A Unified Brain MRI Reporting Framework Built on CoT-Guided VLM and Expert Models

Yuxiao Liu[1†], Xin Lin[1†], Yaping Wu[4,5], Xintong Wu[2], Yichu He[2], Zhaoyu Qiu[1], Qiankun Zheng[1], Yan Bai[4,5], Wei Wei[4,5], Qingxia Wu[4,5], Zengyang Che[4,5], Dijia Wu[2], Feng Shi[2], Kaicong Sun[1], Meiyun Wang[4,5], and Dinggang Shen[1,2,3(✉)]

[1] School of Biomedical Engineering & State Key Laboratory of Advanced Medical Materials and Devices, ShanghaiTech University, Shanghai 201210, China
[2] Shanghai United Imaging Intelligence Co., Ltd., Shanghai 200230, China
[3] Shanghai Clinical Research and Trial Center, Shanghai 201210, China
[4] Department of Radiology, Henan Provincial People's Hospital, Zhengzhou, China
[5] Biomedical Research Institute, Henan Academy of Sciences, Zhengzhou, China
Dinggang.Shen@gmail.com

Abstract. Accurate brain MRI reporting is critical for managing brain disorders but remains time-consuming, motivating the need for automatic reporting. Current VLM-based reporting models often hallucinate unsupported findings or miss subtle abnormalities due to the lack of lesion-wise verification. To address this issue, we introduce **B**rain **R**eporting with **A**djudication and **V**erification–**E**nhanced framework **(BRAVE)** for brain MRI report generation. BRAVE first leverages lesion-specific expert models to generate structured lesion proposals and then performs lesion-wise chain-of-thought (CoT) adjudication to verify, refine, and reconcile these proposals before final report generation. Trained on 150,106 MRI scans and validated on 7,496 scans (spanning 19 lesion subtypes), BRAVE achieves 86.7% accuracy in lesion localization and characterization. The stepwise CoT adjudication reduces false-negative and false-positive errors by 38.1%. In a three-week prospective clinical deployment, BRAVE reduced resident's drafting time by up to 15.4% and consistently shortened attending's sign-off time by 17.4–32.8%, thereby supporting more efficient radiology workflows.

Keywords: Brain MRI reporting, CoT, Lesion-wise verification, VLM, Expert model.

1 Introduction

Radiology reports are indispensable for brain disorder diagnosis[18]. However, manually writing comprehensive reports requires careful lesion localization and characterization across multiple imaging sequences, and it is time-consuming. This underscores the critical need for models capable of generating reliable and interpretable neuroimaging reports [10].

While vision language models (VLMs) have advanced automated medical report generation [11,20], their reliable clinical deployment remains challenging because these

† Equal contribution.

models can still omit subtle abnormalities or hallucinate unsupported findings [24]. To improve lesion localization, some models incorporate lesion-specific expert models to provide lesion proposals [24]. However, these lesion proposals may contain false positives (FPs) and false negatives (FNs) due to imperfect segmentation, imaging artifacts, and lesion mimics [19,14]. Furthermore, in multi-lesion scenarios, different expert models may generate conflicting predictions, making lesion-level reconciliation necessary.

This limitation arises from the lack of an explicit verification mechanism in current VLMs, which differs from the iterative evidence assessment performed by radiologists. As illustrated in Fig. 1a, VLMs generate reports in a single pass without explicitly verifying lesion-level evidence, making them prone to missing subtle lesions or producing hallucinated findings [23]. In contrast, radiologists follow a detect-then-verify process. Specifically, they first identify candidate findings, followed by lesion verification across imaging sequences and final report generation (Fig. 1b) [10]. To bridge this gap, we present **B**rain **R**eporting with **A**djudication and **V**erification–**E**nhanced framework **(BRAVE)** (Fig. 1c), a unified detect-then-verify framework, to integrate lesion-specific expert models with a VLM for reliable brain MRI reporting. BRAVE first employs lesion-specific expert models to generate structured lesion candidates from multi-sequence MRI, where each candidate is associated with lesion evidence, including subtype, anatomical location, spatial relationships, volumetric characteristics, and lesion-aligned visual cues extracted from the global image representation. Building upon these evidence-enriched candidates, BRAVE performs lesion-wise CoT adjudication, where the VLM iteratively verifies, refines, and reconciles lesion findings by integrating local lesion evidence with global imaging context. The final adjudicated lesion set is then used to generate a clinically coherent report with fewer omissions and hallucinations.

