Entity-Constrained CBCT Retrieval for Low-Resource Dental Record Completion

Nhi Ngoc-Yen Nguyen[1,*], Thai Nguyen[1,2,*], Kiet Huynh[1], and Huy-Hieu Pham[1,†]

[1] VinUni-Illinois Smart Health Center, VinUniversity, Hanoi, Vietnam
[2] Vietnam National University Ho Chi Minh City, University of Science, Vietnam
{nhi.nny, thai.nt, hieu.ph}@vinuni.edu.vn
hctkiet22@clc.fitus.edu.vn

Abstract. Completing dental records from cone-beam computed tomography (CBCT) is difficult when annotation is scarce and individual clinical fields are supported by different types of evidence. STSR 2026 Task 3 requires seven-field record completion from only 50 labeled CBCT cases and scores the correctness of structured FDI positions and ICD codes; consequently, a visually plausible retrieved record can still be harmful when it introduces an unsupported entity. We propose *Entity-Constrained CBCT-Guided Retrieval* (ECCR), a parameter-free framework that separates evidence availability from evidence authority. A corpus-derived prior first supplies the complete record. A frozen 3D encoder retrieves image-conditioned Diagnosis evidence, which is appended only if it does not expand the prior FDI or ICD entity set, so the asserted entity set is invariant by construction. On public validation, ECCR reaches a weighted score of 0.3134, improving on both full-record multimodal retrieval (0.2237) and a static text-only prior (0.2915); the guard blocks 63.3% of retrieved candidates, each of which would otherwise have injected an FDI position or ICD code absent from the prior. On the final test evaluation, ECCR obtains 11.37 of a 97.4-point attainable maximum, securing second place overall. The result indicates that, in an extreme low-resource setting, controlling what multimodal evidence is allowed to modify can be more reliable than transferring an entire retrieved record.

Keywords: Multimodal generation · Evidence-constrained retrieval · Cone-beam CT · Clinical record completion · Low-resource learning

1 Introduction

Cone-beam computed tomography (CBCT) is essential for dental diagnosis, yet converting 3D volumes into structured 7-field clinical records is challenging—particularly under the low-resource regime of STSR 2026 Task 3 (50 labeled cases). Granting retrieved evidence equal authority across fields risks hallucinating

* Equal contribution.
† Corresponding author: hieuph@vinuni.edu.vn

findings or overwriting accurate text, directly incurring penalties on strictly evaluated FDI tooth positions and ICD codes.

While medical vision–language models enable image-conditioned retrieval [1,12,3,4], unrestricted transfer fails under sparse supervision. In public validation, replacing records with M3D-CLIP neighbors yields 0.2237, underperforming a text-only BLEU-medoid prior (0.2915). Unrestricted retrieval can therefore provide useful context without being reliable enough to control every clinical field.

We propose *Entity-Constrained CBCT-Guided Retrieval* (ECCR). ECCR establishes a corpus-derived prior for all seven fields, retrieves a CBCT candidate via a frozen 3D encoder, and strictly limits updates to the Diagnosis text—accepting them only if no new FDI positions or ICD codes are introduced. Our contributions are:

1. We formulate low-resource CBCT record completion as an evidence-authority problem, separating retrieval capability from field modification permissions.
2. We introduce a parameter-free entity-constrained admission rule, with a proof that the asserted entity set is invariant under the update.
3. We provide challenge-oriented analysis through authority comparisons, update traces, offline ablations, and a full sub-metric decomposition of our second-ranked result, including a frank account of where the design loses points.

2 Related Work

3D medical image–text modelling. Large 3D vision–language models such as M3D and Med3DVLM connect volumetric images with medical text [1,12], and report generation frameworks such as CT2Rep and SAMF retrieve or synthesize radiological reports [3,4]. These models make image-conditioned evidence accessible but do not delineate which fields can be safely modified under extreme data scarcity. We leverage frozen 3D representations while explicitly governing field-level updates.

Retrieval, factuality, and structured evidence. Retrieval-augmented generation improves access to relevant evidence [6] but risks introducing unsupported clinical findings. Prior work enforces factuality through specialized training objectives [9,2] or structured representations such as RadGraph and knowledge graphs [5,14]. Since training a dedicated factuality model is infeasible with 50 cases, ECCR instead uses extracted FDI positions and ICD codes as a parameter-free, deterministic admission rule.

