CARE: Clinically Aligned Retrieval Evidence for Consistent Radiology Report Generation

Jintao Zhu[1], Yi Gu[2], and Jiaxin Ma[1,3⋆]

[1] Zhejiang CAS Angels Biotechnology Co. Ltd., China
[2] Nara Institute of Science and Technology, Japan
[3] OMRON SINIC X Corp., Japan
jiaxin.ma@sinicx.com

Abstract. Automatic radiology report generation has the potential to assist radiologists by providing preliminary clinical descriptions. Despite recent progress in medical vision-language models, generated reports often contain factual inconsistencies and repetitive templated findings, indicating weak grounding to image evidence. Retrieval-augmented generation (RAG) improves grounding by conditioning on prior reports; however, existing end-to-end optimization suffers from representation inconsistency, as actively updated query representations diverge from stale historical keys. This mismatch destabilizes the retrieval space and disrupts supervision. We propose CARE (Clinically Aligned Retrieval Evidence), a multimodal framework introducing temporally coherent retrieval supervision. CARE maintains a shared memory queue populated by a slowly evolving momentum key encoder to ensure a stable representation space. We further introduce a Disease-Aware Alignment objective that enforces diagnostic agreement between input images and retrieved reports to guide generation. Trained on IU-Xray and evaluated on the MIMIC-CXR-RRG dataset, CARE significantly improves report quality. It nearly doubles clinical entity extraction performance in training-free adaptation settings and substantially reduces repetitive report collapse across multiple backbones. The source code is available at: https://github.com/Levii-zhu/CARE

Keywords: Radiology report generation · Retrieval-augmented generation · Vision-language models · Chest X-ray.

1 Introduction

Automated radiology report generation aims to produce clinically grounded diagnostic descriptions from medical images, alleviating radiologists' workload and reducing diagnostic variability [10,2]. A high-quality report must provide structured and precise interpretations of visual findings to support downstream therapeutic decisions [12]. Recent progress in multimodal large language models (MLLMs) has demonstrated strong biomedical visual understanding [15,21].

⋆ Corresponding author.

Despite this, directly applying these models to radiology reporting remains challenging: generated reports often contain factual inconsistencies or unsupported clinical findings, or fall back to repetitive templates lacking patient-specific clinical details [9,24,33]. Unlike generic image captioning, radiology reporting requires strict alignment between visual evidence and pathological interpretation [13], making even minor factual errors clinically problematic.

Retrieval-Augmented Generation (RAG) introduces external clinical evidence by conditioning report generation on prior cases [4]. While traditional systems employ a frozen retriever, end-to-end RAG jointly optimizes the retriever with the generation objective [14]. However, as the query encoder is updated through backpropagation, its representations may gradually diverge from stale historical keys stored in the retrieval queue, creating inconsistent retrieval supervision. This mismatch can weaken reliable image–pathology associations and encourage the generator to rely on frequent textual patterns rather than retrieved visual evidence. Such shortcut learning [5] is particularly problematic under cross-dataset shifts, where diagnostic statements should remain grounded in the input image. Unlike conventional momentum-contrastive learning, where queued keys primarily serve as negative samples, our method uses the queue as a source of retrieved clinical evidence for report generation. Therefore, temporal drift affects not only contrastive supervision but also the evidence conditioned on by the generator.

Motivated by these challenges, we propose CARE (Clinically Aligned Retrieval Evidence), a retrieval-grounded training framework designed to maintain coherent retrieval supervision during optimization. CARE adopts a MoCo-inspired architecture [6] with a gradient-updated query encoder and a slowly evolving momentum encoder, ensuring that historical keys in the memory queue reside in a smoothly evolving representation space. In addition, we introduce a Disease-Aware Alignment objective using CheXbert [27]. Unlike prior disease-matching retrieval constraints [26], CARE couples diagnostic compatibility with differentiable retrieval weights, encouraging the generator to condition on clinically compatible evidence rather than superficially similar text.