We train BRAVE on a large-scale dataset of 150,106 brain MRI scans. Our main contributions are threefold. **First**, we propose a detect-then-verify architecture to integrate lesion-specific expert models with VLM reasoning, reducing hallucinations through structured lesion-wise CoT verification. **Second**, validated on 7,496 scans spanning 19 lesion subtypes, BRAVE achieves 86.7% lesion characterization accuracy and reduces FN and FP errors by 38.1%, maintaining high reliability in complex multi-lesion cases. **Third**, in a prospective clinical deployment, BRAVE improves workflow efficiency by reducing resident's drafting time by up to 15.4% and shortening attending's sign-off time by 17.4–32.8%, demonstrating its value in alleviating radiology workload.

2 Method

BRAVE unifies a set of lesion-specific expert models and a VLM into a coherent detect-then-verify framework. As shown in Fig. 1c, the workflow comprises two phases: 1) Lesion Proposal Generation, where lesion-specific expert models extract structured lesion candidates; 2) Lesion-wise CoT Adjudication, where the VLM iteratively verifies, refines, and reconciles these proposals against local and global visual context for final report generation.

2.1 Lesion Proposal Generation

To convert the MR image into lesion-level inputs that the VLM can verify, we employ lesion-specific expert models to generate voxel-wise segmentation masks for four lesion categories, including infarction, hemorrhage, tumor, and cerebral small vessel disease (CSVD), covering 19 lesion subtypes. We treat each connected component in the predicted masks as a candidate lesion ℓ, and denote the candidate set as $\mathcal{L}$. For each candidate lesion ℓ, we construct a structured proposal $\mathbf{d}_\ell$ containing three attributes: 1) anatomical localization (via IoU-based matching with brain region masks), 2) spatial relations (i.e., centroid distance and direction to brain region masks), and 3) volumetric size. For each candidate lesion ℓ, we also extract a lesion-level visual embedding $\mathbf{v}_\ell$ as lesion-aligned features pooled from the global image feature map $\mathbf{V}_{\text{img}}$, allowing the VLM to verify each proposal using localized image evidence rather than relying on global context alone. Together, the structured attributes and visual embeddings form the lesion proposals for subsequent reasoning: $\mathcal{D} = \{\mathbf{d}_\ell, \mathbf{v}_\ell\}_{\ell \in \mathcal{L}}$.

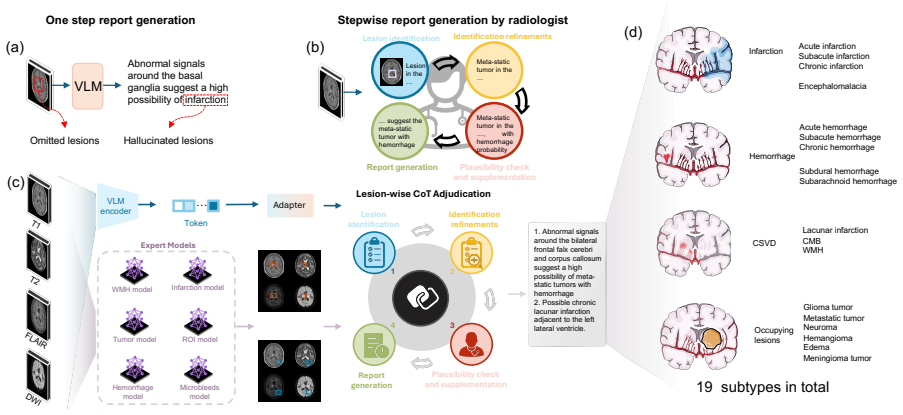

Fig. 1. Overview of BRAVE. (a) One-step VLM reporting with possible omissions and hallucinations. (b) Radiologist-style stepwise reasoning. (c) BRAVE with multi-sequence encoding, expert-guided lesion proposal generation, and lesion-wise CoT adjudication. (d) Lesion taxonomy evaluated in this work.

