Unleashing Video Language Models for Fine-grained HRCT Report Generation

Yingying Fang[1,2*], Huichi Zhou[1*], KinHei Lee[1*], Yijia Wang[1*], Zhenxuan Zhang[1], Jiahao Huang[1], and Guang Yang[1,2,3,4]

[1] Bioengineering Department and Imperial-X, Imperial College London, London, UK
[2] National Heart and Lung Institute, Imperial College London, London, UK
[3] Cardiovascular Research Centre, Royal Brompton Hospital, London, UK
{y.fang, g.yang}@imperial.ac.uk

Abstract. Generating precise diagnostic reports from high-resolution computed tomography is critical for clinical workflows, yet remains challenging due to the high pathological diversity and spatial sparsity within 3D volumes. While video language models (VideoLMs) have demonstrated remarkable spatio-temporal reasoning in general domains, their adaptability to domain-specific, high-volume medical interpretation remains underexplored. In this work, we present AbSteering, an abnormality-centric framework that adapts general-domain VideoLMs to precise report generation. Specifically, AbSteering introduces: (i) an abnormality-centric Chain-of-Thought scheme that encourages finding-first abnormality reasoning, and (ii) a Direct Preference Optimization objective that uses clinically confusable abnormalities as hard negatives to enhance fine-grained discrimination. Our results demonstrate that general-purpose VideoLMs exhibit strong transferability to high-volume medical imaging when guided by AbSteering. Notably, VideoLMs outperform state-of-the-art domain-specific CT foundation models, achieving superior detection sensitivity while mitigating hallucinated findings. Our data and model weights are released at https://github.com/ayanglab/CTRG_VideoLMs.

1 Introduction

High-Resolution Computed Tomography (HRCT) is a key modality for the definitive diagnosis and longitudinal monitoring of thoracic and cardiopulmonary diseases, providing volumetric scans with rich texture and fine anatomical detail. In parallel, AI-driven medical report generation has become an active research direction, as it can reduce clinical workload, standardize diagnostic narratives, and mitigate inter-observer variability. While report generation for 2D chest X-rays has progressed rapidly, extending these advances to 3D HRCT remains substantially more challenging. Compared with 2D images, HRCT introduces both (i) prohibitive computational and memory costs due to hundreds of slices per study, and (ii) a more difficult visual understanding problem, where clinically critical

* Equal contribution.

abnormalities are spatially localized and diverse, often being overwhelmed by dominant normal anatomical patterns.

Early attempts at CT report generation built on X-ray paradigms by compressing CT volumes into lower-dimensional representations and then reusing X-ray–oriented report generators [9,15,4]. With the emergence of large language models (LLMs), subsequent work (e.g., Dia-LLaMA [7]) designed CT-specific visual encoders and connected them to LLM decoders to improve language quality and clinical style. More recently, modality-specific radiology foundation models have been explored to better handle 3D medical inputs and support report generation, such as RadFM [18], CT-CHAT [10] and M3D [2]. Building on these backbones, a line of work further focuses on strengthening CT feature extraction and representation learning based on the foundation models [5,11,8].

Despite steady progress, most existing approaches have primarily focused on developing CT-tailored encoders, which can be data- and compute-intensive. However, their ability to comprehensively capture abnormalities distributed across the whole volume, which is critical for report correctness, remains less systematically explored. In a different vein, recent advances in **video-language models (VideoLMs)** have significantly improved their ability to reason over long visual sequences by modeling temporal and spatial dependencies [12,17,19,16,23]. These VideoLMs are trained to align dynamic visual content with language, enabling strong performance on captioning and multi-step video reasoning. Importantly, an HRCT volume can be naturally viewed as a video-like slice sequence, which makes VideoLMs a promising candidate for volumetric report generation. However, despite their strong performance, VideoLMs are typically pretrained on natural video data and therefore lack domain-specific knowledge—especially for medical 3D imaging where recognizing diverse abnormalities requires specialized clinical understanding. This mismatch raises three key questions: (1) *Can the encoders in VideoLMs capture clinically relevant 3D features and support accurate CT report generation?* (2) *How can we adapt general-domain VideoLMs to domain-specific medical reporting in a highly efficient manner?* (3) *How does such transfer compare with modality-specific CT foundation models in terms of report accuracy and clinical fidelity?*

