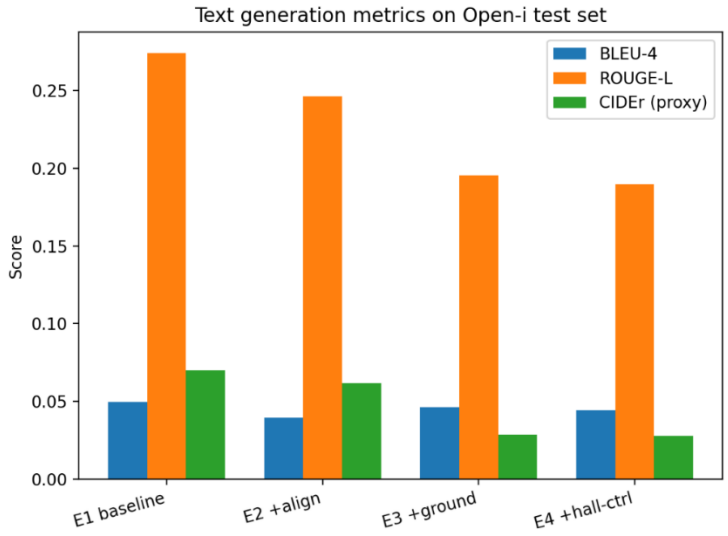

Figure 3. Text generation metrics on the Open-i test set across variants.

b. Hallucination reduction and content fidelity.: Evidence grounding (E3) substantially reduces the hallucination rate proxy (61.6% → 22.2%) and increases entity overlap (0.307 → 0.502), indicating that conditioning on retrieved evidence helps constrain the clinical

content. However, E3 may reduce ROUGE/CIDEr because citation constraints and evidence-conditioned phrasing can diverge from reference wording, particularly when the dataset contains repetitive normal templates. Reranking (E4) further reduces hallucination to 8.2% and yields the highest entity overlap (0.522), showing that inference-time selection based on groundedness can suppress unsupported entities beyond what prompting alone achieves. This inverse relationship is visualized in Figure 4, where stronger groundedness proxies correspond to lower hallucination rates.

c. Latency trade-off and deployment implications.: Grounding improvements are not free: retrieval and reranking increase end-to-end latency substantially, as shown in Figure 5. On T4, E3 incurs retrieval overhead (dense index lookup and additional context serialization) and can also increase decoding time due to longer conditioning inputs. E4 adds further overhead by reranking multiple beam candidates, which scales with the number of candidates and the cost of computing the groundedness score. The log-scale visualization in Figure 5 emphasizes the multiplicative nature of these costs and highlights a practical trade-off for deployment: systems seeking lower hallucination may require stronger.

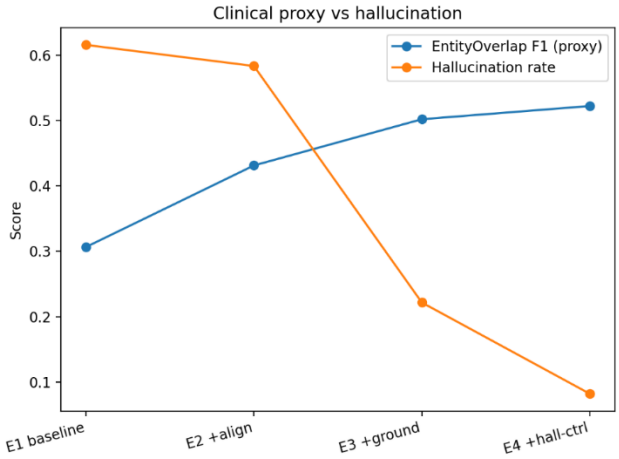

Figure 4. Clinical proxy (entity overlap F1) versus hallucination rate across variants. Grounding and reranking improve.