We train on the IU-Xray dataset [3] and evaluate out-of-domain generalization via training-free adaptation on the MIMIC-CXR benchmark [11]. **The contributions** of this work are threefold: 1) We propose CARE, a retrieval-grounded framework that maintains temporally coherent retrieval supervision through a momentum-updated key encoder and a shared memory queue. Unlike conventional contrastive queues, the stored representations are associated with reports used as generation evidence, making retrieval consistency important for both representation learning and report generation. 2) We introduce a differentiable disease-aware evidence alignment objective that couples retrieval weighting with diagnostic compatibility, encouraging the model to select clinically compatible retrieved reports rather than merely textually similar evidence. 3) Extensive evaluations demonstrate superior clinical efficacy on IU-Xray and robust out-of-domain generalization on MIMIC-CXR. Empirically, CARE nearly doubles the clinical entity extraction performance of its base model in training-free adaptation settings and remains competitive with fully supervised models.

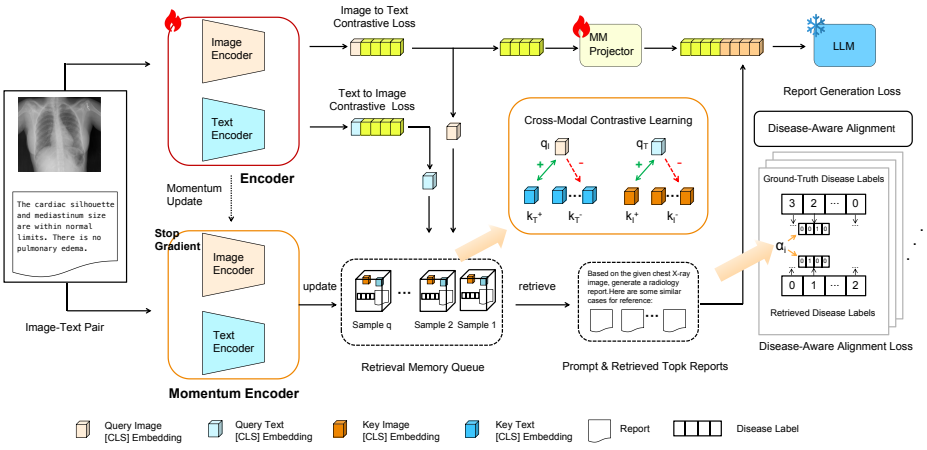

Fig. 1. Overview of the CARE framework. A chest X-ray image is processed by the query encoder to extract query features q_I and q_T, while the momentum key encoder (updated via exponential moving average) provides temporally coherent key representations k_I and k_T to populate a shared Retrieval Memory Queue. We enforce a Disease-Aware Alignment Loss using soft retrieval weights α_i to ensure the retrieved top-K reports are diagnostically consistent with the visual input. Finally, the projected image features H_{proj} and the retrieved clinical evidence R are fed into a frozen LLM to autoregressively generate the diagnostic report.

2 Method

Consistent Memory Queue via Momentum Encoders. To prevent the retrieval representation space from becoming inconsistent during active end-to-end optimization, CARE adopts a momentum-driven memory architecture [6]. As illustrated in Fig. 1, a gradient-updated query encoder f_q extracts query features $q_I \in \mathbb{R}^d$ and $q_T \in \mathbb{R}^d$ from an input image I and its corresponding report T. Concurrently, a slowly evolving momentum encoder f_k processes the same inputs to produce key features k_I and k_T. We maintain a First-In-First-Out (FIFO) dynamic memory queue $\mathcal{Q}$ of size M. For each historical sample, $\mathcal{Q}$ stores the momentum-encoded keys (k_I, k_T), the ground-truth disease label vector y, and the raw report text r. Crucially, the momentum encoder parameters θ_k are updated strictly via an exponential moving average (EMA) of the query parameters ($\theta_k \leftarrow m\theta_k + (1-m)\theta_q$). This mathematical constraint decouples the memory keys from abrupt backpropagation updates, ensuring that the historical representations in $\mathcal{Q}$ reside in a smoothly evolving, consistent space.