To address these questions, we investigate the transferability of pretrained VideoLMs to high-volume radiology interpretation and propose ***AbSteering***, an abnormality-centric framework that steers VideoLMs toward precise radiology interpretation by structured reasoning pathways and discriminative preference optimization. Concretely, our framework introduces: **(i) an abnormality-centric Chain-of-Thought (CoT) training scheme** built on curated reports with structured templates. This explicitly prompts clinically critical reasoning before final report generation, constraining the model to learn diverse disease categories while suppressing hallucinated or normal-tissue-dominated descriptions; **(ii) a fine-grained abnormality discrimination objective based on Direct Preference Optimization (DPO)**, which introduces clinically confusable abnormalities within the same anatomical region as hard negatives to better resolve subtle inter-abnormality differences and reduce hallucinations.

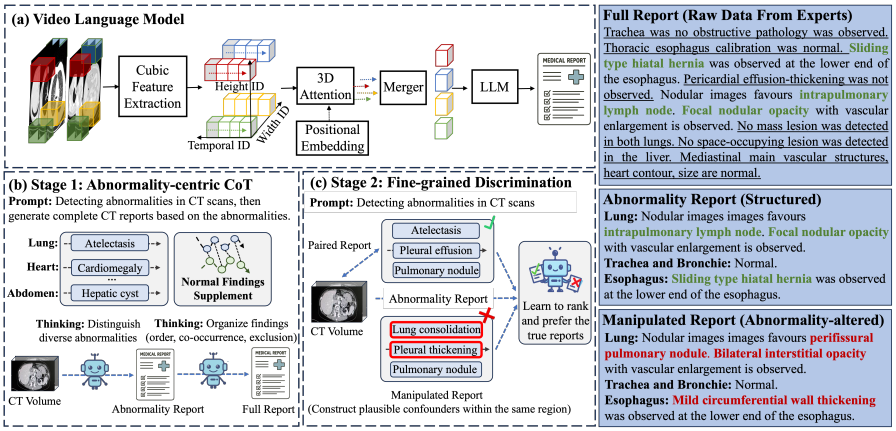

Fig. 1: Left: overall framework for adapting mainstream VideoLMs with the proposed AbSteering in two stages: (a) the typical backbone of current VideoLMs; (b) Stage 1, abnormality-centric CoT training; and (c) Stage 2, fine-grained abnormality discrimination via DPO. **Right:** examples of abnormality-centric and manipulated report samples.

Our contributions are three-fold: **(i) Cross-modal transferability.** We systematically study the transferability of general-domain VideoLMs and show that, under limited data, they transfer effectively to 3D medical imaging, providing an efficient alternative to training modality-specific foundation models from scratch; **(ii) Method.** We propose an efficient abnormality-centric language-steering framework that successfully adapts VideoLMs and achieves state-of-the-art performance on fine-grained clinical efficacy metrics; and **(iii) Dataset.** We curate ***CT-RATE-AB***, a dataset designed to facilitate CoT training in medical imaging and to evaluate diversity and fine-grained abnormality recognition for high-volume chest CT understanding.

2 Method

In this section, we first describe how volumetric CT scans are represented as long visual sequences and decoded into radiology reports, and then introduce AbSteering, a two-stage training strategy consisting of abnormality-centric CoT training and DPO-based fine-grained abnormality preference optimization.

