FARSIGHT: A Dynamic Multi Modal Medical Image Analysis Framework Powered by a Prior Knowledge Guided Vision Language Model

Jinghan Yang[1,2††], Qinghua Li[2,3†], Jingxin Liu[1,2], Li Qiu[4], Jie Tian[1,5], Jiaojiao Zhou[4✉], and Kun Wang[1✉]

[1] CAS Key Laboratory of Molecular Imaging, Institute of Automation, Chinese Academy of Sciences, Beijing, China
kun.wang@ia.ac.cn
[2] School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China
[3] School of Advanced Interdisciplinary Sciences, University of Chinese Academy of Sciences, Beijing, China
[4] Department of Medical Ultrasound, West China Hospital, Sichuan University, Chengdu, China
zhoujiaojiao@wchscu.cn
[5] School of Engineering Medicine, Beihang University, Beijing, China

Abstract. Long-term retrospective medical imaging archives cannot be directly utilized due to the lack of fine-grained annotations, requiring extensive manual curation by expert physicians in most previous medical imaging studies. Furthermore, the absence of standardized archiving protocols and the heavy reliance on subjective judgment for image retention lead to significant variations in the number of images per patient, while the presence of irrelevant frames introduces substantial noise. We propose FARSIGHT, an end-to-end archive-to-prediction framework designed to overcome these universal bottlenecks, evaluated here on renal transplant chronic allograft injury assessment. FARSIGHT integrates the Modality Aware Governance and Optimization System (MAGOS), an offline vision-language-model-guided curation agent, with hierarchical multi-modal multi-instance learning (HMMIL) for patient-level prediction. On a retrospective cohort of 4,576 patients comprising 41,203 images across four distinct modalities, MAGOS achieves 0.999 routing accuracy with complete parsing success. Using these curated inputs, HMMIL achieves a test AUC of 0.816 and an accuracy of 0.746 for patient-level prediction, while handling variable-length studies and missing modalities. The code is available at: https://github.com/hanlliumu/Farsight.

Keywords: Automated Data Curation, Vision-Language Models, Multiple Instance Learning, Real-World Data.

---

† These authors contributed equally.
✉ Corresponding authors: Jiaojiao Zhou and Kun Wang.

1 Introduction

Translating AI advances into large-scale clinical practice remains challenging because real-world medical archives are not model-ready: retrospective data often lack fine-grained annotations, contain unreliable metadata and non-diagnostic vendor overlays, and therefore demand substantial expert curation. Moreover, routine clinical archives are structurally irregular—studies vary widely in image counts, include redundant or noisy frames, and frequently miss or misalign modalities—complicating robust multi-modal learning.

Vision-language models (VLMs) provide a practical bridge from archives to models by recognizing modality-specific cues, interface elements, and layout patterns without task-specific training, enabling archive-level modality routing and ROI-oriented curation [1,2]. Yet free-form generation is difficult to audit, and institution-specific fine-tuning is often infeasible under strict offline privacy constraints [3,4]. These constraints motivate an agentic design that combines VLM perception with prior-knowledge rules and structured, verifiable outputs to enable reliable archive-level curation [5,6]. However, even after curation, clinical studies remain variable in length, partially observed across modalities, and sparsely informative at the frame level. For such settings, multiple-instance learning (MIL) provides a natural weakly supervised formulation in which each patient is represented as a bag of instances with a single patient-level label [7-11]. While MIL has shown strong performance in sparse-evidence tasks such as whole-slide imaging, it typically operates within a single-modality feature space. Clinical archives, in contrast, are multi-modal and heterogeneous, with imbalanced modality distributions and frequent missing modalities. Consequently, effective prediction requires modality-aware aggregation mechanisms that can account for cross-modality variability.

We propose FARSIGHT, an end-to-end archive-to-prediction framework integrating MAGOS with HMMIL. MAGOS performs modality routing and rule-constrained ROI extraction with schema-checked outputs, producing structured inputs without foundation-model fine-tuning [12,13]. HMMIL then aggregates evidence within and across modalities, accommodating variable-length studies and missing modalities without forced alignment. On renal transplant ultrasound for chronic allograft injury assessment, MAGOS achieves 0.999 routing accuracy on over 40,000 images with complete parsing success, and HMMIL improves test AUC from 0.777 to 0.816 and test accuracy from 0.700 to 0.746 over a strong baseline.

Our main contributions are as follows: We introduce FARSIGHT, a archive-to-prediction framework designed for real-world medical imaging repositories. We present MAGOS, an offline and auditable VLM-plus-rules curation agent that performs modality identification and ROI extraction without foundation-model fine-tuning. We propose HMMIL tailored to variable-length and incompletely paired clinical data, and we demonstrate robust performance under real-world curation noise and missing-modality conditions.