Contrastive Retrieval Learning. Building upon recent successes in cross-modal representation learning (e.g., ALBEF [17]) and medical vision-language pre-training [35,30], we optimize the retrieval space using the coherent keys established in $\mathcal{Q}$. Specifically, we apply a symmetric InfoNCE loss [22] to contrast

the active queries against their corresponding positive keys, while pushing away negative keys sampled from both the dynamic memory queue and the current mini-batch. The objective is defined as:

$$\mathcal{L}_{contrast} = -\frac{1}{2} \sum_{v \in \{I,T\}} \log \frac{\exp(q_v \cdot k_{\bar{v}}^{+} / \tau_{nce})}{\sum_{k \in \{k_{\bar{v}}^{+}\} \cup \mathcal{Q}_{\bar{v}} \cup \mathcal{B}_{\bar{v}}} \exp(q_v \cdot k / \tau_{nce})}, \tag{1}$$

where $v \in \{I, T\}$ denotes the query modality, and $\bar{v}$ denotes its cross-modal counterpart. For a given query q_v, $k_{\bar{v}}^{+}$ represents the positive key feature from the paired sample. The denominator computes the similarity over the positive key and a comprehensive set of negative keys k, drawn from the union of the cross-modal memory queue $\mathcal{Q}_{\bar{v}}$ and the current mini-batch $\mathcal{B}_{\bar{v}}$. By leveraging the momentum queue, this contrastive objective effectively decouples the negative sample size from the mini-batch limit while maintaining representation consistency.

Disease-Aware Clinical Evidence Alignment. For enforcing diagnostic consistency, we retrieve the top-K candidates from $\mathcal{Q}$ based on cosine similarity $s_i = \text{sim}(q_I, k_{T_i})$ and assign soft weights via softmax relaxation as $\alpha_i = \exp(s_i/\tau_{align}) / \sum_{j=1}^{K} \exp(s_j/\tau_{align})$, where τ_{align} is a temperature hyperparameter controlling the sharpness of the retrieval weight distribution. Let y and y_{T_i} denote the clinical state vectors [27] of the query and the i-th retrieved sample. The alignment objective minimizes the expected diagnostic penalty computed by Mean Squared Error (MSE) as $\mathcal{L}_{align} = \sum_{i=1}^{K} \alpha_i \cdot \text{MSE}(\text{OneHot}(y), \text{OneHot}(y_{T_i}))$. Crucially, gradients propagate through α_i to the query encoder, coupling retrieval weighting with visual representation learning.

Retrieval-Grounded Report Generation & Optimization. The frozen LLM is conditioned on the retrieved reports $R = \{r_1, \ldots, r_K\}$ and the image patch tokens H_p. Specifically, H_p is mapped to the LLM's text embedding space via a trainable Multimodal (MM) Projector [19], yielding the projected features H_{proj}. For seamless integration of visual and textual evidence, we construct the multimodal input using a fixed instruction prompt: *"Based on the given chest X-ray image, generate a radiology report. Here are some similar cases for reference: [R]"*, where [R] represents the concatenated text of the top-K retrieved reports. The generation loss is computed as the standard autoregressive negative log-likelihood $\mathcal{L}_{gen} = -\sum_t \log P(w_t \mid w_{<t}, H_{proj}, R)$, where w_t denotes the t-th token of the ground-truth report and $w_{<t}$ represents all preceding tokens. During training, the LLM remains frozen, and the momentum encoders are updated strictly via exponential moving average. Only the query encoders and the MM Projector are updated by minimizing the joint objective. The three objectives respectively regularize representation consistency, clinical compatibility, and generative fidelity, combined as $\mathcal{L}_{total} = \mathcal{L}_{contrast} + \mathcal{L}_{align} + \mathcal{L}_{gen}$.