3 Dataset Analysis and Design Motivation

The challenge data contain labeled training, unlabeled training, and public-validation splits, with one NIfTI CBCT volume per case. The labeled CSV comprises 161 visit rows from 50 cases; we aggregate rows sharing a filename by concatenating their unique, non-empty field values, forming one structured record per volume.

Table 1. Field heterogeneity motivates entity-constrained retrieval update. Statistics cover 161 labeled visits; Filled (%) denotes non-empty rows and Unique counts distinct non-empty strings. *Oral check* corresponds to the *Oral exam* aspect of the official metric.

Field	Filled (%)	Unique	Observed structure
Handle	96	145	Near visit-specific
Oral check	93	146	Near visit-specific
Doctor advices	80	59	Strongly templated
Diagnosis	76	99	Diverse, recurrent groups
Treatment plan	59	88	Diverse, incomplete
Main appeal	34	50	Sparse
Present medical history	28	43	Sparse

Field heterogeneity. Table 1 shows why a uniform update policy is unsuitable. Doctor advices is strongly templated (59 unique values, one phrase appearing 33 times), Handle and Oral check are almost visit-specific, and Main appeal and Present medical history are frequently absent. Image evidence should therefore not be granted the same authority across all fields.

Structured-entity skew. The 759 tooth mentions cover 45 distinct two-digit tooth codes and the 116 ICD mentions cover 29 codes, both concentrated: teeth 16, 46, 17, 26, 36, 21, 38 and codes K00.002, K05.300, K07.302 dominate. This corpus statistic is computed with a broader two-digit extractor than the permanent range 11–48 on which the admission rule of Section 4.3 is defined, so the two counts are not directly comparable; the count reported here is descriptive of the corpus and is not the vocabulary over which the guarantee of Section 4.3 holds.

Geometry. All 50 labeled volumes share geometry $640 \times 640 \times 400$ at isotropic 0.25 mm spacing, so geometry does not explain variation in this split. Global resizing and pooling can obscure tooth-localized findings, motivating retrieval as optional evidence rather than complete-record replacement.

Field-to-metric mapping. The official Record-completeness aspect scores four sub-fields: main appeal, present history, past history, and doctor advices. Three map onto columns of our aggregated schema (Table 1), but past history has no corresponding column among the seven fields we aggregate, so our filename-level aggregation emits it empty for every case. The zero in Table 7 is therefore a property of the aggregation step, not of retrieval or admission, and any non-empty default would recover the credit.

4 Method

4.1 Problem Formulation

We distinguish evidence availability from evidence authority. Let x_i denote the CBCT volume for case i and $y_i = \{y_i^{(f)}\}_{f=1}^{7}$ its structured record. The system has

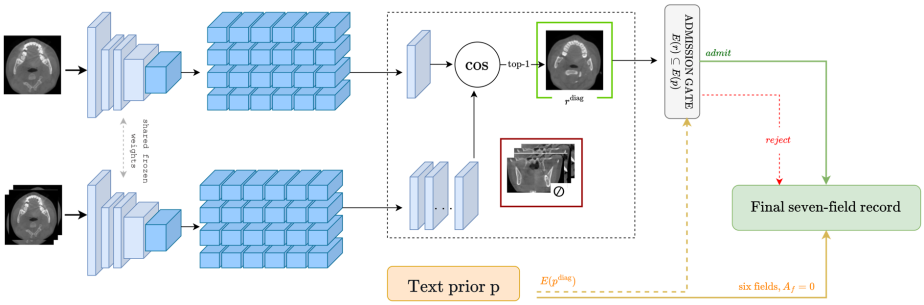

Fig. 1. Retrieved evidence is controlled rather than copied. A frozen 3D encoder retrieves a neighbour, while prior p supplies all seven fields. Retrieval may propose a *Diagnosis* update only, and admission requires $\mathcal{E}(r) \subseteq \mathcal{E}(p)$; rejection retains the prior.

access to a training-derived text prior $p = \{p^{(f)}\}$ and image-conditioned retrieved text $r_i = \{r_i^{(f)}\}$. An admission function $A_f(p, r_i, x_i) \in \{0, 1\}$ determines whether the retrieved evidence may modify field f:

$$\hat{y}_i^{(f)} = \begin{cases} U\big(p^{(f)}, r_i^{(f)}\big), & A_f(p, r_i, x_i) = 1, \\ p^{(f)}, & \text{otherwise}, \end{cases} \tag{1}$$

where U is deterministic. Full authority replaces the record, absent authority retains p, and bounded authority restricts admission by field and content.