Qualitative analysis

Qualitative analysis Quantitative metrics summarize trends but can hide important behaviors such as incorrect citation attachment, evidence copying, or failure to mention subtle abnormalities. We therefore recommend including a compact qualitative panel in the final manuscript (e.g., 3–5 test cases) showing: (i) the retrieved evidence sentences with identifiers, (ii) the generated report with citations, and (iii) the reference report. Such examples make grounding interpretable and allow readers to assess whether citations are meaningful or merely

decorative. In particular, the panel should highlight both successes (correctly grounded abnormal findings) and failures (e.g., overly generic “normal” sentences, partial mismatches between evidence and generated claims, missing abnormality mentions, or citation omissions).

Limitations and safety considerations

Dataset size and bias. Open-i is relatively small and contains repetitive normal findings, which can encourage template-like generation and inflate overlap metrics. Models trained on such data may under-represent rare pathologies and may not generalize to broader clinical distributions (Demner-Fushman et al., 2016).

Proxy metrics and imperfect grounding signals. Entity overlap and hallucination rate are approximations. They may miss nuanced clinical correctness (e.g., severity, laterality, uncertainty language) and depend on the quality of entity extraction tools (Boag et al., 2020; Jain et al., 2021).

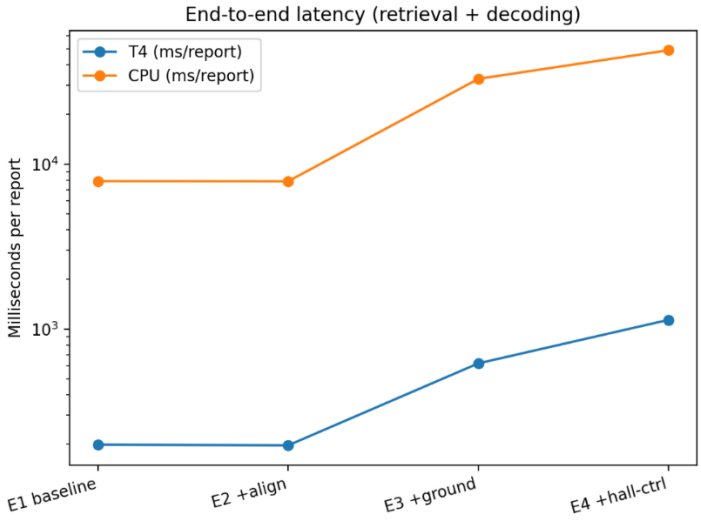

Figure 5. End-to-end latency (retrieval + decoding) for GPU (T4) and CPU, plotted on a log scale.

Retrieval noise and citation gaming. Retrieval can inject irrelevant or contradictory sentences, and the generator can attach citations incorrectly. While reranking reduces unsupported entities, it remains a heuristic control rather than a guarantee of clinical validity.

Non-clinical scope. This system is a research prototype and is not validated for clinical use. It must not be used for diagnosis or treatment decisions; over-reliance on fluent generated text can cause harm, especially when grounding is imperfect (Bender et al., 2021).

Table 2. Prediction intervals on TEST (α = 0.1). Higher coverage is better; lower width is

Method	Coverage (mean)	Width (mean)	Shift coverage
Static conformal	0.671	15.310	0.630
Rolling conformal	0.758	16.323	0.666

Comparison

Direct numerical comparison against prior work is challenging because reported IU X-ray results often differ in preprocessing details (e.g., section parsing, normalization), tokenization choices, image handling (single-view vs. multi-view), and even split definitions. These factors can shift overlap metrics substantially, making headline BLEU/ROUGE/CIDEr numbers difficult to compare fairly across papers (Demner-Fushman et al., 2016). We therefore provide a method-level comparison that focuses on what capabilities are added by evidence grounding, explicit citations, and groundedness-aware selection, rather than claiming absolute SOTA on a potentially non-matched evaluation protocol.

a. Prior report generators.: Most prior radiology report generators improve fidelity by strengthening the mapping from image representations to long-form clinical text. For example, memory-driven Transformers augment the decoder with memory mechanisms to better capture report structure and common clinical patterns (Chen et al., 2020). Cross modal memory networks extend this approach by explicitly aligning visual cues with textual memory to improve coverage of clinically salient findings and reduce generic phrasing (Chen et al., 2021). Multi-modal Transformer variants further strengthen cross-attention between visual features and generated tokens, improving coherence and long-range dependencies in report text (Liu et al., 2021). Collectively, these approaches represent the modern trend: increasing.