2.1 Architecture of Video–Language Models

Most mainstream VideoLMs follow a unified visual–language pipeline for processing long visual sequences. Given an input video or volumetric image $X \in \mathbb{R}^{T\times H\times W\times C}$, where T denotes the temporal or axial dimension and (H, W) denote the spatial dimensions, the visual backbone first partitions the input into non-overlapping volumetric patches of size (p_t, p_h, p_w), which are flattened and

projected into visual token embeddings through a volumetric patch embedding layer: $Z^{(0)} = E_{\text{patch}}(X)$, where $Z^{(0)} \in \mathbb{R}^{N \times d_v}$, $N = \frac{T}{p_t}\frac{H}{p_h}\frac{W}{p_w}$ denotes the number of visual tokens after patchification, and d_v is the visual embedding dimension. The visual tokens are then processed by stacked Transformer blocks:

$$Z^{(l+1)} = \mathcal{T}^{(l)}\left(Z^{(l)} + E_{\text{pos}}\right),$$

where $\mathcal{T}^{(l)}(\cdot)$ denotes the l-th Transformer block and E_{pos} represents positional information over the temporal/axial, height, and width dimensions. To reduce the large number of volumetric tokens before language modeling, a projector or resampler $\mathcal{P} : \mathbb{R}^{N \times d_v} \to \mathbb{R}^{M \times d_t}$ compresses the dense visual representation into a smaller set of language-aligned tokens:

$$V = \mathcal{P}\left(Z^{(L)}\right),$$

where $M \ll N$ is the reduced token number and d_t denotes the embedding dimension of the language model. Finally, the compressed visual tokens are provided to an autoregressive language decoder:

$$p(y_t \mid y_{<t}, V) = \mathcal{D}(y_{<t}, V),$$

where $\mathcal{D}(\cdot)$ denotes the large language model and y_t is the generated report token at decoding step t. From an architectural perspective, current VideoLMs and modality-specific CT foundation models are largely similar. Both typically follow the same high-level pipeline: volumetric or spatiotemporal tokenization, positional encoding, 3D attention-based visual encoding, token merging, and language decoding for text generation. The main difference therefore does not lie in the backbone structure itself, but in the training domain and supervision.

2.2 Abnormality-centric chain-of-thought training

To enable abnormality-centric reasoning, we normalize ***CT-RATE*** reports into a unified `(region: abnormality)` template. Using this structured-to-raw pairing, we implement a CoT process that decouples the task: the model first distills diagnostic findings as a reasoning anchor before synthesizing the final report.

Construction of $R_{\mathbf{AB}}$. To reorganize raw reports into an abnormality-preserving format, we adopted the anatomical taxonomy defined in [21], which divides chest CT reports into ten regions: Lung, Trachea and Bronchi, Mediastinum, Heart, Esophagus, Pleura, Bone, Thyroid, Breast, and Abdomen. GPT-4o [1] was used to assign each abnormality-related sentence to the corresponding anatomical region. Different from [21], we only retained sentences describing abnormalities and introduced an additional "Others" category to include abnormal findings that do not fall into the ten predefined regions. To ensure the completeness of the reorganized reports, we further introduced two auxiliary audit categories, "uncategorized" and "repetitive", to identify abnormality sentences that were either not assigned to any region or assigned to multiple regions. These

cases were manually reviewed and corrected by assigning missing sentences to the appropriate region or removing duplicated sentences from redundant regions, resulting in the curated ***CT-RATE-AB*** dataset, as in Fig. 1.

CoT training. Given ***CT-RATE-AB***, each training scan contains a curated abnormality-centric summary R_{AB} and the corresponding complete radiology report R_{Full}. The abnormality-centric summary R_{AB} includes only clinically relevant abnormal findings, whereas R_{Full} retains the complete reporting content, including normal descriptions and standard report structures. To encourage the model to identify abnormalities before generating the full report, we serialize the target response into a structured CoT-style sequence:

$$A_{\text{CoT}} = [\texttt{<AB>}, R_{\text{AB}}, \texttt{<REPORT>}, R_{\text{Full}}],$$

where `<AB>` marks the beginning of the abnormality-focused intermediate stage and `<REPORT>` marks the transition to the final report-generation stage.