3 Experiments

Datasets and evaluation metrics. CARE is evaluated under two settings: *in-domain* performance and *out-of-domain* generalization. IU-Xray [3] serves as the primary training corpus and is evaluated using the official patient-level split. For out-of-domain assessment, we adopt the MIMIC-CXR-RRG benchmark [33,34], a standardized evaluation subset derived from the official MIMIC-CXR test set [11]. This cross-dataset setting introduces substantial variations in disease prevalence, institutional reporting styles, and imaging protocols. To assess generalization without prior target-domain supervised training, we employ the LLaVA-Med [15] backbone, whose pre-training data excludes MIMIC-CXR. All parameters optimized on IU-Xray remain frozen, and adaptation to the target reporting style is achieved solely by populating the external memory queue with text-only reports from MIMIC-CXR. We report standard natural language generation (NLG) metrics (BLEU, METEOR, ROUGE-L) [25,1,18], together with clinically grounded evaluations, including RadGraph F1 and Exact Reward (RG-F_1, RG_{ER}) [8], as well as CheXbert-based micro- and macro-F1 over five CheXpert categories [27,7], denoted ${}^5\text{Mi-F}_1$ and ${}^5\text{Ma-F}_1$, respectively. We also report RadCliQ0 (CliQ_0; $\downarrow$), a composite metric reflecting radiologist preferences [31].

Implementation details. We instantiate CARE on LLaVA-Med [15] and LLaVA-Rad [32]. For fair comparison, we freeze the core LLM and optimize only the query encoder, multimodal projector, and momentum retrieval components. Models are trained on 8 NVIDIA A100 GPUs for 20 epochs using AdamW [20] (lr=5×10^{-5}, batch size=32, bf16). Retrieval parameters follow the configuration of $K = 5$, $m = 0.999$, $|\mathcal{Q}| = 1000$, and $\tau_{nce} = \tau_{align} = 0.07$. Open-source baselines are locally re-evaluated using official weights unless otherwise noted.

In-domain performance. Evaluation on the IU-Xray dataset demonstrates the effectiveness of CARE across both the general-domain LLaVA-Med backbone and the medically optimized LLaVA-Rad backbone. As shown in Table 1, CARE consistently outperforms the corresponding SFT variants on most evaluation metrics, while remaining competitive with the compared baselines(XrayGPT, R2GenGPT, and EVCAP [28,29,16]). An interesting observation is that the improvements are not always reflected in RG-F_1 and CliQ_0, two clinically important report-level metrics. In several cases, CARE achieves comparable or slightly lower scores than the corresponding SFT models despite outperforming them on most other metrics. We hypothesize that this behavior is related to differences in report diversity. Standard SFT tends to imitate dominant normal-style reporting templates, whereas retrieval-augmented CARE generates more diverse reports conditioned on retrieved evidence. To examine this effect, Fig. 2 reports the cumulative number of unique N-grams generated across the test set. For both LLaVA-Med and LLaVA-Rad, CARE exhibits substantially higher vocabulary diversity, while the corresponding SFT models produce markedly flatter

Table 1. Performance evaluation on the IU-Xray dataset. Models are grouped by their exposure to the training set.

Dataset: IU-Xray									
Model	**Clinical Metrics**					**Lexical Metrics**			
	^{5}Mi-F$_1$	**^{5}Ma-F$_1$**	**RG-F$_1$**	**RG$_{ER}$**	**CliQ$_0$ (↓)**	**B-1**	**B-4**	**MTR**	**R-L**
Zero- or Few-Shot Models									
MedFlamingo [21] (Few-shot)	5.1	3.2	5.0	5.1	4.86	14.2	1.3	8.8	10.0
LLaVA-Med [15]	3.9	1.0	6.4	7.0	4.51	13.4	0.7	12.4	10.1
Supervised Models (Either Pretrained or Finetuned on IU-Xray)									
XrayGPT [28]	12.2	12.3	30.4	32.4	3.02	32.3	5.6	23.9	24.1
R2GenGPT [29]	3.9	3.5	24.8	24.8	3.27	38.2	12.8	**30.9**	24.6
EVCAP [16]	11.2	5.4	20.2	32.1	3.72	16.8	3.3	25.2	16.5
LLaVA-Med (SFT) [15]	1.0	0.3	30.1	25.3	3.18	25.7	8.6	19.6	27.9
CARE (LLaVA-Med)	20.4	15.4	27.5	29.5	3.17	30.3	6.9	24.7	24.0
LLaVA-Rad [32]	27.9	<u>23.9</u>	28.5	30.0	3.03	37.0	8.9	25.3	27.3
LLaVA-Rad (SFT) [32]	<u>34.4</u>	23.2	**38.0**	<u>34.0</u>	**2.55**	<u>39.1</u>	<u>13.3</u>	28.0	**34.4**
CARE (LLaVA-Rad)	**35.1**	**26.9**	<u>32.2</u>	**35.4**	<u>2.75</u>	**40.7**	**14.6**	<u>30.3</u>	<u>31.9</u>

