CIRCA: Maximising Per-Slice Detail for Head-CT Report Generation with a Windowed-RGB and Split-Slab Bone-Projection Representation for the HEADLINE Challenge

Louis Bellmann*, Sören Spiegel*, and Maximilian Nielsen

Institute for Applied Medical Informatics, University Medical Center Hamburg-Eppendorf (UKE), Hamburg, Germany
{l.bellmann, s.spiegel, m.nielsen}@uke.de
* These authors contributed equally.

Abstract. We address automatic radiology-report generation from non-contrast head CT for the HEADLINE challenge, studying the input representation rather than model scale as a lever for the small, low-contrast findings that drive clinical value. We present **CIRCA** (*CT Imaging and Radiograph Composition for Autoreporting*), a representation that (i) fuses the three redundant soft-tissue CT windows (brain, subdural, stroke) into the R, G, B channels of one image, (ii) presents the per-slice sequence as a short video so that all axial slices fit the deployment token budget at 448×448, and (iii) gives the sharp bone reconstruction its own six-view split-slab maximum-intensity projection ("skull radiographs") instead of a colour channel. A Qwen2.5-VL-7B backbone is LoRA-fine-tuned on both the vision encoder and the language model and deployed in 8-bit on a 24 GB GPU; it ranked #1 on the challenge validation leaderboard (automated score **0.2675**, +0.0214 over the runner-up) and placed among the three top teams in the final test phase. Internally, the representation beats an axial-slice montage at a matched epoch, 8-bit is lossless relative to bfloat16, and dropping the video pair-merge does not help – which, with a finding-level error analysis, points to in-plane low-contrast perception rather than through-plane packaging as the limiting factor.

Keywords: Radiology report generation · Head CT · Vision–language models · LoRA · Medical image representation.

1 Introduction

The HEADLINE challenge[1] (CCI Bonn / University Hospital Bonn) targets automatic generation of radiology reports from non-contrast head CT. The reference reports are German clinical reports translated into English, describing intracranial haemorrhage, ischaemic infarction, skull fractures and mass

[1] https://headline.ccibonn.ai/

effect; submissions are ranked by an automated score combining a rescaled BERTScore [1] with lexical-overlap metrics. The task is hard for two reasons. First, many of the most clinically important findings are small, low-contrast and confined to one or two slices – an early infarct, a subtle subarachnoid bleed – so they are easily missed. Second, the reports use terse, centre-specific phrasing that the model must reproduce without overfitting to it.

Our contribution is a recipe for adapting a general-purpose vision–language model to this task within the challenge's tight deployment limits. Holding the 7B backbone fixed, we treat the per-slice detail reaching the encoder as our design variable and contribute CIRCA: an input representation that (a) fuses three soft-tissue Hounsfield windows (brain, subdural/blood, stroke) into the R, G and B channels of one image, since all three are contrast stretches of the same anatomy and together cost no extra tokens, (b) gives the bone reconstruction its own projection views rather than a colour channel, and (c) maximises in-plane resolution within the deployment budget.

2 Related work

3-D CT report generation. Native volumetric encoders dominate CT-to-report generation: CT2Rep is an auto-regressive 3-D report generator [2], CT-CLIP/CT-RATE aligns whole chest-CT volumes with reports at scale [3], and RadFM [4] trains a 3-D vision–language foundation model on a large volume–text corpus. These are predominantly chest-CT, and each requires pretraining a volumetric encoder. We take a different route, for two reasons: the challenge runs offline under a fixed GPU and time budget, which puts such pretraining out of reach, and the target is free-text reporting style, which rewards a pretrained language model. We therefore keep a general 2-D vision–language model (Qwen2.5-VL [5]) and adapt it through the input representation alone, complementary to the 3-D route rather than a replacement.

Visual encoders and windowed inputs. Two design choices are easy to conflate, so we separate them. The *visual encoder* is the pretrained network mapping images to features; medically pretrained encoders such as MedGemma [6] aim to make those features domain-aligned. The *input representation* is what we feed that encoder: which Hounsfield windows, at which resolution, packed into how many images. Our experiments vary only the second, holding the encoder fixed (Section 4.2), so we make no claim about how the two compare; a controlled encoder swap remains future work. Encoding Hounsfield windows as RGB channels of one image is established for head-CT haemorrhage detection [7,8]; we adapt this window-as-channel idea to reporting (Section 3.2).