2.3 Fine-grained abnormality discrimination

CT abnormalities often exhibit subtle and visually confusable patterns, making fine-grained discrimination highly dependent on domain-specific clinical knowledge. To improve the model's ability to distinguish these subtle distinctions, we introduce a fine-grained preference learning strategy in Stage 2 that constructs fake reports by replacing true abnormalities with clinically confusable alternatives. We then apply DPO to encourage the model to assign higher preference to the clinically correct report than to the corrupted counterpart. By learning which report is more faithful, the model is encouraged to attend to subtle visual cues that determine report quality, thereby improving both fine-grained abnormality discrimination and hallucination suppression.

Construction of $R_{\text{AB\_Fake}}$. We construct $R_{\text{AB\_Fake}}$ automatically from R_{AB} using GPT-4o. For each abnormality-centric report, GPT-4o is instructed to replace the target abnormality with an anatomically matched but clinically distinct abnormality that could plausibly occur in the same region, while keeping the original region label, sentence template, and all other positional information unchanged. This yields hard negative reports that preserve fluency and structural consistency but alter the clinically critical abnormality semantics, thereby providing an effective contrastive signal for preference optimization.

Direct Preference Optimization. We further refine the Stage 1 model as the reference model π_{ref}, and optimize the target model π_θ using DPO. For each CT volume X, we use the DPO prompt p_{DPO}: "Identify the clinically meaningful abnormalities in the chest CT volume". We construct a preference pair (y_w, y_l), where $y_w = R_{\text{AB}}$ is the curated abnormality summary and $y_l = R_{\text{AB-Fake}}$ is the corrupted counterpart. We minimize the following DPO objective:

$$\mathcal{L}_{\text{DPO}}(\pi_\theta; \pi_{\text{ref}}) = -\log \sigma \left[\beta \log \frac{\pi_\theta(y_w \mid X, p_{\text{DPO}})}{\pi_{\text{ref}}(y_w \mid X, p_{\text{DPO}})} - \beta \log \frac{\pi_\theta(y_l \mid X, p_{\text{DPO}})}{\pi_{\text{ref}}(y_l \mid X, p_{\text{DPO}})} \right],$$

where $\sigma(\cdot)$ is the logistic function and β controls the deviation from the reference.

Table 1: Benchmark of different methods across various evaluation metrics on the CT-RATE benchmark. Rows with the suffix "AbSteer" denote models fine-tuned with the proposed AbSteering strategy using ***CT-RATE-AB***. **Bold*** indicates the best, **bold** indicates the second best, and underline indicates the third best.

	NLG Metrics				CE Micro			CE Macro			GEMA Abnormality		
Method	**BL-4**	**MT-R**	**RG-L**	**BERT**	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
CT2Rep [9]	**28.04***	**45.62***	**45.43***	**88.10***	26.39	10.50	14.10	17.03	12.59	10.65	10.65	16.03	7.60
RadFM [18]	17.02	36.81	30.46	86.17	36.10	13.48	19.63	28.51	10.16	13.05	8.19	14.76	6.92
Reg2RG [6]	21.08	38.25	24.41	86.18	28.47	11.06	15.93	19.76	8.17	10.48	7.37	13.27	5.58
CTCHAT [10]	17.63	37.62	32.50	86.35	25.13	37.48	30.08	19.97	29.68	21.66	9.12	19.32	7.98
M3D-8B [2]	22.98	40.75	37.76	87.52	47.60	28.54	35.69	41.62	22.41	26.74	16.31	16.49	11.93
Qwen2.5-VL-7B [16]	21.25	40.73	36.71	87.30	48.06	25.88	33.64	39.85	20.06	25.57	15.82	22.33	15.79
InternVL3-8B [23]	22.05	41.82	38.49	87.40	**53.57**	37.99	44.45	**48.28**	33.71	**38.91**	**24.05**	23.47	**20.01**
M3D-AbSteer	23.09	41.26	38.58	**87.83**	44.95	41.66	43.24	39.94	34.75	36.18	18.98	20.73	16.28
Qwen2.5-VL-AbSteer	21.40	41.61	37.99	87.13	49.15	**43.22**	**45.99**	43.13	**35.78**	37.90	18.16	**25.32**	19.28
InternVL3-AbSteer	**23.58**	**44.13**	**40.49**	87.59	**57.88***	**51.58***	**54.55***	**54.08***	**45.01***	**47.66***	**24.26***	**32.25***	**22.29***