4.2 Text Prior and Image Retrieval

For each field, we select its BLEU-medoid as a stable text prior and no-image control. Retrieval then supplies optional evidence; in the submitted ECCR system, only Diagnosis may be updated. We preprocess every volume with percentile clipping, min–max normalization, and trilinear resizing to $128 \times 256 \times 256$. A frozen Med3DVLM VisionTower [12] produces a feature tensor that is mean-pooled and ℓ_2-normalized. We standardize query and bank features with the bank mean and standard deviation before normalizing again. For query embedding z_i and bank embedding z_j, top-1 cosine retrieval selects

$$j^* = \arg\max_j \frac{z_i^\top z_j}{\|z_i\|_2 \|z_j\|_2}, \tag{2}$$

and returns $r_i = y_{j^*}$. The encoder is frozen and the 50 bank features are cached, so the pipeline requires no task-specific training.

4.3 Entity-Constrained Diagnosis Update

As illustrated in Figure 1, image evidence proposes rather than replaces a Diagnosis. For every non-Diagnosis field, $A_f = 0$. If the candidate is admitted, we append it to the prior after exact-duplicate removal; otherwise the prior is retained. With $k = 1$, neighbour agreement is degenerate, so Diagnosis-only authority and entity consistency are the operative controls.

Table 2. System configuration. The proposed pipeline is parameter-free and requires only frozen encoder inference over the cached bank and query volumes.

Component	Configuration
GPU	1× NVIDIA GeForce RTX 4090 (24 GB)
Encoder	Med3DVLM VisionTower [12] (frozen)
Optimization	No training; parameter-free inference

Entity extraction and admission. Let $\mathcal{E}(s)$ extract FDI positions 11–48 and dental ICD-10 codes matching `K`$dd\,.\,ddd$ from a Diagnosis string s; the guarantee is deliberately limited to these measured entity types and establishes no recall property for free-text clinical concepts. The candidate is eligible only when

$$\mathcal{E}\left(r_i^{(\mathrm{diag})}\right) \subseteq \mathcal{E}\left(p^{(\mathrm{diag})}\right), \tag{3}$$

so an admitted candidate may repeat or refine an asserted entity but cannot add a new one.

Entity-set invariant. Let $d = p^{(\mathrm{diag})}$ and $c = r_i^{(\mathrm{diag})}$. The update returns either d or $d \oplus c$, where $\oplus$ denotes concatenation with a separator. Since $\mathcal{E}(d \oplus c) = \mathcal{E}(d) \cup \mathcal{E}(c)$, Equation (3) gives

$$\mathcal{E}\left(\hat{y}_i^{(\mathrm{diag})}\right) = \mathcal{E}(d). \tag{4}$$

Retrieval therefore cannot alter the detected entity set. The prior bounds entity precision, while its coverage caps entity recall.

5 Experiments

5.1 Protocol and Implementation

Our primary endpoint is the public-validation weighted score computed by Codabench against hidden references [13]; because the split is fixed, it has no variance estimate. We additionally use a 50-case leave-one-out (LOO) proxy averaging four smoothed BLEU scores, ROUGE-L F1, tooth F1, and diagnosis-code F1. The proxy omits the official corpus-level diversity factor and is used only to compare update rules. The environment is given in Table 2.

Retrieval spaces and encoder choice. Full-record transfer uses the joint image–text space of M3D-CLIP [1] to match complete reports, whereas bounded admission retrieves candidate Diagnosis text from pooled Med3DVLM features [12] and relies on the text prior elsewhere. Comparing ECCR against the text-only prior therefore isolates the effect of image-conditioned evidence, since both share identical non-Diagnosis fields.

Official scoring. Four aspects are scored on a 100-point scale with weights 0.45 (Diagnosis), 0.20 (Oral exam), 0.20 (Treatment plan), 0.15 (Record completeness). A submission reproducing the ground truth exactly attains 97.4 rather than 100,

Table 3. Public-validation comparison of evidence authority (0–1 scale). The update column states what each system may modify.