3 Methods

3.1 Dataset and task

The HEADLINE dataset collects real non-contrast head-CT studies acquired and reported in a single neuroradiology department, each paired with its clinical report in a *Findings / Assessment* structure (81% carry an explicit Findings section, 70% an Assessment section; typical length 115–140 words). For internal comparisons we hold out a fixed random 500-study development split and train on the remaining 15,968 released studies; leaderboard-comparable numbers come only from the organiser's separate official validation set (322 patients, 500 studies, no patient overlap).

Each study provides two reconstructions, and we use both. The smooth soft-tissue reconstruction (4 mm slices) is our source for every soft-tissue view; the sharp bone reconstruction ($\sim$1 mm) carries true fracture detail but has a $\sim$12$\times$ higher soft-tissue noise floor, so we use it only for the bone views. About 46% of studies are follow-ups whose reports may mention prior imaging; as we process each study in isolation, the model reproduces such references only implicitly. The score S_{Auto} combines a rescaled BERTScore [1] with BLEU-4 [9], METEOR [10] and ROUGE-L [11], per study and averaged:

$$S_{\text{Auto}} = 0.7\,\text{BERTScore}_{\text{rescaled}} + 0.3\,\text{mean}(\text{BLEU-4}, \text{METEOR}, \text{ROUGE-L}). \quad (1)$$

3.2 Input representation

Each study becomes a compact set of images matched to how the encoder consumes detail. Every soft-tissue (4 mm) axial slice is cropped to the head, padded square, and resized to 448×448; we keep all of them, as the series is short and one- or two-slice findings are easily lost to subsampling.

Three-window RGB fusion. The same slice is windowed three ways – brain (level 40, width 80), subdural/blood (65, 200), and stroke (35, 30) – and the three grayscale results become the R, G and B channels of one image (Fig. 1, top). One fused image lets the model attend to parenchyma, extra-axial blood and early stroke at the token cost of a single slice.

Bone as composed skull radiographs. A common head-CT convention stacks brain, subdural and bone windows as R/G/B; we instead free the third channel for the stroke window and give bone its own images. Decisive bone findings – fractures, lucent or sclerotic lesions, hardware – are contour and density features read most naturally in projection, like a plain skull radiograph. We therefore *compose* the sharp 1 mm reconstruction into radiograph-like views (the "Radiograph Composition" of CIRCA). A whole-head MIP would superimpose the near and far skull tables, so we split the head into anterior/posterior, left/right and superior/inferior half-slabs and MIP each half, giving six 448×448 "skull radiographs" (Fig. 1, middle).