3 Experimental details

Dataset. We use the ***CT-RATE*** dataset [9], which contains paired non-contrast chest CT volumes and radiology reports collected from Istanbul Medipol University. The dataset includes 25,692 CT studies from 21,304 unique patients, resulting in 50,188 CT volumes after applying multiple reconstruction settings. Each volume is accompanied by the corresponding radiology report and metadata. We follow the data split and preprocessing protocol in [9]. Specifically, each HRCT volume is resampled into 240 slices of size 480×480, with a uniform spacing of 0.75 mm along the x- and y-axes and 1.5 mm along the z-axis, and clipped within a Hounsfield Unit (HU) window of $[-1000, 200]$. The dataset is split into a training set of 46,717 CT scans from 20,000 patients and a validation set of 3,039 CT scans from 1,314 patients. To accommodate VideoLMs, we further convert each CT scan into MP4 format at a frame rate of 18 fps.

Experimental Setting. We benchmark our method against existing CT-specific baselines with open-source training code, including the non-LLM report generator CT2Rep [9] and four CT foundation model baselines: RadFM [18], M3D [2], CT-CHAT [10], and Reg2RG [6]. We further include two representative general-purpose VideoLMs, Qwen2.5-VL and InternVL3, to evaluate the transferability of video-pretrained vision–language models to HRCT report generation. For a fair comparison, all baseline models are initialized from their publicly released pretrained weights and fine-tuned on the ***CT-RATE*** dataset for the report-generation task. To evaluate AbSteering, we apply the proposed two-stage strategy to M3D, Qwen2.5-VL, and InternVL3. The first stage performs supervised fine-tuning on the abnormality CoT data, while the second stage applies DPO-based abnormality preference optimization. For Qwen2.5-VL and InternVL3, we freeze the visual encoder and update only the multimodal projector and the LoRA-adapted LLM to preserve their pretrained visual representations. All models are trained on two 80GB NVIDIA A100 GPUs with a total batch size of 4.

Evaluation Metrics. To comprehensively compare the generated reports, we employ both natural language generation (NLG) metrics and clinical-efficacy

Full Report from Experts: The thyroid gland is enlarged with a heterogeneous appearance. A subcutaneous pacemaker and electrodes are present on the left anterior chest wall. A 13x9 mm oval calcified lesion is seen at the retroareolar level of the left breast. No occlusive pathology was observed in the trachea and lumen of both main bronchi. The ascending aorta is dilated and the descending aorta is mildly enlarged. Heart size is at the upper limit of normal. Calcified atheroma plaques are present in the thoracic aorta and coronary arteries. Pericardial effusion-thickening was not observed. Thoracic esophagus calibration was normal. No pathologically enlarged lymph nodes were detected. Minimal peribronchial thickening is seen bilaterally. A mosaic attenuation pattern is present in both lungs. Minimal bronchiectatic changes with fibrotic recessions are observed in the posterior segment of the right upper lobe. Passive atelectasis is present in the posterobasal and laterobasal segments of both lower lobes. No mass lesion or active infiltration was detected. Liver contours are lobulated with caudate lobe hypertrophy, consistent with chronic parenchymal disease. A millimetric gallstone is noted. Nodular thickening is present in the left adrenal gland. Right-facing rotoscoliosis and osteodegenerative changes are present at the thoracic level.

M3D [M:1, O:0]: Pericardial effusion-thickening was not observed. Thoracic esophagus calibration was normal. There are mild atelectatic changes in the lower lobes of both lungs. Pleural effusion-thickening was not detected.