2 Method

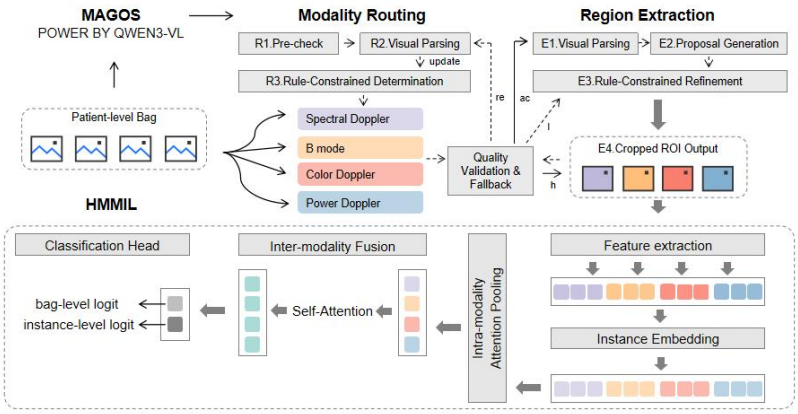

Fig. 1. MAGOS encodes radiologist-derived visual priors as explicit modality-routing rules: R1 validates inputs, R2 detects modality-specific cues, and R3 assigns B, C, P, or S labels. E1–E4 parse images, generate and refine modality-specific ROI proposals, and crop the resulting regions. A quality-validation and fallback module accepts, rejects, or revises unreliable outputs. HMMIL then aggregates the curated instances through intra-modality pooling and inter-modality fusion for patient-level prediction. re, reject; ac, accept; h, high quality; l, low quality.

2.1 Dataset and Cohort

This retrospective study included renal transplant recipients evaluated for chronic allograft injury (CAI). The patient-level prediction task was binary classification of CAI versus normal controls, with CAI as the positive class. The primary cohort included 4,576 patients from Hospital A between 2008 and 2025 and was split at the patient level to prevent data leakage into 3,204 training, 686 validation, and 686 internal test patients. These subsets were used exclusively to develop and evaluate HMMIL. The dataset contained 41,203 ultrasound images across four modalities, including 16,456 from patients with CAI and 24,747 from normal controls. Spectral Doppler (S) was the dominant modality with 22,981 images, followed by B-mode (B; 9,669), Color Doppler (C; 8,277), and Power Doppler (P; 276). Because of its limited sample size, Power Doppler was excluded from subsequent predictive analyses. Two additional image-level sets were reserved for independent MAGOS evaluation: Routing Evaluation Set A included 400 images for assessing modality routing, and ROI Evaluation Set B included 200 images for assessing ROI extraction. Neither set was used to optimize MAGOS or to train, validate, or test HMMIL. The cohort retained the irregular structure of real-world clinical examinations. Per-patient frame counts ranged from 1 to 46, with a median of 8. Most patients had three modalities, 133 had all four, and the remainder had one or two, reflecting routine missing-modality patterns in retrospective data. This variability provided a realistic and challenging setting for evaluating the robustness of the proposed multi-instance framework.

2.2 Overview of FARSIGHT System

As illustrated in Fig.1, FARSIGHT integrates MAGOS with HMMIL to enable robust patient-level diagnosis from raw dataset archives without fine-grained image-level annotations.

MAGOS. To address missing modality labels and ROI annotations in retrospective archives, we develop MAGOS, which decomposes curation into VLM-guided modality routing and deterministic, modality-specific ROI extraction. MAGOS employs Qwen3-VL-8B [14,15] for privacy-preserving on-prem deployment and reliable structured generation under instruction-following and constrained decoding.

Robust Vision Language Model Guided Modality Routing. The modality router maps each input image X to one of four standardized modality spaces $Y \in \{S, C, P, B\}$. To ensure reproducible and auditable decisions under highly heterogeneous ultrasound interfaces, we encode clinical prior knowledge as an expert-defined visual state space S_{vis} and apply a deterministic rule mapping implemented through prompting. The VLM first predicts key visual states, and the final modality label is obtained by applying the predefined decision rules with greedy decoding at zero temperature. To guarantee valid structured outputs, MAGOS uses JSON-Schema-constrained decoding. Rare residual parsing failures are handled by a lightweight fallback parser consisting of automatic syntax repair, abbreviation normalization, and regex-based recovery from clinical shorthand, enabling stable large-scale deployment within our dataset without crashes.

Spatial Heuristic Adaptive Region Extraction. After modality routing, MAGOS applies a modality-specific computer vision pipeline to extract regions of interest. Importantly, the VLM-based agent does not directly generate or crop bounding boxes. ROI localization is performed exclusively by a deterministic OpenCV-driven module using predefined spatial heuristics. Extracted ROIs are subsequently analyzed by the agent and reviewed by physicians to identify systematic errors or edge cases. Based on this feedback loop, heuristic rules are refined and updated. Thus, the closed-loop optimization updates rule parameters rather than delegating pixel-level localization to the VLM. To enhance cross-device generalization, spatial heuristics are defined using relative image dimensions instead of absolute pixel thresholds. Specifically, Color Doppler and Power Doppler regions are extracted from color overlays while suppressing UI artifacts; Spectral Doppler regions are localized via waveform panel detection; and B-mode regions are obtained by segmenting the principal grayscale scanning field using connected-component and morphology-based criteria [16].