V**ariant**	+**Alignment**	**+Grounding (retrieval+citation)**	**+Rerank (halluc. ctrl)**
E1 Baseline	0	0	0
E2 + Alignment	1	0	0
E3 + Grounding	0	1	0
E4 + Hallucination Ctrl	0	1	1

However, two practical limitations are common across these families. First, grounding is typically implicit: while text is conditioned on image features, the system does not expose which specific evidence supports each statement, making auditing difficult without external tools or manual review. Second, evaluation in many studies is dominated by n-gram metrics and related overlap measures, which can reward template-like phrasing and may not adequately penalize unsupported entities (Lin, 2004; Papineni et al., 2002;

Vedantam et al., 2015). As a result, improvements in overlap metrics do not necessarily translate to reduced hallucination risk.

b. Grounded generation as an explicit interface.: Retrieval-augmented generation offers an alternative framing: instead of requiring the model to internalize all relevant knowledge in parameters, it retrieves supporting text and conditions generation on that retrieved context (Izacard & Grave, 2021; Lewis et al., 2020). In our setting, the retrieved text is not general-domain knowledge but a curated evidence pool extracted from training reports. This shifts the question from "can the model generate a plausible report?" to "can the model generate a report with explicit attribution?". This distinction matters operationally because attribution supports inspection, debugging, and policy enforcement (e.g., requiring citations for abnormal findings).

c. Our contribution relative to SOTA.: Relative to prior encoder–decoder report generators, our grounding pipeline adds two practical features that are not standard in the classical IU X-ray generation setup:

   a) Attribution via explicit citations. We require the generator to output citations [E#] that point to retrieved evidence sentences. This converts grounding into a visible, verifiable artifact: readers can check whether key claims are supported by the provided evidence, and the system can compute simple coverage checks (e.g., whether high-risk sentences contain at least one citation). In contrast, implicit grounding approaches require external post-processing to approximate support.

   b) A control knob for factuality–latency trade-offs. We introduce groundedness-aware reranking that trades inference-time compute for reduced hallucination. This can be tuned by (i) beam size or the number of candidates, and (ii) the groundedness weight in the reranking score. The result is a configurable mechanism: latency budgets can be respected by selecting smaller candidate sets, while safety-oriented settings can allocate more compute to reduce unsupported entities.

   These features align with a safety-first perspective on language generation: because hallucination is a known risk in NLG systems, mechanisms that expose evidence and enable controllable suppression of unsupported content are practically valuable (Ji et al., 2023).

d. What our approach does not claim.: We do not claim that explicit citations alone guarantee correctness. A model can still attach citations incorrectly, retrieve irrelevant evidence, or omit clinically important findings even when evidence is present. Instead, we position the contribution as a concrete, implementable framework for measuring and reducing

unsupported entities using proxy groundedness signals, and for making the model's conditioning information transparent for downstream audit.

e. When to prefer which approach.: The preferred approach depends on the target objective:
   a) Maximizingoverlapmetricsundertightlatency budgets. If the primary goal is maximal BLEU/ROUGE/CIDEr proxy on IU X-ray and runtime constraints dominate, a strong baseline generator (E1) or an aligned generator (E2) may be sufficient, especially when outputs are used only for low-stakes summarization or as a drafting aid.
   b) Reducing hallucination and improving content consistency. If the goal is to reduce unsupported entities and improve entity-level consistency, evidence grounding (E3) and groundedness-aware reranking (E4) are preferable. This is particularly relevant when attribution is required, when auditing is part of the workflow, or when safety concerns dominate system design choices (Ji et al., 2023).
f. Ablation and component-level interpretation.: To support this method-level comparison, we include an ablation table that isolates how each component (alignment, retrieval grounding, citation constraints, and reranking) affects text overlap, entity overlap, hallucination proxy, and latency. Table II provides this component-wise view and clarifies which gains come from retrieval grounding versus inference-time reranking.