2.2 Lesion-wise CoT Adjudication

Unlike typical one-pass VLMs for reporting, BRAVE performs a lesion-wise CoT adjudication through explicit, lesion-level operations. Given global image tokens $\mathbf{V}_{\text{img}}$ and a set of structured lesion proposals $\mathcal{D} = \{\mathbf{d}_\ell, \mathbf{v}_\ell\}_{\ell \in \mathcal{L}}$, the model iteratively verifies and updates the lesion set via an action set:

$$u \in \{\texttt{KEEP}(\ell),\ \texttt{SUPPRESS}(\ell),\ \texttt{MERGE}(\ell_i, \ell_j),\ \texttt{RELABEL}(\ell, t),\ \texttt{ADD}(\ell)\}. \tag{1}$$

Here, SUPPRESS removes unsupported proposals to reduce FPs. MERGE consolidates duplicated or fragmented proposals. RELABEL corrects lesion subtype assignments when the structured lesion attributes $\mathbf{d}_\ell$ conflict with visual evidence. ADD supplements missed lesions only when consistent evidence is identified under the global imaging context. With these operations, the CoT proceeds in four stages, where the early stages focus on filtering obvious mismatches and retaining plausible candidates, and the later stages refine attributes, resolve conflicts, and supplement omissions.

Stage 1: Lesion identification. Given the candidate lesion set $\mathcal{D}$, the model performs an initial verification stage governed primarily by the lesion attributes $\mathbf{d}_\ell$. It applies SUPPRESS to eliminate candidates exhibiting anatomical implausibility. For instance, proposals whose volumetric size falls below a noise threshold, or whose anatomical localization conflicts with the global image context $\mathbf{V}_{\text{img}}$ (e.g., a lesion localized entirely within cerebrospinal fluid), are suppressed. Candidates with valid spatial and volumetric attributes are assigned a KEEP action for later stages.

Stage 2: Identification refinements. For each retained candidate, the model checks whether multiple proposals correspond to the same lesion by comparing their anatomical localization and spatial relations in $\mathcal{D}_\ell$, and applies MERGE to consolidate highly overlapping or adjacent proposals into a single lesion entity. Furthermore, it applies RELABEL when sequence-specific visual cues in $\mathbf{v}_\ell$ (e.g., distinct hyperintensity on DWI) contradict the initial subtype in $\mathcal{D}_\ell$.

Stage 3: Plausibility check and supplementation. In this stage, the model performs a global consistency check over the refined lesion set. It applies SUPPRESS to remove residual FPs that are inconsistent with the overall imaging context, such as candidates whose location or extent conflict with the global image features $\mathbf{V}_{\text{img}}$. To mitigate FNs, it re-examines high-risk regions with no expert-predicted lesion masks under $\mathbf{V}_{\text{img}}$. When convergent evidence across sequences suggests an omitted lesion (e.g., a missed lacunar infarction), the model applies ADD and infers its location based on visual evidence from $\mathbf{V}_{\text{img}}$, so that the supplemented finding remains grounded in image evidence rather than free-form generation.

Stage 4: Report generation. Finally, the model synthesizes a clinician-style radiology report from the fully adjudicated lesion set. It generates a coherent narrative that follows routine reading conventions, describing lesion type and key attributes in context and with appropriate uncertainty when needed.

Loss Functions We formulate the detect-then-verify paradigm as an end-to-end report generation task. The target sequence is formed by concatenating the intermediate CoT reasoning trajectory $\mathbf{Y}_{\text{CoT}}$ and the final diagnostic report $\mathbf{Y}_{\text{rep}}$. The model is jointly optimized using the standard autoregressive cross-entropy loss $\mathcal{L}_{\text{CE}}$:

$$\mathcal{L}_{\text{total}} = \lambda_{\text{CoT}}\mathcal{L}_{\text{CE}}\left(\mathbf{Y}_{\text{CoT}} \mid \mathbf{V}_{\text{img}}, \mathcal{D}\right) + \lambda_{\text{rep}}\mathcal{L}_{\text{CE}}\left(\mathbf{Y}_{\text{rep}} \mid \mathbf{V}_{\text{img}}, \mathcal{D}, \mathbf{Y}_{\text{CoT}}\right), \tag{2}$$

where λ_{CoT} and λ_{rep} are the balancing weights. The first term supervises the generation of stage-based verification actions, while the second optimizes the report generation.