RadFM [M:0, O:6]: Pericardial-pleural effusion was not detected. Areas of increase in density consistent with consolidation are observed. Viral pneumonias are considered. There are sequela parenchymal changes. There are minimal emphysematous changes in both lungs. There is a 3 mm diameter hyperdense stone. No lymph node was observed in intraabdominal pathological. There is a compression fracture in the T12, L1 and L2 vertebra corpus superior end plate, which causes a height loss of less than 50%.

M3D Steered [M:3, O:1]: Calcified atheroma plaques were observed in the thoracic aorta and coronary arteries. Pericardial effusion-thickening was not observed. There are passive atelectatic changes in both lung lower lobe posterobasal segments. There are mild emphysematous changes in both lungs. No mass or infiltrative lesion was detected. Osteodegenerative changes were observed in the vertebral bone structures within the sections.

Qwen-VL Steered [M:4, O:1]: The thyroid gland is larger than normal and has a heterogeneous nodular appearance. Pericardial effusion-thickening was not observed. Calcific atheroma plaques are observed in the aorta and coronary arteries. Thoracic esophagus calibration was normal. Subpleural atelectasis and sequela fibrotic changes are observed in both lungs, more prominently in the left lower lobe. There are millimetric nonspecific nodules in both lungs. Degenerative changes are observed in the vertebrae.

Intern_VL Steered [M:5, O:1]: There is a pacemaker on the anterior chest wall on the left. Pericardial effusion-thickening was not observed. Calcific plaques are present in the aorta and coronary arteries. No enlarged lymph nodes in prevascular. Linear atelectasis and mosaic density differences are observed. There are minimal atelectasis and mosaic density differences in the lower lobe of the left lung. Peribronchial ground glass densities are present in the posterior lower lobe of the left lung. There are degenerative changes in the vertebrae.

Fig. 2: Case-study comparison of generated reports. Green text indicates predicted abnormalities that are present in the ground-truth report shown in blue text, whereas red text indicates generated findings that should not have been included. $\boldsymbol{M}$ and $\boldsymbol{O}$ denote the total numbers of matched and overpredicted findings, respectively.

(CE) metrics. For NLG evaluation, we report BLEU (BL) [14], METEOR (MT-R) [3], ROUGE-L (RG-L)[13] and BERTScore (BERT)[20]. For CE evaluation, we use a pretrained RadBERT classifier [9] to classify 18 abnormalities from reports. We evaluate abnormality recognition using micro-, macro and sample-averaged precision, recall, and F1 scores across the 18 abnormalities. In addition, we report the LLM-based GEMA score [22] to measure case-level open-vocabulary abnormality accuracy beyond the predefined 18 classes.

4 Experimental results

Numerical results. *VideoLMs exhibit significant potential for high-volume medical report generation, achieving performance comparable to modality-specific foundation models.* As illustrated in Table 1, among the CT foundation model baselines, M3D-8B achieves the leading performance when fine-tuned on the CT-RATE dataset. Notably, the general-domain Qwen2.5-VL-7B achieves compara-

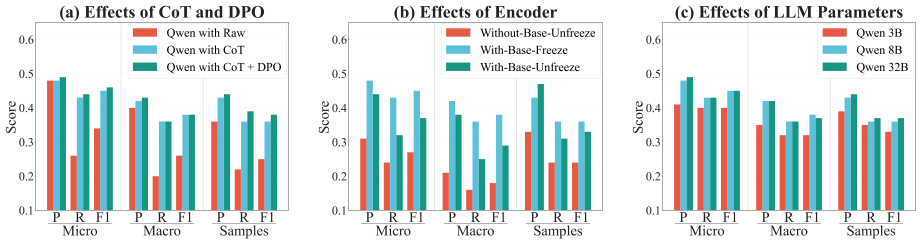

Fig. 3: Ablation studies of (a) the AbSteering strategy, (b) video encoder pretraining with different initialization and tuning settings, and (c) LLM parameters.

ble results to M3D-8B, with only a marginal gap in recall while InternVL3-8B demonstrates even stronger baseline capabilities, occasionally surpassing M3D-8B across CE metrics. It is also worth noting that relying solely on NLG metrics can be unreliable when generated reports are dominated by abundant normal findings. *Steering unlocks strong medical report generation ability in general VideoLMs.* As shown in the "AbSteered" section of Table 1, our steering method improves both M3D and the two general VideoLM backbones. The gains are particularly pronounced for Qwen2.5-VL and InternVL3, which surpass the strongest modality-specific report generators by a clear margin. This result not only validates the effectiveness of our abnormality-centric steering framework, but also suggests that the volumetric medical reasoning capability of large-scale VideoLMs can be effectively unlocked through domain-specific guidance.