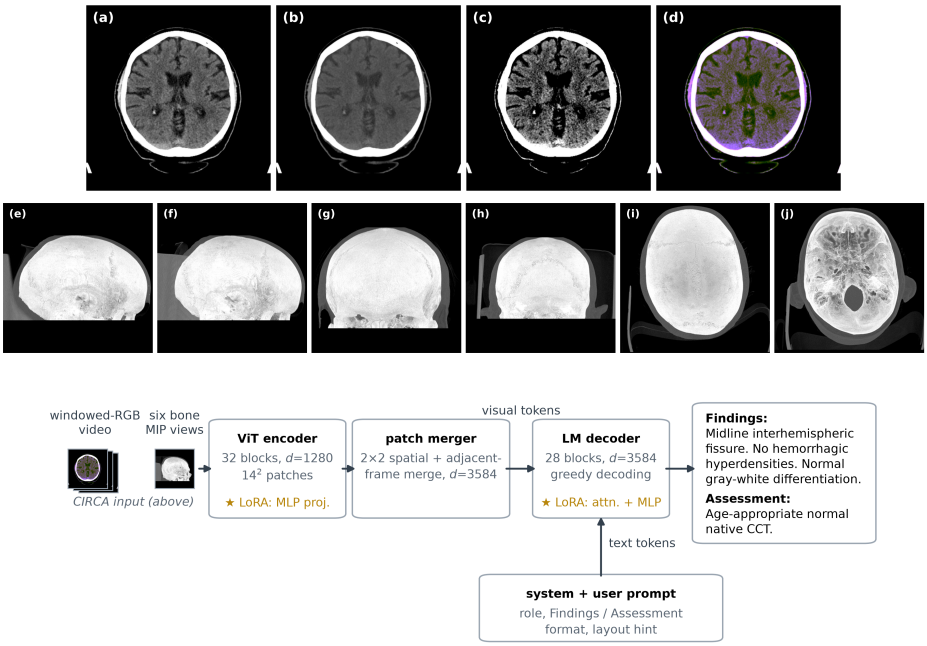

Fig. 1. CIRCA. *Top:* three-window RGB fusion of one soft-tissue slice – (a) brain, (b) subdural/blood, (c) stroke window, (d) the three stacked as R/G/B. Shown in standard radiological orientation (anterior up, patient left on the image right); the model consumes the unrotated frames. *Middle:* six-view split-slab MIP of the sharp 1 mm bone reconstruction – (e) left, (f) right, (g) anterior, (h) posterior, (i) superior, (j) inferior. *Bottom:* the pipeline of Section 3.3; stars mark the LoRA-adapted projections, all other weights stay frozen.

Video packing for resolution. Rather than passing each slice as a separate image, we use the backbone's native video input, which merges adjacent slices and roughly halves their visual-token cost. We spend the saved budget on resolution, keeping every slice at 448×448 where a montage would have to shrink it. The merge blends a little through-plane detail; Section 4.3 ablates this.

3.3 Model, prompt and fine-tuning

The backbone is Qwen2.5-VL-7B-Instruct [5], a vision–language model with a dynamic-resolution encoder and native multi-image and video input. Figure 1 (bottom) shows how the CIRCA images enter it: the ViT encoder produces patch tokens, a patch merger projects them to the language-model width, and the visual tokens are interleaved with the prompt's text tokens in the decoder input.

Instruction. Training and inference share one prompt, so the model is always queried exactly as it was trained. The system prompt fixes the role, output structure and target length, and does specify the Findings/Assessment format explicitly:

"You are a neuroradiologist writing a clinical CT head report. Output ONLY two sections: `**Findings:**` 2–5 sentences describing what you observe anatomically (brain parenchyma, ventricles, extra-axial spaces, skull/calvaria, sinuses). Report findings, NOT image channels. Name any hemorrhage, infarct, mass effect, midline shift, hardware, or postoperative change you see. `**Assessment:**` 1–2 sentence clinical summary. Rules: no [Insert ...] placeholders, no disclaimers or recommendations to consult a physician, no meta-commentary about imaging technique or channels. Target length: 80–150 words total."

The user turn carries the vision blocks, followed by a one-line hint naming them in their fixed order – "This head CT is shown as a video of all axial soft-tissue slices (inferior to superior), followed by six bone-window MIP views of the skull (left, right, anterior, posterior, superior, inferior)." – and then the request, "Report the findings on this head CT." Loss is computed on the assistant turn only.

LoRA configuration. We apply LoRA [12] with rank 16, $\alpha = 32$, dropout 0.05 and no bias terms, targeting the seven projection names `q_proj`, `k_proj`, `v_proj`, `o_proj`, `gate_proj`, `up_proj` and `down_proj`. In Qwen2.5-VL these match every attention and MLP projection of the 28 language-model blocks, and the MLP projections of the 32 vision-encoder blocks; the encoder's fused attention projections (`qkv`, `proj`) carry different names and stay frozen, as do the patch merger and all base weights. This yields 47 589 376 trainable parameters (0.57% of 8.34 B), 15% of them inside the vision encoder, enough to adapt it to the false-colour input while preserving vision–language alignment. Rank, α, dropout and learning rate follow common LoRA practice rather than a sweep of our own; the compute went to the representation comparisons instead.