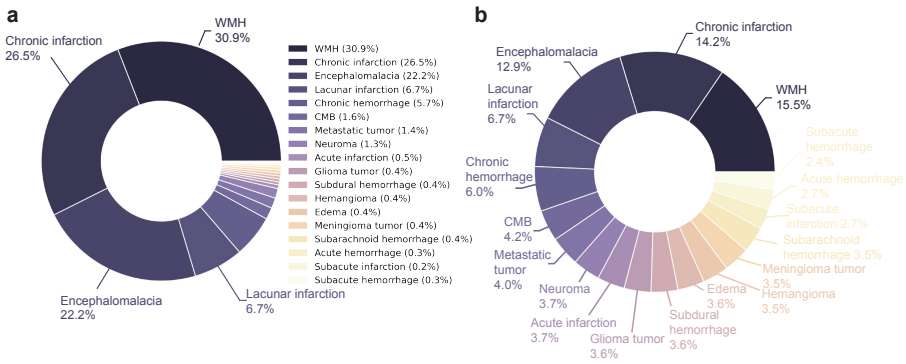

Fig. 2. Distributions of the training (a) and evaluation (b) datasets.

3 Datasets and Implementation Details

3.1 Datasets

We collect data from our in-house dataset and public datasets, including BraTS 2025 [15] (glioma, meningioma, metastatic tumor) and ISLES 2022 [6] (infarction). Fig. 2 summarizes the lesion subtype distributions of the training and evaluation sets, which contain 150,106 and 7,496 scans, respectively, covering 19 lesion subtypes. All brain MRI data undergo a standardized preprocessing pipeline. Raw DICOM scans are first converted to NIfTI using dcm2niix[13]. Each volume is then intensity-normalized to $[-1, 1]$ via percentile-based rescaling. All volumes are subsequently resampled to a fixed size of $40 \times 224 \times 224$ for model input. We use the Qwen-72B LLM [1] to refine the original reports by removing irrelevant content and to generate the stage-wise CoT supervision used in our framework. Given the ground-truth report and the corresponding expert-model lesion proposals, Qwen-72B produces a consistent multi-stage verification trajectory that decomposes the report into four corresponding CoT stages.

3.2 Implementation Details

We implement the model using the PyTorch [16] framework. For the global visual encoding, we employ a pre-trained CLIP encoder [17] based on our training dataset. The CLIP vision and text encoders are based on ViT [3] and BERT [2], respectively. The LLM backbone is initialized with Qwen-7B [1]. To obtain the structured proposals $\mathcal{D}$, the lesion-specific expert models are instantiated as independent 3D U-Net models[9,22], which are pre-trained on domain-specific datasets for the four predefined lesion categories. The visual encoder and expert models are kept frozen during VLM training. We only update the cross-modal linear projector and fine-tune the LLM backbone using Low-Rank Adaptation (LoRA) [8] with a rank of $r = 64$ and $\alpha = 128$. The training follows a two-stage curriculum. In the initial vision-language alignment stage, only the linear projector is trained for 1 epoch to map the visual tokens and lesion-level visual embeddings $\mathbf{v}_\ell$ into the LLM's text embedding space, using a learning rate of 1×10^{-3}. In