5. CONCLUSIONS

We presented an evidence-grounded chest X-ray report generation pipeline that combines image-to-text generation, optional multimodal alignment, retrieval-based evidence prompting with explicit citations, and groundedness-aware reranking for hallucination control. The design is motivated by a practical concern in high-stakes natural language generation: fluent reports can still contain unsupported clinical entities, and purely overlap-based optimization does not reliably prevent this failure mode (Bender et al., 2021; Ji et al., 2023). Rather than treating grounding as an implicit byproduct of conditioning on image features, our pipeline makes the supporting context explicit by retrieving sentence-level evidence and requiring the model to cite that evidence in the generated report.

Across four controlled variants on Open-i (IU X-ray) (Demner-Fushman et al., 2016), evidence grounding substantially reduced the entity-level hallucination-rate proxy and improved entity overlap, while increasing end-to-end latency. Specifically, hallucination decreased from 61.6% (E1) to 22.2% (E3) and 8.2% (E4), while entity-overlap F1 increased from 0.307 (E1) to 0.502 (E3) and 0.522 (E4). These factuality-proxy gains came with higher inference cost (199 ms/report on T4 for E1 versus 619 ms for E3 and 1137 ms for E4), making

the quality groundedness latency trade-off explicit. Notably, the results also illustrate an important evaluation caveat emphasized in prior work: text-overlap metrics can remain high for template-heavy outputs even when entity-level support is weak, reinforcing the need for complementary factuality-oriented measurements (Lin, 2004; Papineni et al., 2002; Vedantam et al., 2015).

a. Implications.: Our findings suggest that the preferred modeling choice depends on the operational objective and the tolerance for inference-time compute:
    a) Fluency/overlap-oriented settings. If the primary goal is to maximize surface similarity to reference reports under tight runtime constraints (e.g., batch summarization, low-stakes drafting aids), a strong baseline generator (E1) or an aligned generator (E2) may be sufficient. In such settings, the added latency of retrieval and reranking may not be justified.
    b) Safety- and auditability-oriented settings. If the goal is to reduce unsupported entities and provide transparent attribution for key findings, evidence retrieval with citations (E3) and grounded reranking (E4) offers a practical alternative. Citations provide a simple auditing interface (coverage checks and post-hoc inspection), while reranking provides a tunable mechanism for further suppressing unsupported entities at an explicit inference cost.

        More broadly, the results reinforce a safety-first framing: when outputs can be over-trusted, systems should prefer designs that expose evidence and enable controllable reductions in hallucination risk (Bender et al., 2021; Ji et al., 2023).

b. Limitations and future work.: This work has several limitations. First, Open-i is relatively small and exhibits dataset-specific biases (including repetitive normal patterns), which can inflate overlap metrics and limit generalization to broader clinical distributions (Demner-Fushman et al., 2016). Second, our groundedness and hallucination measurements are proxy metrics; while entity-based proxies are more content-sensitive than n-gram overlap, they do not fully capture nuanced clinical correctness (e.g., laterality, severity, uncertainty language) and depend on the accuracy of extraction tools (Boag et al., 2020; Jain et al., 2021). Third, we do not provide clinical validation; the system is a research prototype and is not suitable for clinical decision-making.

    Research prototype and is not suitable for clinical decision-making. Future work should therefore prioritize: (i) evaluation on larger, more diverse datasets (e.g., MIMIC-CXR) to stress-test retrieval and grounding under distribution shift (Johnson et al., 2019); (ii) stronger groundedness scoring that goes beyond bag-of-entities to entity–relation

matching and structured consistency checks (Jain et al., 2021); and (iii) adaptive decoding policies that explicitly trade citation strictness, uncertainty phrasing, and fluency based on runtime budgets and risk tolerance. On the representation side, future work may also explore stronger vision–language pretraining to improve retrieval quality and alignment robustness (Nguyen et al., 2022; Radford et al., 2021).

c. Reproducibility.: To support reproducibility, our experimental pipeline produces a companion artifact bundle that includes trained checkpoints, evaluation tables, and figures generated from a fixed preprocessing and evaluation script (Paszke et al., 2019).