Training, inference and deployment. Training runs for 3 epochs (batch size 1, gradient accumulation 8, learning rate 2×10^{-4}, cosine schedule with 5% warmup, weight decay 0.01, bfloat16, gradient checkpointing) on a single NVIDIA RTX PRO 6000 at ∼18 h/epoch; training and inference share one preprocessing path. At inference we decode greedily. If generation fails or no 4 mm series is found we substitute a fixed *fallback report* rather than emit an empty string, which the evaluator scores as zero: a generic, structurally valid Findings/Assessment text stating no acute intracranial abnormality, assembled from the most frequent sentences in the training corpus. It is a failure guard only and was never triggered – 0 of 500 studies internally and 0 of 500 on the official validation submission. Of the two containerized deployment GPUs (16 GB and 24 GB, 20-minute budget per job) we use the 24 GB one in 8-bit integer quantization [13] at 25–40 s per study; at full per-slice coverage the 8-bit model does not fit 16 GB, which would need a 4-bit fallback.

Table 1. Official validation leaderboard (500 studies), final model: ranked #1, +0.0214 over the second-placed entry.

S_{Auto}	Rescaled BERTScore-F1	ROUGE-L	METEOR	BLEU-4
0.2675	0.2968	0.3033	0.1734	0.1206

Table 2. Representation comparison at matched epoch 2 (internal 500-study split, 8-bit, raw metrics).

Representation	BERTScore-F1 (raw)	S_{Auto} (raw)
Axial montage (bone in colour channel, ∼224 px)	0.7478	0.5751
CIRCA (RGB video + bone MIP, 448 px)	**0.7562**	**0.5915**

4 Results

4.1 Official validation leaderboard

Our fine-tuned model (epoch 3) ranked #1 on the challenge validation leaderboard (Table 1, all 500 validation studies). The composition follows Eq. (1); the +0.0214 margin over the second-placed entry is concentrated in the rescaled BERTScore, rescaled against a fixed baseline $b = 0.7727$ confirmed from the official per-study values. Our entry also placed among the three top teams in the final test phase; all numbers reported here are from the validation phase and our internal development set. The choice between the two deployment GPUs is a quantization choice and costs accuracy: internally, varying only quantization on the shipped checkpoint, 8-bit reaches raw S_{Auto} 0.5936 (BERTScore-F1 0.7594) against 0.5890 (0.7568) for the 4-bit (NF4) fallback.

4.2 Internal development set: checkpoint and representation

On the full 500-study development split, in the shipped 8-bit configuration, raw S_{Auto} improves monotonically over training – 0.5788, 0.5915 and 0.5936 at epochs 1, 2 and 3 – so we ship epoch 3; 8-bit is lossless, as int8 slightly exceeded bfloat16 in a matched 100-study check (BERTScore-F1 0.7621 vs 0.7600). Against an axial-slice montage at matched epoch 2 (Table 2), CIRCA gains +0.0084 BERTScore-F1 and +0.0164 raw S_{Auto}. This bundles three changes – bone in dedicated projection views, the added stroke window and 448-px resolution – so it evidences the representation as a whole, not resolution alone; the visual encoder is identical on both sides. Our internal BERTScore backbone (a biomedical BERT [14]) is on a different raw scale than the official evaluator, so these numbers rank only relatively.

4.3 Video pair-merge ablation and error analysis

The video merge blends ∼8 mm through-plane, the scale of the one- or two-slice findings the task turns on, making it the prime suspect for the residual error.

We test it directly, at epoch 3 on the full 500-study set with identical 448×448 slices and six bone views, packed either as the shipped video or as one image per slice (≈10.7k tokens). Dropping the merge moves raw S_{Auto} from 0.5890 to 0.5912 (+0.0022) – smaller than merely re-quantizing the identical checkpoint from bfloat16 to int8, i.e. within run-to-run noise. The merge is therefore not what limits accuracy here, and we keep the cheaper video representation.

A finding-level analysis on 100 held-out studies points the same way. Detecting a 12-category finding taxonomy by negation-aware keyword matching, applied identically to generated and reference reports, recall splits sharply by contrast: high-contrast structural findings are recovered well (hardware 0.85, post-surgical change 0.75, mass effect 0.68), whereas low-contrast parenchymal ones are largely missed (haemorrhage ≈0.5, infarct 0.29, oedema ≈0); cases with at least one missed finding score 0.237 rescaled S_{Auto} against 0.307 for cases with none. Keyword matching over-counts misses where the model uses other vocabulary (oedema as "effacement" or "hypodensity"), and the analysis is correlational; with those caveats the pattern points to an in-plane, low-contrast perception limit rather than to the packaging or to wording alone.