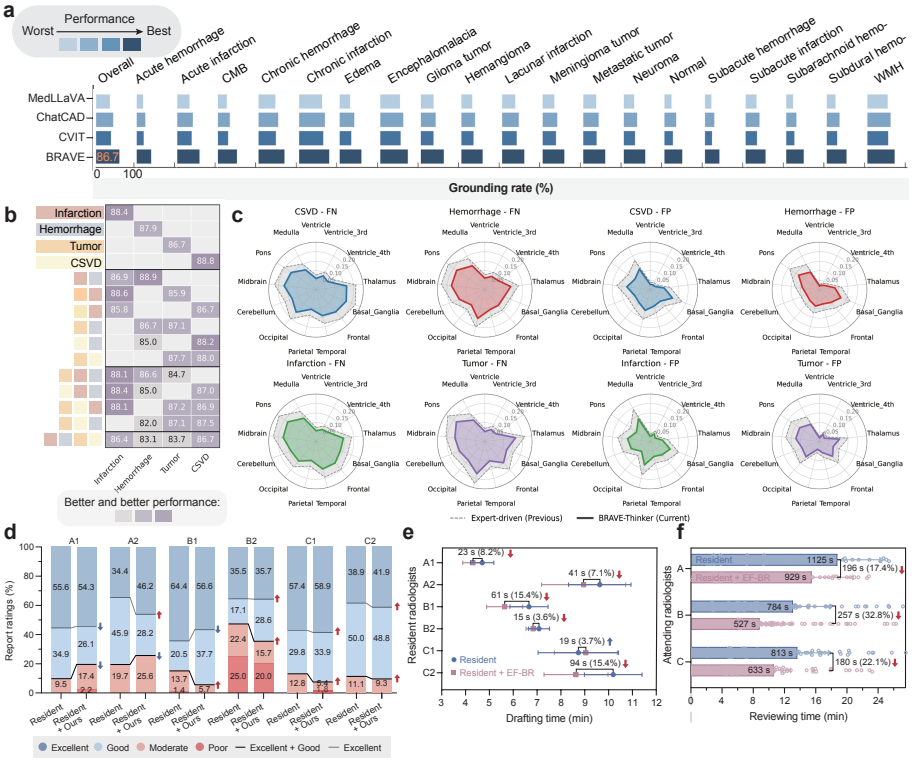

Fig. 3. Quantitative and clinical evaluation of BRAVE. (a) Lesion-specific reporting performance across different methods. (b) Performance under multi-lesion coexistence scenarios. (c) Correction of false-negative and false-positive errors. (d) Report quality assessment in prospective reader studies. (e) Resident drafting time reduction. (f) Attending radiologist review time reduction.

the subsequent CoT instruction-tuning stage, we jointly optimize the model using $\mathcal{L}_{\text{total}}$ with $\lambda_{\text{CoT}} = 0.5$ and $\lambda_{\text{rep}} = 1.0$. The model is fine-tuned for 3 epochs using the AdamW optimizer with a learning rate of 2×10^{-5}, weight decay of 0.01, and a cosine annealing learning rate scheduler with a 3% linear warmup. All experiments are conducted on NVIDIA A100 (80GB) GPUs utilizing `bfloat16` mixed-precision training.

4 Results

We evaluate BRAVE from three complementary aspects: 1) comparison with representative models across diverse lesion subtypes and multi-lesion coexistence patterns, 2) ablation on lesion-wise CoT adjudication to quantify FN and FP correction, and 3) prospective deployment to validate clinical faithfulness and workflow efficiency.

Comparison with representative baseline methods We evaluate report generation using lesion-specific metrics, where a prediction is counted as correct only if it matches