Case Study. A case study is presented in Fig. 2 to illustrate the precision–recall trade-offs across different models in fine-grained abnormality detection. As observed, all models benefit significantly from the proposed steering framework. Notably, VideoLMs achieve the highest recall among all tested models without increasing hallucinations.

Ablation Study. *(CoT and DPO)* Fig. 3 (a) shows that abnormality-centric CoT significantly boosts recall. Adding fine-grained DPO enhances both precision and recall, effectively suppressing hallucinations while increasing sensitivity—a synergy often elusive in conventional methods. *(Effect of the encoder)* Using Qwen with Stage 1 CoT as a baseline, we compared: (1) training from scratch, (2) a frozen pre-trained encoder for Qwen2.5-VL, and (3) Low-Rank Adaptation (LoRA) fine-tuning ($rank = 8$). Fig. 3 (b) confirms that discarding pretrained weights causes a sharp performance drop, indicating that general-domain video pretraining provides an important visual foundation. Unexpectedly, LoRA yields no gains over the frozen encoder, suggesting the pre-trained features from the VideoLMs are sufficiently robust to generalize to medical textures without further adaptation. *(LLM Scaling)* We explored 3B, 7B, and 32B LLM scales. Performance increases substantially from 3B to 7B, but shows only marginal gains at 32B (Fig. 3 (c)), suggesting that the current bottleneck lies more in visual-textual alignment and abnormality supervision than in LLM capacity.

5 Conclusion

HRCT plays a critical role in diagnosing and monitoring thoracic and cardiopulmonary diseases, yet automated report generation remains challenging due to the need for comprehensive abnormality recognition across high-volume CT scans. In this work, we systematically investigate the transferability of general-domain vision–language models pretrained on videos to HRCT report generation and propose an abnormality-steering framework that adapts pretrained VideoLMs to medical reporting by encouraging finding-first generation and fine-grained abnormality discrimination. Our results show that, with appropriate domain-specific guidance, large-scale VideoLMs can serve as effective backbones for volumetric CT report generation and improve the capture of diverse clinically relevant abnormalities. Future work will extend the evaluation to external datasets and incorporate more clinically oriented metrics to further assess cross-dataset generalization and abnormality-specific performance.

Acknowledgements Guang Yang was supported in part by the ERC IMI (101005122), H2020 (952172), the MRC (MC/PC/21013), the Royal Society (IEC/NSFC/211235), the NVIDIA Academic Hardware Grant Program, the SABER project supported by Boehringer Ingelheim Ltd, the NIHR Imperial Biomedical Research Centre (RDA01), the Wellcome Leap Dynamic Resilience program co-funded by Temasek Trust, UKRI guarantee funding for Horizon Europe MSCA Postdoctoral Fellowships (EP/Z002206/1), the UKRI MRC Research Grant, TFS Research Grants (MR/U506710/1), the Swiss National Science Foundation (Grant No. 220785), and the UKRI Future Leaders Fellowship (MR/V023799/1, UKRI2738).