System	Retrieval update	Score
M3D-CLIP retrieval	Full: replaces the complete record	0.2237
BLEU-medoid	None: text prior only	0.2915
Ours	Diagnosis only; entity-consistent	**0.3134**

so all totals should be read against that ceiling. The weighted content sum is then multiplied by a corpus-level template-diversity factor

$$\phi = 0.5 + 0.5\, d, \tag{5}$$

where d is the fraction of distinct Diagnosis texts across the 50 cases; identical long texts repeated across cases additionally have their tooth-level content excluded. A single repeated template thus incurs a double penalty, which shapes the analysis below.

5.2 Evidence-Authority Comparison and Update Trace

Table 3 tests the paper's central design choice. Replacing the full record with an M3D-CLIP neighbour gives 0.2237, 0.0678 below the text-only prior. Allowing only entity-compatible Diagnosis evidence raises the score by 0.0219 to 0.3134, second on the public leaderboard. The result supports selective evidence use; it does not, because of the encoder difference above, isolate authority as the sole cause of the full-record gap.

Update-rule ablation. Table 4 examines why the final update appends rather than replaces. Replacement is clearly harmful, costing 0.0289. Naive and entity-safe appending differ by 0.0002—below the resolution of a proxy with no term penalising an unsupported entity. The rightmost column supplies that missing term: across the 50 LOO folds the retrieved candidate carries an FDI or ICD entity absent from the prior in 40 cases, which both replacement and naive appending propagate, whereas Equation (4) guarantees zero injections. Since the official Diagnosis aspect is scored by entity-level F1, each injected entity is a false positive: the two rules are equivalent on the proxy and not on the metric that decides the challenge.

Admission trace. Table 5 traces the guard on the 49 public-validation queries: the constraint intervenes in 63.3% of retrievals, blocking 31 candidates and admitting 18 (36.7%). The two traces are computed on different sets—50 LOO folds over the labeled split and 49 validation queries—and are reported separately throughout.

Blocked candidates are not concentrated at low similarity: they occur throughout the observed top-1 cosine range and overlap the admitted ones throughout, so a similarity threshold cannot substitute for entity-aware verification. Retrieval capability and evidence authority therefore address distinct failure modes.

Table 4. Update-rule ablation on the strict LOO proxy (50 labeled cases, 0–1 scale, no corpus-level term). *Injections* counts folds whose emitted Diagnosis contains an FDI or ICD entity absent from the prior.

Method	Proxy	Inject.
Text prior	0.1831	0/50
Diagnosis replace	0.1542	40/50
Naive append	**0.1833**	40/50
Entity-safe append (ours)	0.1831	**0/50**

Table 5. Admission trace over the 49 public-validation cases. Conflict denotes an FDI or ICD entity absent from the prior. The gate row is degenerate at $k = 1$.

Stage / conflict	Cases	%
Retrieved evidence	49	100.0
Gate passed	49	100.0
Blocked: FDI only	6	12.2
Blocked: ICD only	16	32.7
Blocked: both	9	18.4
Final admitted	18	36.7

5.3 Reconciling the Offline Proxy with the Official Score

Smoothed BLEU and ROUGE-L reward n-gram overlap with a single reference and impose no penalty for an unsupported entity, so appending a short compatible clause is near metric-neutral by construction. The parity in Table 4 is therefore consistent with local updates preserving lexical quality; it is not evidence of the safety property, which Equation (4) establishes by construction.

The official validation gain (+0.0219) is largely attributable to Equation (5). The text-only prior emits one Diagnosis template for all cases ($d = 1/50$, $\phi = 0.51$); admitting entity-safe evidence expands this into six distinct phrasing patterns ($d = 0.12$, $\phi = 0.56$). Rescaling the prior by $0.56/0.51$ gives $0.2915 \times 1.098 \approx 0.320$, which brackets the observed 0.3134. We therefore do not claim the gain reflects improved clinical content; it reflects a legitimate reduction in template repetition obtained without altering the asserted entity set. The same mechanism exposes a structural limit of a static prior: because repeated baseline text is excluded from tooth-level scoring, a medoid prior caps tooth-level Diagnosis F1 at 0.016 (Table 7), the single largest source of forfeited Diagnosis credit.