diversity curves, indicating stronger reliance on recurrent report templates. We further argue that this trade-off is generally favorable from a clinical perspective. Although template-driven generation can maintain competitive RG-F$_1$ and CliQ$_0$ scores, excessive reliance on normal-style templates may impair abnormality recognition. This effect is particularly evident for LLaVA-Med, whose SFT version exhibits extremely low diversity and even achieves lower ^{5}Mi-F$_1$ and ^{5}Ma-F$_1$ scores than the original backbone, suggesting degraded abnormality-label agreement after fine-tuning. A similar but less severe trend can be observed for LLaVA-Rad, where the SFT version attains a lower ^{5}Ma-F$_1$ than the original model. In contrast, CARE avoids this performance regression. These findings suggest that CARE achieves a better balance between report-level similarity and pathology-level correctness, mitigating the mode collapse commonly observed in standard SFT.

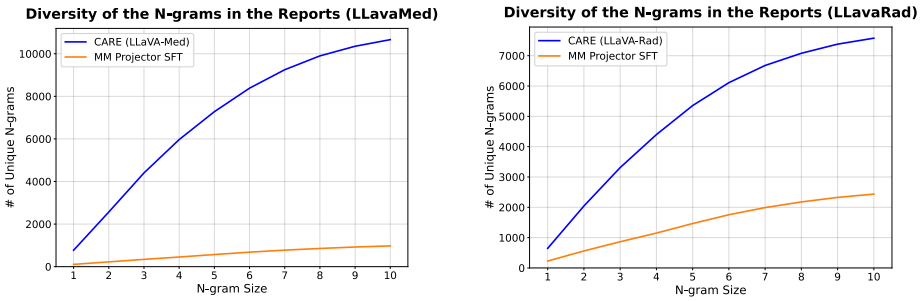

Fig. 2. Cumulative count of unique N-grams generated across the test set. CARE maintains substantially higher vocabulary diversity.

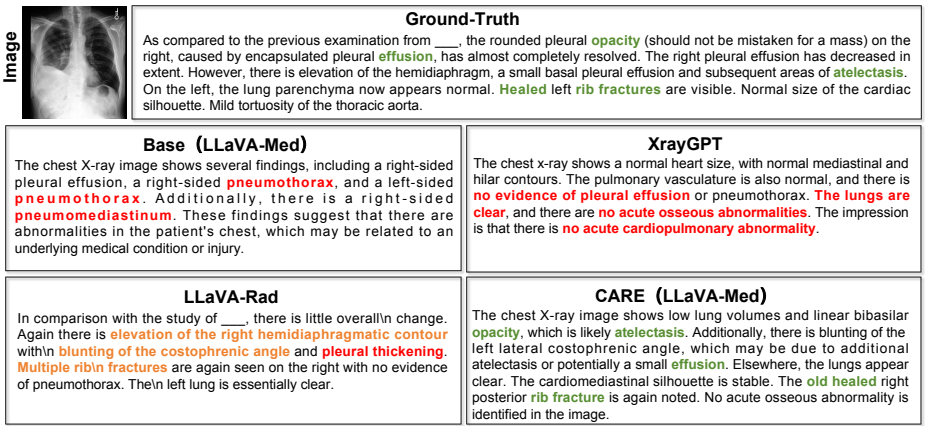

Fig. 3. Qualitative comparison under the training-free adaptation setting. Unlike baselines that suffer from mode collapse (e.g., outputting rigid "clear lungs" templates) or unsupported clinical findings (e.g., inventing pneumothorax), CARE effectively aligns with genuine clinical pathology, correctly identifying atelectasis, small effusion, and the subtle "old healed" attribute of the rib fracture.

Table 2. Out-of-domain generalization on the MIMIC-CXR-RRG benchmark. Baseline results for GPT-4V, MedFlamingo, and LLaVA are reported from [32]. Best and second-best results under the training-free adaptation setting are highlighted in bold and underlined, respectively.