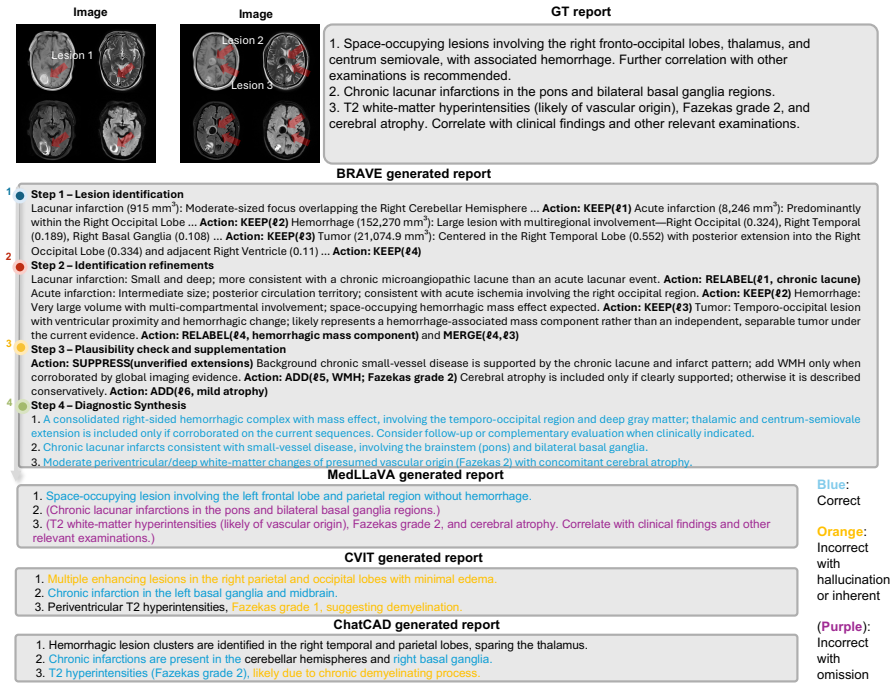

Fig. 4. Qualitative example of lesion-wise CoT adjudication and report generation. BRAVE progressively verifies and refines lesion proposals through four-stage CoT reasoning, reducing omissions and hallucinations compared with baseline VLMs.

both the anatomical location and the lesion subtype. As shown in Fig. 3a, BRAVE consistently outperforms representative baseline methods, including MedLLaVA [12], ChatCAD [21], and CVIT [11], across all lesion subtypes. Overall, BRAVE surpasses the strongest baseline (ChatCAD) by 8–12% on hemorrhage- and infarction-related lesions and achieves a mean lesion-specific accuracy of 86.7%. These results indicate improved lesion localization and characterization. Notably, the gains remain consistent on rare subtypes, including CMB and metastatic tumors, suggesting improved robustness under long-tailed distributions.

Since multiple abnormalities frequently co-occur in real clinical cases, we further assess performance under increasing diagnostic complexity (Fig. 3b). BRAVE maintains strong lesion-specific accuracy across diverse coexistence patterns, with an average accuracy of 87.1% under two-lesion coexistence and 86.6% under three-lesion coexistence, and remains reliable even in highly complex cases with four distinct coexisting lesion subtypes (85.0%). These results support the robustness of lesion-wise verification and cross-lesion integration enabled by our CoT.

Finally, we perform qualitative case studies to evaluate BRAVE. Fig.4 shows that one-pass VLMs often generate reports containing omissions or hallucinations. In contrast, BRAVE applies lesion-wise CoT adjudication to reduce such errors. Overall, these cases

Table 1. Ablation study of lesion-wise CoT adjudication and input evidence.

Block A: Backbone × lesion-wise CoT adjudication					**Block B: Evidence necessity**	
Backbone	**K/S**	**+M**	**+R**	**+A (Full)**	**Evidence**	**Acc. (%)**
Evidence fixed to $V_{\text{img}} + d_\ell + v_\ell$					Backbone fixed to Qwen2.5-7B-Instruct	
Qwen2.5-7B-Instruct	84.5	85.4	86.1	**86.7**	V_{img} only	76.8
Llama-3.1-8B-Instruct	83.6	84.6	85.2	85.7	$V_{\text{img}} + d_\ell$	80.9
GLM-4-9B-Chat	83.2	84.1	84.8	85.3	$V_{\text{img}} + v_\ell$	82.1
DeepSeek-R1-Distill-Qwen-7B	83.8	84.7	85.5	86.0	$V_{\text{img}} + d_\ell + v_\ell$	**86.7**

demonstrate that lesion-wise CoT adjudication yields more complete and internally consistent reports that better match radiologist descriptions.

Ablation study We first quantify the contribution of input evidence and action-level CoT adjudication in Table1. In Block A, lesion accuracy increases monotonically as we cumulatively enable KEEP/SUPPRESS (K/S), MERGE (+M), RELABEL (+R), and ADD (+A). For the best-performing backbone (Qwen2.5-7B-Instruct), accuracy improves from 84.5% (K/S) to 85.4% (+M), 86.1% (+R), and 86.7% with the full action set (+A). The progressive gains are attributable to each action: K/S suppresses anatomically implausible candidates to reduce FPs, MERGE consolidates duplicated detections, RELABEL fixes subtype mismatches using localized evidence, and ADD recovers missed lesions in high-risk regions. This monotonic improvement is consistent across Llama-3.1-8B [4], GLM-4-9B [7], and DeepSeek-R1-Distill-Qwen-7B[5], suggesting the benefit comes from the detect-then-verify design rather than a specific backbone. In Block B (evidence necessity), relying on global image tokens V_{img} alone yields the lowest lesion accuracy (76.8%), indicating that global context without explicit lesion-level evidence is insufficient for faithful localization and subtype characterization. Adding structured lesion attributes d_ℓ improves accuracy to 80.9%. Using lesion-aligned visual embeddings v_ℓ achieves a higher accuracy (82.1%), highlighting that localized appearance cues are particularly important for distinguishing subtle lesions. Combining d_ℓ and v_ℓ produces the strongest evidence-based performance, supporting their complementarity: d_ℓ constrains anatomical plausibility, while v_ℓ strengthens appearance-driven subtype discrimination.