5.4 Final Leaderboard Result and Score Decomposition

Under strict penalties for entity mismatch and template repetition, scores are heavily compressed: a ground-truth submission would score 97.4, yet the top entry reaches 17.70 (Table 6). ECCR obtains 11.37, ranking second, ahead of the third-placed system despite a lower Diagnosis sub-score because Record completeness contributes 3.76 against 1.68.

Table 7 makes the failure modes explicit rather than implicit. Diagnosis carries the submission at 4.89 points, but almost all of it comes from condition-level and whole-mouth detection: tooth–condition matching contributes 0.16 of a possible 10, the direct consequence of the medoid cap in Section 5.3. The Oral exam aspect scores zero, forfeiting 20% of the available weight. This is a genuine cost of restricting A_f to Diagnosis, not a design benefit: Table 1 shows Oral check is near visit-specific (146 unique values in 161 visits), so a corpus medoid transfers almost nothing, and we did not extend the admission rule to it because no entity guard was defined for oral findings.

Table 6. Final Task 3 leaderboard under the 100-point protocol (attainable maximum 97.4). ECCR ranks second.

Rank	Team	Diagnosis	Oral exam	Treatment	Completeness	Total
1	abc-abc	5.56	1.49	4.71	5.95	17.70
2	**Ours**	4.89	0.00	2.72	3.76	**11.37**
3	sjtu426lab-task3	5.29	0.00	3.12	1.68	10.08
4	jlshen	4.50	0.00	2.07	3.50	10.06
Aspect weight		0.45	0.20	0.20	0.15	1.00

Table 7. Sub-metric decomposition of ECCR's final score. Points are on the 100-point scale before the diversity factor; Credited applies $\phi = 0.56$. Completeness sub-metrics combine a completion rate and ROUGE-L and are reported at aspect level only.

Aspect (Points / Credited)		Sub-metric	n	F1	Points
Diagnosis	8.73 / **4.89**	Condition detection	47	0.2345	4.69
		Tooth–condition match	32	0.0159	0.16
		ICD codes	40	0.1630	1.63
		Whole-mouth	20	0.4500	2.25
Oral exam	0.00 / **0.00**	Findings detection	34	0.0000	0.00
		Tooth-level findings	21	0.0000	0.00
		Whole-mouth findings	7	0.0000	0.00
Treatment	4.85 / **2.72**	Action detection	42	0.4762	4.76
		Tooth-level actions	32	0.0000	0.00
		Handled actions	29	0.0172	0.09
Completeness	6.72 / **3.76**	(main appeal 3.18, present 1.91, past 0.00, advices 1.64)			
Total					20.30 → **11.37**

Finally, Past history scores zero on both completion and ROUGE-L for the structural reason given in Section 3—an implementation gap worth 1.5 points before the diversity factor. Two of the three largest losses are thus addressable without changing the framework.

6 Conclusion and Limitations

ECCR combines a text prior with CBCT retrieval under governed update authority. Bounding candidate Diagnosis updates with entity preservation enriches clinical detail while preventing unverified information injection, giving a parameter-free framework for low-resource clinical record completion that requires no task-specific training.

Four limitations should be stated plainly. (i) The guarantee of Equation (4) covers only FDI positions in the permanent range 11–48 and dental ICD-10 codes, and establishes no property for free-text clinical concepts. Two-digit codes outside this range are not recognised by $\mathcal{E}(\cdot)$ and therefore fall outside the invariant; widening the entity vocabulary is a direct, parameter-free extension. (ii) The

invariant is one-sided: it bounds entity precision by the prior but caps recall by the prior's coverage (tooth-level Diagnosis F1 = 0.016). (iii) Restricting authority to Diagnosis forfeits the Oral exam aspect (20% of the weight); a per-field guard for oral findings is the most valuable extension we identify. (iv) The comparison against full-record retrieval confounds authority with encoder choice (M3D-CLIP versus Med3DVLM), and the fixed single split admits no variance estimate. Future work will address region-aware volumetric representations, per-field admission rules with field-specific entity vocabularies, and learned confidence thresholds. Code will be released publicly upon publication.

Acknowledgements

This work was developed and evaluated within the ODIN 2026 challenge cluster, including the STS, ToothFairy, and Bite2Text challenges [11,7,8,10].