REFERENCES

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT), 610–623. https://doi.org/10.1145/3442188.3445922

Boag, W., Wittenberg, E., Folkman, L., Khosla, S., & Manrai, A. (2020). Chexbert: Combining automatic labelers and expert annotations for accurate radiology report labeling using BERT. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)

Chen, Z., Song, Y., Chang, T.-H., & Wan, X. (2020). Generating radiology reports via memory-driven transformer. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 1439–1449. https://doi.org/10.18653/v1/2020.emnlp-main.112

Chen, Z., Song, Y., Chang, T.-H., & Wan, X. (2021). Cross-modal memory networks for radiology report generation. Pattern Recognition, 118, 108050. https://doi.org/10.1016/j.patcog.2021.108050

Demner-Fushman, D., Kohli, M. D., Rosenman, M. B., Shooshan, S. E., Rodriguez, L., Antani, S., Thoma, G. R., & McDonald, C. J. (2016). Preparing a collection of radiology examinations for distribution and retrieval. Journal of the American Medical Informatics Association, 23(2), 304–310. https://doi.org/10.1093/jamia/ocv080

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 770–778. https://doi.org/10.1109/CVPR. 2016.90

Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilona, S., Chute, C., Marklund, H., Haghgoo, B., Ball, R., Shpanskaya, K., et al. (2019). Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. Proceedings of the AAAI Conference on Artificial Intelligence, 33(1), 590–597. https://doi.org/10.1609/aaai. v33i01.3301590 Izacard,

G., & Grave, E. (2021). Leveraging passage retrieval with generative models for open domain question answering. Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics (EACL).

Jain, S., Delbrouck, J.-B., Vakil, P., Chi, P., Ommer, B., & Langlotz, C. (2021). Radgraph: Extracting clinical entities and relations from radiology reports. arXiv preprint arXiv:2106.14463.

Ji, Z., Lee, N., Frieske, R., et al. (2023). Survey of hallucination in natural language generation. ACM Computing Surveys, 55(12), 1–38. https://doi.org/10.1145/3571730

Jing, B., Xie, P., & Xing, E. (2018). On the automatic generation of medical imaging reports. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL), 2577–2586. https://doi.org/10.18653/ v1/P18-1240

Johnson, A. E. W., Pollard, T. J., Berkowitz, S. J., Greenbaum, N. R., Lungren, M. P., Deng, C.-y., Mark, R. G., & Horng, S. (2019). Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. Scientific Data, 6, 317. https://doi.org/10.1038/s41597-019-0322-0

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Stoyanov, V., & Riedel, S. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. Advances in Neural Information Processing Systems, 33, 9459–9474.

Li, Y., Liang, X., Hu, Z., & Xing, E. (2018). TieNet: Text-image embedding network for common thorax disease classification and reporting in chest x-rays. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 9049–9058. https://doi.org/10.1109/CVPR.2018.00943

Lin, C.-Y. (2004). Rouge: A package for automatic evaluation of summaries. Text Summarization Branches Out: Proceedings of the ACL-04 Workshop.

Liu, F., You, C., Wu, X., Xu, G., Liu, Y., Liu, T., & Wang, J. (2021). Multi-modal transformer for radiology report generation. Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP).

Nguyen, H., et al. (2022). BioViL: Self-supervised vision–language pretraining for biomedical imaging. arXiv preprint arXiv:2204.09817.

Oord, A. v. d., Li, Y., & Vinyals, O. (2018). Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748.

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). Bleu: A method for automatic evaluation of machine translation. Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), 311–318.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al. (2019). PyTorch: An imperative style, high-performance deep learning library. Advances in Neural Information Processing Systems, 32.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). Learning transferable visual models from natural language supervision. Proceedings of the 38th International Conference on Machine Learning (ICML), 8748–8763.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30.

Vedantam, R., Zitnick, C. L., & Parikh, D. (2015). Cider: Consensus-based image description evaluation. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 4566–4575. https://doi.org/10. 1109/CVPR.2015.7299087 14

Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., & Summers, R. M. (2017). Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2097–2106