We then localize the error patterns via a region-wise analysis over four major lesion categories (Fig. 3c) by removing lesion-wise CoT adjudication and directly generating reports from the same inputs. In this setting, FN errors concentrate on the anatomically complex deep structures, with higher rates in the thalamus, basal ganglia, and brainstem (8–15%) than in cortical regions (2–5%). FP errors show complementary, lesion-specific spatial biases: hemorrhage and tumor have higher FP rates near periventricular and extra-axial zones (8–12%), whereas infarction shows lower FP rates in frontal and parietal cortices (2–4%). These patterns indicate that single-pass reporting without verification is vulnerable to FN-prone high-risk regions and artifact-driven confounders in anatomically ambiguous areas. With lesion-wise CoT adjudication, BRAVE mitigates these errors, thereby reducing error propagation into the final report.

Prospective clinical workflow evaluation To validate clinical utility, we integrated BRAVE into the daily workflow of a radiology department for three weeks. The study involved six residents (drafters) and three attending radiologists (reviewers). As shown in Fig. 3d, AI assistance shifts report quality ratings toward higher quality categories, increasing the proportion of acceptable drafts (Excellent + Good) and reducing Moderate/Poor cases across resident groups. Also, this quality gain translates into measurable time savings: resident drafting time decreases in 5/6 groups by 3.6–15.4% (Fig. 3e), and attending review time is reduced for all three attendings by 17.4–32.8% (Fig. 3f). This confirms that BRAVE does not merely generate text, but produces review-ready drafts that alleviate the correction burden on attending radiologists.

5 Conclusion

In this work, we present BRAVE, a lesion-grounded VLM that couples lesion-specific expert proposals with lesion-wise CoT reasoning to generate clinically faithful brain MRI reports. BRAVE consistently improves lesion-level fidelity and performs robustly across diverse lesion subtypes, while preserving high accuracy under multi-lesion coexistence. Blinded reader assessments further support its reliability. Overall, BRAVE moves automated neuroimaging reporting beyond one-step text synthesis toward explicit lesion-grounded verification, providing a scalable foundation for trustworthy radiology report generation.

6 Acknowledgements

This work was supported in part by National Natural Science Foundation of China (grant numbers 82441023, U23A20295, 62131015, 82394432), the China Ministry of Science and Technology (STI2030-Major Projects-2022ZD0213100), The Key R&D Program of Guangdong Province, China (grant number 2023B0303040001), and HPC Platform of ShanghaiTech University.

References

1. Bai, J., Bai, S., Chu, Y., Cui, Z., Dang, K., Deng, X., Fan, Y., Ge, W., Han, Y., Huang, F., et al.: Qwen technical report. arXiv preprint arXiv:2309.16609 (2023)
2. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers). pp. 4171–4186 (2019)
3. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
4. Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., et al.: The llama 3 herd of models. arXiv preprint arXiv:2407.21783 (2024)