Dataset: MIMIC-CXR-RRG									
	Clinical Metrics					Lexical Metrics			
Model	^{5}**Mi-F$_1$**	^{5}**Ma-F$_1$**	**RG-F$_1$**	**RG$_{ER}$**	**CliQ$_0$** (↓)	**B-1**	**B-4**	**MTR**	**R-L**
Supervised Models (Trained on MIMIC-CXR)									
XrayGPT [28]	39.5	30.0	19.7	18.1	3.66	30.3	5.9	19.1	21.9
R2GenGPT [29]	22.7	19.2	17.1	17.6	3.68	32.3	7.8	23.3	20.5
LLaVA-Rad [32]	55.7	46.6	21.2	20.7	3.41	33.4	9.7	23.0	25.1
Zero- or Few-Shot Models									
GPT-4V [23]	25.8	<u>19.6</u>	**13.2**	–	–	16.4	<u>1.9</u>	–	13.2
MedFlamingo [21] (Few-shot)	21.1	12.1	9.2	<u>7.1</u>	4.37	10.6	1.3	9.6	14.4
LLaVA [19]	23.4	17.5	4.5	–	–	21.0	1.3	–	13.8
LLaVA-Med [15]	<u>35.2</u>	17.5	7.0	6.3	<u>4.26</u>	<u>23.2</u>	1.2	<u>12.0</u>	<u>15.6</u>
CARE (LLaVA-Med)	**37.7**	**30.4**	<u>13.0</u>	**13.2**	**3.88**	**31.4**	**4.3**	**17.3**	**17.8**

Out-of-domain generalization. Evaluation of whether CARE captures invariant diagnostic semantics—rather than overfitting to dataset-specific patterns—is conducted via out-of-domain assessment on the MIMIC-CXR-RRG benchmark [11] (Table 2). For training-free adaptation, all IU-Xray optimized parameters remain frozen. Leveraging RAG's plug-and-play nature, report-style

adaptation is achieved simply by populating the memory queue with 500 randomly sampled text-only reports from MIMIC-CXR-RRG. As inference relies solely on cross-modal retrieval, this setting requires no target-domain images or labels and evaluates adaptation under the absence of target-domain supervision. Unlike specialized medical baselines (e.g., LLaVA-Rad [32], XrayGPT [28], and R2GenGPT [29]) that inherently include MIMIC-CXR in their training corpora, the LLaVA-Med [15] backbone was never exposed to MIMIC-CXR data. Evaluated under this training-free domain adaptation setting, CARE demonstrates robust generalization, nearly doubling the clinical entity extraction performance of the LLaVA-Med baseline (RG-F1 [8] improving from 7.0 to 13.0). Furthermore, as qualitatively demonstrated in Figure 3, CARE effectively overcomes the memorization of dataset-specific lexical priors, successfully identifying subtle clinical pathologies while baseline models suffer from severe mode collapse.

Table 3. Ablation study evaluating the step-wise contributions of CARE components using the LLaVA-Med [15] backbone on the IU-Xray dataset.

Dataset: IU-Xray									
	Clinical Metrics					**Lexical Metrics**			
Ablation Setting	^{5}**Mi-F$_1$**	^{5}**Ma-F$_1$**	**RG-F$_1$**	**RG$_{\text{ER}}$**	**CliQ$_0$** (↓)	**B-1**	**B-4**	**MTR**	**R-L**
Backbone: LLaVA-Med									
Pre-trained Base	3.9	1.0	6.4	7.0	4.51	13.4	0.7	12.4	10.1
+ In-batch CRL	13.8	8.5	25.4	28.9	3.44	22.0	3.8	23.5	19.9
+ SFT (MM Projector) [19]	14.6	12.2	26.3	29.0	3.20	28.1	5.5	24.3	22.8
+ Disease-Aware Alignment	14.9	12.4	26.9	29.3	3.22	28.2	6.6	24.6	23.1
→ Momentum Memory (**CARE**)	**20.4**	**15.4**	**27.5**	**29.5**	**3.17**	**30.3**	**6.9**	**24.7**	**24.0**