4.4 Qualitative examples

In Table 3(a) the model recovers the high-contrast structural content: it places the new external ventricular drain on the correct side and horn, carries over the known subarachnoid haemorrhage, and adds an Assessment section the reference lacks. In (b) it names the main finding – a left basal-ganglia haemorrhage with ventricular rupture – but shows both dominant error modes: it invents a comparison with a prior study, a default learned from the 46% of follow-ups, and omits the subtler 3 mm midline shift and the older post-ischaemic defect.

5 Conclusion

For HEADLINE, where many clinically decisive findings are small, low-contrast and confined to one or two axial slices, we held the 7B backbone fixed and

Table 3. Two held-out studies: reference (left) vs. generated report (right). Abridged; "[. . .]" marks elided routine negatives.

(a) Typical success: hardware and known blood localised correctly.	
Findings: [. . .] A newly inserted steel-needle EVD from the right frontal, passing through the right lateral ventricle anterior horn [. . .]	**Findings:** [. . .] a steel-needle EVD was placed in the right frontal region, tip in the right lateral ventricle anterior horn. [. . .] **Assessment:** Normal position of the newly inserted EVD.
(b) Typical failure: invented prior comparison; subtle findings missed.	
Findings: No comparable previous imaging. Basal ganglia hemorrhage on the left with ventricular rupture and discrete midline shift up to 3 mm to the right [. . .]	**Findings:** For comparison, an external native CCT from [DATE] is available. In comparison, a slight redistribution of the known left basal ganglia hemorrhage [. . .]

worked on the input representation. CIRCA fuses the three soft-tissue windows into one RGB image, gives the sharp bone reconstruction six split-slab skull radiographs rather than a colour channel, and packs the slices as a video to keep 448-px resolution within the deployment budget; LoRA-fine-tuned, it ranked #1 on the validation leaderboard and finished among the three top teams in the final test phase. Our evidence varies the representation under one fixed encoder, on one centre's data, with three factors changing at once; isolating them, comparing against a medically pretrained or native 3-D encoder, and suppressing the invented prior-study comparisons are the natural next steps.

References

1. Zhang, T., et al.: BERTScore: Evaluating text generation with BERT. In: International Conference on Learning Representations (ICLR) (2020)
2. 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). https://doi.org/10.1007/978-3-031-72390-2_45
3. Hamamci, I.E., et al.: A foundation model utilizing chest CT volumes and radiology reports for supervised-level zero-shot detection of abnormalities (CT-CLIP / CT-RATE) (2024)
4. Wu, C., Zhang, X., Zhang, Y., Wang, Y., Xie, W.: Towards generalist foundation model for radiology (RadFM) (2023)
5. Bai, S., et al.: Qwen2.5-VL technical report (2025)
6. Google Research and Google DeepMind: MedGemma technical report (2025)
7. Umapathy, S., et al.: Automated computer-aided detection and classification of intracranial hemorrhage using ensemble deep learning techniques. Diagnostics **13**(18), 2987 (2023)
8. Songsaeng, D., et al.: HU to RGB transformation with automatic windows selection for intracranial hemorrhage classification using ncCT. PLOS ONE **20**(8), e0327871 (2025)
9. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: BLEU: a method for automatic evaluation of machine translation. In: Annual Meeting of the Association for Computational Linguistics (ACL) (2002)
10. Banerjee, S., Lavie, A.: METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In: ACL Workshop on Intrinsic and Extrinsic Evaluation Measures (2005)
11. Lin, C.Y.: ROUGE: A package for automatic evaluation of summaries. In: Text Summarization Branches Out (2004)
12. Hu, E.J., et al.: LoRA: Low-rank adaptation of large language models. In: International Conference on Learning Representations (ICLR) (2022)
13. Dettmers, T., et al.: LLM.int8(): 8-bit matrix multiplication for transformers at scale. In: Advances in Neural Information Processing Systems (NeurIPS) (2022)
14. Gu, Y., et al.: Domain-specific language model pretraining for biomedical natural language processing. ACM Transactions on Computing for Healthcare **3**(1) (2021)