5. Guo, D., Yang, D., Zhang, H., Song, J., Wang, P., Zhu, Q., Xu, R., Zhang, R., Ma, S., Bi, X., et al.: Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. arXiv preprint arXiv:2501.12948 (2025)
6. Hernandez Petzsche, M.R., De La Rosa, E., Hanning, U., Wiest, R., Valenzuela, W., Reyes, M., Meyer, M., Liew, S.L., Kofler, F., Ezhov, I., et al.: Isles 2022: A multi-center magnetic resonance imaging stroke lesion segmentation dataset. Scientific data **9**(1), 762 (2022)
7. Hong, W., Yu, W., Gu, X., Wang, G., Gan, G., Tang, H., Cheng, J., Qi, J., Ji, J., Pan, L., et al.: Glm-4.5 v and glm-4.1 v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning. arXiv preprint arXiv:2507.01006 (2025)
8. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al.: Lora: Low-rank adaptation of large language models. Iclr **1**(2), 3 (2022)
9. Hua, R., Huo, Q., Gao, Y., Sui, H., Zhang, B., Sun, Y., Mo, Z., Shi, F.: Segmenting brain tumor using cascaded v-nets in multimodal mr images. Frontiers in computational neuroscience **14**, 9 (2020)
10. Leming, M.J., Bron, E.E., Bruffaerts, R., Ou, Y., Iglesias, J.E., Gollub, R.L., Im, H.: Challenges of implementing computer-aided diagnostic models for neuroimages in a clinical setting. NPJ Digital Medicine **6**(1), 129 (2023)
11. Li, C.Y., Chang, K.J., Yang, C.F., Wu, H.Y., Chen, W., Bansal, H., Chen, L., Yang, Y.P., Chen, Y.C., Chen, S.P., et al.: Towards a holistic framework for multimodal llm in 3d brain ct radiology report generation. Nature Communications **16**(1), 2258 (2025)
12. Li, C., Wong, C., Zhang, S., Usuyama, N., Liu, H., Yang, J., Naumann, T., Poon, H., Gao, J.: Llava-med: Training a large language-and-vision assistant for biomedicine in one day. Advances in Neural Information Processing Systems **36**, 28541–28564 (2023)
13. Li, X., Morgan, P.S., Ashburner, J., Smith, J., Rorden, C.: The first step for neuroimaging data analysis: Dicom to nifti conversion. Journal of neuroscience methods **264**, 47–56 (2016)
14. Liévin, V., Hother, C.E., Motzfeldt, A.G., Winther, O.: Can large language models reason about medical questions? Patterns **5**(3) (2024)
15. Menze, B.H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., Burren, Y., Porz, N., Slotboom, J., Wiest, R., et al.: The multimodal brain tumor image segmentation benchmark (brats). IEEE transactions on medical imaging **34**(10), 1993–2024 (2014)
16. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: Pytorch: An imperative style, high-performance deep learning library. Advances in neural information processing systems **32** (2019)
17. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)
18. Schwartz, L.H., Panicek, D.M., Berk, A.R., Li, Y., Hricak, H.: Improving communication of diagnostic radiology findings through structured reporting. Radiology **260**(1), 174–181 (2011)
19. Singhal, K., Azizi, S., Tu, T., Mahdavi, S.S., Wei, J., Chung, H.W., Scales, N., Tanwani, A., Cole-Lewis, H., Pfohl, S., et al.: Large language models encode clinical knowledge. Nature **620**(7972), 172–180 (2023)
20. Tu, T., Azizi, S., Driess, D., Schaekermann, M., Amin, M., Chang, P.C., Carroll, A., Lau, C., Tanno, R., Ktena, I., et al.: Towards generalist biomedical ai. arXiv preprint arXiv:2307.14334 (2023)
21. Wang, S., Zhao, Z., Ouyang, X., Liu, T., Wang, Q., Shen, D.: Interactive computer-aided diagnosis on medical image using large language models. Communications Engineering **3**(1), 133 (2024)
22. Zhang, Z., Ding, Z., Chen, F., Hua, R., Wu, J., Shen, Z., Shi, F., Xu, X.: Quantitative analysis of multimodal mri markers and clinical risk factors for cerebral small vessel disease based on deep learning. International Journal of General Medicine pp. 739–750 (2024)

23. Zhao, Q., Lu, Y., Kim, M.J., Fu, Z., Zhang, Z., Wu, Y., Li, Z., Ma, Q., Han, S., Finn, C., et al.: Cot-vla: Visual chain-of-thought reasoning for vision-language-action models. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 1702–1713 (2025)
24. Zhao, Z., Wang, S., Gu, J., Zhu, Y., Mei, L., Zhuang, Z., Cui, Z., Wang, Q., Shen, D.: Chatcad+: Towards a universal and reliable interactive cad using llms. IEEE Transactions on Medical Imaging (2024)