Ablation study. We conduct an ablation study on the LLaVA-Med [15] backbone to analyze the incremental contributions of CARE components (Table 3). Starting from the *Pre-trained Base*, we first introduce retrieval-augmented generation optimized via In-batch Contrastive Retrieval Learning (+ In-batch CRL), where training relies solely on in-batch negatives without a memory queue. The retriever is trained within the batch but remains fixed at inference. We then incorporate Supervised Fine-Tuning (+ SFT) to optimize the multimodal projector for end-to-end report generation, which substantially improves clinical extraction (e.g., RG-F_1 reaching 26.3). However, in-batch sampling limits the diversity of retrieval supervision and does not provide a temporally consistent representation space for dynamic optimization. Adding the Disease-Aware Alignment objective further enforces diagnostic compatibility between images and retrieved reports, yielding consistent gains. Finally, replacing in-batch sampling with a momentum-driven memory queue (→ Momentum Memory, **CARE**) introduces a dynamically maintained yet temporally stable retrieval space. This design enlarges the negative pool beyond batch constraints while ensuring smooth representation

evolution during joint training. As a result, CARE achieves the best overall clinical performance, demonstrating the importance of coherent and stable retrieval supervision.

4 Conclusion

We introduced CARE, a retrieval-grounded framework mitigating representation inconsistency for end-to-end radiology report generation. By integrating a momentum memory queue and a Disease-Aware Alignment Objective, CARE encourages reports to be conditioned on clinically compatible visual evidence rather than dataset-specific priors. Evaluations show that CARE reduces shortcut learning [5] and template collapse while improving clinical entity extraction in training-free adaptation settings. CARE still lacks explicit filtering for severely erroneous retrieved evidence and has not been evaluated under adversarially corrupted retrieval evidence. It also lacks entity- or region-level grounding and expert radiologist evaluation. Future work will address these limitations through retrieval verification, fine-grained grounding, and parameter-efficient LLM adaptation.

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

References

1. Banerjee, S., Lavie, A.: METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In: Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization. pp. 65–72 (2005)
2. Chen, Z., Song, Y., Chang, T.H., Wan, X.: Generating radiology reports via memory-driven transformer. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 1439–1449 (2020)
3. Demner-Fushman, D., Kohli, M.D., Rosenman, M.B., et al.: Preparing a collection of radiology examinations for distribution and retrieval. Journal of the American Medical Informatics Association **23**(2), 304–310 (2016)
4. Endo, M., Krishnan, R., Krishna, V., Ng, A.Y., Rajpurkar, P.: Retrieval-based chest x-ray report generation using a pre-trained contrastive language-image model. In: Proceedings of Machine Learning for Health. vol. 158, pp. 209–219 (2021)
5. Geirhos, R., Jacobsen, J.H., Michaelis, C., et al.: Shortcut learning in deep neural networks. Nature Machine Intelligence **2**(11), 665–673 (2020)
6. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9729–9738 (2020)
7. Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., et al.: CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 33, pp. 590–597 (2019)