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

References

1. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023)
2. 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)
3. 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)
4. Cao, W., Zhang, J., Xia, Y., Mok, T.C., Li, Z., Ye, X., Lu, L., Zheng, J., Tang, Y., Zhang, L.: Bootstrapping chest ct image understanding by distilling knowledge from x-ray expert models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11238–11247 (2024)

5. Chen, H., Zhao, W., Li, Y., Zhong, T., Wang, Y., Shang, Y., Guo, L., Han, J., Liu, T., Liu, J., et al.: 3d-ct-gpt: Generating 3d radiology reports through integration of large vision-language models. arXiv preprint arXiv:2409.19330 (2024)
6. Chen, Z., Bie, Y., Jin, H., Chen, H.: Large language model with region-guided referring and grounding for ct report generation. IEEE Transactions on Medical Imaging (2025)
7. Chen, Z., Luo, L., Bie, Y., Chen, H.: Dia-llama: Towards large language model-driven ct report generation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 141–151. Springer (2025)
8. Deng, X., He, X., Bao, J., Zhou, Y., Cai, S., Cai, C., Chen, Z.: MvKeTR: chest ct report generation with multi-view perception and knowledge enhancement. IEEE Journal of Biomedical and Health Informatics (2025)
9. Hamamci, I.E., Er, S., Menze, B.: CT2Rep: Automated radiology report generation for 3d medical imaging. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 476–486. Springer (2024)
10. Hamamci, I.E., Er, S., Wang, C., Almas, F., Simsek, A.G., Esirgun, S.N., Dogan, I., Durugol, O.F., Hou, B., Shit, S., et al.: Generalist foundation models from a multimodal dataset for 3D computed tomography. Nature Biomedical Engineering pp. 1–19 (2026)
11. Li, S., Qin, P., Wu, H., Nie, D., Thirunavukarasu, A.J., Yu, J., Zhang, L.: μ^2 tokenizer: Differentiable multi-scale multi-modal tokenizer for radiology report generation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 3–12. Springer (2025)
12. Lin, B., Ye, Y., Zhu, B., Cui, J., Ning, M., Jin, P., Yuan, L.: Video-llava: Learning united visual representation by alignment before projection. arXiv preprint arXiv:2311.10122 (2023)
13. Lin, C.Y.: Rouge: A package for automatic evaluation of summaries. In: Text summarization branches out. pp. 74–81 (2004)
14. 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)
15. Tang, Y., Yang, H., Zhang, L., Yuan, Y.: Work like a doctor: Unifying scan localizer and dynamic generator for automated computed tomography report generation. Expert Systems with Applications **237**, 121442 (2024)
16. Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W., et al.: Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. arXiv preprint arXiv:2409.12191 (2024)
17. Wang, X., Zhang, X., Luo, Z., Sun, Q., Cui, Y., Wang, J., Zhang, F., Wang, Y., Li, Z., Yu, Q., et al.: Emu3: Next-token prediction is all you need. arXiv preprint arXiv:2409.18869 (2024)
18. Wu, C., Zhang, X., Zhang, Y., Hui, H., Wang, Y., Xie, W.: Towards generalist foundation model for radiology by leveraging web-scale 2D&3D medical data. Nature Communications **16**(1), 7866 (2025)
19. Zhang, J., Huang, J., Jin, S., Lu, S.: Vision-language models for vision tasks: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence (2024)
20. Zhang, T., Kishore*, V., Wu*, F., Weinberger, K.Q., Artzi, Y.: Bertscore: Evaluating text generation with bert. In: International Conference on Learning Representations (2020)
21. Zhang, X., Wu, C., Zhao, Z., Lei, J., Tian, W., Zhang, Y., Xie, W., Wang, Y.: Development of a large-scale grounded vision language dataset for chest ct analysis. Scientific Data **12**(1), 1636 (2025)

22. Zhang, Z., Lee, K., Jing, P., Deng, W., Zhou, H., Jin, Z., Huang, J., Gao, Z., Marshall, D.C., Fang, Y., et al.: Gema-score: granular explainable multi-agent scoring framework for radiology report evaluation. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 40, pp. 13025–13033 (2026)
23. Zhu, J., Wang, W., Chen, Z., Liu, Z., Ye, S., Gu, L., Tian, H., Duan, Y., Su, W., Shao, J., et al.: InternVL3: Exploring advanced training and test-time recipes for open-source multimodal models. arXiv preprint arXiv:2504.10479 (2025)