Disclosure of Interests

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

References

1. Bai, F., Du, Y., Huang, T., Meng, M.Q.H., Zhao, B.: M3D: Advancing 3d medical image analysis with multi-modal large language models. arXiv preprint arXiv:2404.00578 (2024) 2, 5
2. Delbrouck, J.B., Chambon, P., Blüthgen, C., Tsai, E.B., Almusa, O., Langlotz, C.P.: Improving the factual correctness of radiology report generation with semantic rewards. In: Findings of the Association for Computational Linguistics: EMNLP 2022. pp. 4348–4360. Association for Computational Linguistics (2022). https://doi.org/10.18653/v1/2022.findings-emnlp.319, https://aclanthology.org/2022.findings-emnlp.319/ 2
3. Hamamci, I.E., Er, S., Menze, B.: Ct2rep: Automated radiology report generation for 3d medical imaging. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2024. pp. 476–486. Springer Nature Switzerland (2024). https://doi.org/10.1007/978-3-031-72390-2_45 2
4. Hosseini, A., Ibrahim, A., Serag, A.: From slices to volumes: Multi-scale fusion of 2d and 3d features for ct scan report generation. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2025. Springer Nature Switzerland (2025). https://doi.org/10.1007/978-3-032-04978-0_26 2
5. Jain, S., Agrawal, A., Saporta, A., Truong, S.Q.H., Duong, D.N., Bui, T., Chambon, P., Lungren, M., Ng, A., Langlotz, C., Rajpurkar, P.: Radgraph: Extracting clinical entities and relations from radiology reports. PhysioNet (2021). https://doi.org/10.13026/hm87-5p47, https://physionet.org/content/radgraph/1.0.0/ 2
6. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., tau Yih, W., Rocktäschel, T., Riedel, S., Kiela, D.: Retrieval-augmented generation for knowledge-intensive NLP tasks. In: Advances in Neural Information Processing Systems. vol. 33, pp. 9459–9474 (2020) 2
7. Lumetti, L., Di Bartolomeo, M., Pellacani, A., Anesi, A., Grana, C., Bolelli, F.: Ontology-grounded structured prediction for dental CBCT reporting. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2026 (2026) 9
8. Lumetti, L., Rizzo, F., Cremonini, F., Candeloro, E., Luca, L., Grana, C., Bolelli, F.: Do multimodal LLMs understand intraoral dental data? dataset, platform, and baselines. In: European Conference on Computer Vision (ECCV) (2026) 9
9. Miura, Y., Zhang, Y., Tsai, E.B., Langlotz, C.P., Jurafsky, D.: Improving factual completeness and consistency of image-to-text radiology report generation. In: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics (2021). https://doi.org/10.18653/v1/2021.naacl-main.416, https://aclanthology.org/2021.naacl-main.416/ 2
10. Wang, Y., Li, Z., Wu, C., Liu, J., Zhang, Y., Ni, J., Luo, Q., Chen, J., Zhang, H., Liu, J., et al.: MICCAI STS 2024 challenge: Semi-supervised instance-level tooth segmentation in panoramic X-ray and CBCT images. Medical Image Analysis p. 103986 (2026) 9
11. Wang, Y., Zhang, Y., Chen, X., Wang, S., Qian, D., Ye, F., Xu, F., Zhang, H., Dan, R., Zhang, Q., et al.: MICCAI 2023 STS challenge: A retrospective study of semi-supervised approaches for teeth segmentation. Pattern Recognition **170**, 112049 (2026) 9
12. Xin, Y., Ates, G.C., Gong, K., Shao, W.: Med3DVLM: An efficient vision-language model for 3d medical image analysis. arXiv preprint arXiv:2503.20047 (2025) 2, 4, 5

13. Xu, Z., Escalera, S., Pavão, A., Richard, M., Tu, W.W., Yao, Q., Zhao, H., Guyon, I.: Codabench: Flexible, easy-to-use, and reproducible meta-benchmark platform. Patterns **3**(7), 100543 (2022) 5
14. Zhang, Y., Wang, X., Xu, Z., Yu, Q., Yuille, A., Xu, D.: When radiology report generation meets knowledge graph. arXiv preprint arXiv:2002.08277 (2020), https://arxiv.org/abs/2002.08277 2