8. Jain, S., Agrawal, A., Saporta, A., Truong, S.Q.H., et al.: RadGraph: Extracting clinical entities and relations from radiology reports. In: Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks. vol. 1 (2021)
9. Ji, Z., Lee, N., Frieske, R., Yu, T., et al.: Survey of hallucination in natural language generation. ACM Computing Surveys **55**(12), 248:1–248:38 (2023)
10. Jing, B., Xie, P., Xing, E.: On the automatic generation of medical imaging reports. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 2577–2586 (2018)
11. Johnson, A.E.W., Pollard, T.J., Berkowitz, S.J., et al.: MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. Scientific Data **6**(1), 317 (2019)
12. Kahn, Jr., C.E., Langlotz, C.P., Burnside, E.S., et al.: Toward best practices in radiology reporting. Radiology **252**(3), 852–856 (2009)
13. Langlotz, C.P.: The Radiology Report: A Guide to Thoughtful Communication for Radiologists and Other Medical Professionals. CreateSpace Independent Publishing Platform (2015)
14. Lewis, P., Perez, E., Piktus, A., Petroni, F., et al.: Retrieval-augmented generation for knowledge-intensive NLP tasks. In: Advances in Neural Information Processing Systems. vol. 33, pp. 9459–9474 (2020)
15. Li, C., Wong, C., Zhang, S., Usuyama, N., et al.: LLaVA-Med: Training a large language-and-vision assistant for biomedicine in one day. In: Advances in Neural Information Processing Systems. vol. 36, pp. 28541–28564 (2023)
16. Li, J., Vo, D.M., Sugimoto, A., Nakayama, H.: EVCap: Retrieval-augmented image captioning with external visual-name memory for open-world comprehension. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13733–13742 (2024)
17. Li, J., Selvaraju, R., Gotmare, A., Joty, S., et al.: Align before fuse: Vision and language representation learning with momentum distillation. In: Advances in Neural Information Processing Systems. vol. 34, pp. 9694–9705 (2021)
18. Lin, C.Y.: ROUGE: A package for automatic evaluation of summaries. In: Text Summarization Branches Out. pp. 74–81 (2004)
19. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. In: Advances in Neural Information Processing Systems. vol. 36, pp. 34892–34916 (2023)
20. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: International Conference on Learning Representations (2019)
21. Moor, M., Huang, Q., Wu, S., Yasunaga, M., et al.: Med-Flamingo: A multimodal medical few-shot learner. In: Proceedings of the 3rd Machine Learning for Health Symposium. vol. 225, pp. 353–367 (2023)
22. van den Oord, A., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 (2018)
23. OpenAI: GPT-4V(ision) system card. Tech. rep., OpenAI (2023)
24. Pal, A., Umapathi, L.K., Sankarasubbu, M.: Med-HALT: Medical domain hallucination test for large language models. arXiv preprint arXiv:2307.15343 (2023)
25. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: BLEU: A method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. pp. 311–318 (2002)
26. Park, S.J., Heo, K.S., Shin, D.H., Son, Y.H., Oh, J.H., Kam, T.E.: Dart: Disease-aware image-text alignment and self-correcting re-alignment for trustworthy radiology report generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15580–15589 (2025)

27. Smit, A., Jain, S., Rajpurkar, P., Pareek, A., et al.: CheXbert: Combining automatic labelers and expert annotations for accurate radiology report labeling using BERT. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 1500–1519 (2020)
28. Thawkar, O., Shaker, A., Mullappilly, S.S., et al.: XrayGPT: Chest radiographs summarization using medical vision-language models. arXiv preprint arXiv:2306.07971 (2023)
29. Wang, Z., Liu, L., Wang, L., Zhou, L.: R2GenGPT: Radiology report generation with frozen LLMs. Meta-Radiology **1**(3), 100033 (2023)
30. Wang, Z., Wu, Z., Agarwal, D., Sun, J.: MedCLIP: Contrastive learning from unpaired medical images and text. In: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. pp. 3876–3887 (2022)
31. Yu, F., Endo, M., Krishnan, R., Pan, I., et al.: Evaluating progress in automatic chest x-ray radiology report generation. Patterns **4**(9), 100802 (2023)
32. Zambrano Chaves, J.M., Huang, S.C., Xu, Y., et al.: Towards a clinically accessible radiology foundation model: Open-access and lightweight, with automated evaluation. arXiv preprint arXiv:2403.08002 (2024)
33. Zhang, X., Meng, Z., Lever, J., Ho, E.S.L.: CCD: Mitigating hallucinations in radiology MLLMs via clinical contrastive decoding. arXiv preprint arXiv:2509.23379 (2025)
34. Zhang, X., Meng, Z., Lever, J., Ho, E.S.L.: Libra: Leveraging temporal images for biomedical radiology analysis. In: Findings of the Association for Computational Linguistics: ACL 2025. pp. 17275–17303 (2025)
35. Zhang, Y., Jiang, H., Miura, Y., Manning, C.D., Langlotz, C.P.: Contrastive learning of medical visual representations from paired images and text. In: Proceedings of the 7th Machine Learning for Healthcare Conference. vol. 182, pp. 2–25 (2022)