HMMIL. After MAGOS provides modality-standardized, ROI-cropped images, we cast patient-level diagnosis as a weakly supervised multi-modal multi-instance learning problem to handle variable-length exams without image-level annotations. Our key innovations are a two-level hierarchical attention mechanism [17,18] aligned

with clinical modality structure, a masked two-stage Transformer fusion that yields auditable cross-modality interactions under missingness, and an auxiliary Top-K hard-negative instance supervision that improves instance discriminability and enables reliable saliency analysis. For each patient, we build a bag $B = \{(x_{m,i}, m)\}$, where $x_{m,i}$ denotes the *i*-th ROI image from modality m. The bag is naturally partitioned into modality-specific sub-bags B_m with size N_m. Missing modalities are represented by a modality mask and excluded from aggregation, enabling inference without fixed modality combinations. Following a lightweight backbone comparison, we adopt ShuffleNet as the instance encoder to produce embeddings $f_{m,i}$ with low memory overhead [19]. To avoid the brittleness of max pooling under noisy exams, we learn modality-aware instance importance within each sub-bag. The network first computes intra modality attention weights $\alpha_{m,i}$ to aggregate instances into a modality specific representation F_m, where w_{intra} and V_{intra} are learnable parameters of the intra modality attention network. This design explicitly matches the clinical acquisition process: instances are first filtered within modality before any cross-modality interaction, improving robustness to modality-specific redundancy and artifacts.

$$\alpha_{m,i} = \frac{exp\,(w_{intra}^{T}\,tanh\,(V_{intra} f_{m,i}))}{\sum_{j=1}^{N_m} exp\,(w_{intra}^{T}\,tanh\,(V_{intra} f_{m,j}))}, \quad F_m = \sum_{i=1}^{N_m} \alpha_{m,i}\, f_{m,i} \tag{1}$$

We introduce a two-stage Transformer fusion pipeline. An intra-modality Transformer first refines the instance sequence within each modality via self-attention, and the resulting modality tokens are then fused by an inter-modality Transformer to model cross-modality interactions. Each modality token is scaled by a learnable scalar weight and aggregated into a patient-level representation for the final classification head. Missing modalities are handled by masking and are excluded from aggregation, enabling robust inference under incomplete modality availability. Standard MIL can assign high attention to non-diagnostic frames, especially in negative patients where "hard negatives" resemble positives. We therefore propose an auxiliary Top-*K* Hard Negative Instance Loss to sharpen instance-level discriminability under weak supervision. For each bag, we select the Top-*K* instances by current instance score/attention. In positive bags, these instances serve as pseudo-positives and are optimized with a foreground objective; in negative bags, the Top-*K* highest-scoring instances are treated as hard negatives and are explicitly penalized. This targeted supervision improves the reliability of learned instance weights, directly benefiting downstream saliency reporting and per-image importance analysis.

3 Results

3.1 Implementation Details

All data preprocessing and model development were conducted on a Linux workstation powered by an Intel Core Ultra 9 processor. The environment was equipped with two NVIDIA GeForce RTX 5090 GPUs, each featuring 32 GB of

dedicated memory. The entire experimental framework was implemented utilizing PyTorch version 2.10.0 with comprehensive CUDA acceleration support. Further details on hyperparameters and VLM training available via public links.

3.2 Evaluation of MAGOS Framework

MAGOS Performance. MAGOS was evaluated using the two independent image-level sets defined in Section 2.1. An ultrasound physician with five years of clinical experience provided the reference modality labels for Independent Set A and manually delineated the regions of interest for Independent Set B. A second physician independently reviewed a subset of these annotations to assess interrater consistency. All MAGOS components were finalized before evaluation. Modality-routing performance was assessed on Independent Set A, whereas ROI extraction performance was assessed on Independent Set B.

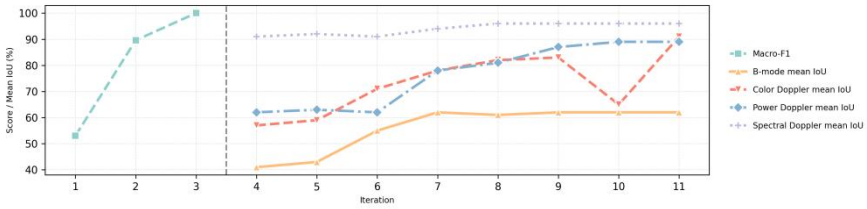

Fig. 2. Performance of the Magos across different iterations, including Macro-F1 score (cyan dashed line with squares), B mode mean IoU (orange solid line with triangles), Color Doppler mean IoU (red dashed line with circles), Power Doppler mean IoU (blue dash-dot line with diamonds), and Spectral Doppler mean IoU (purple dotted line with plus signs).

We trained a lightweight supervised baseline utilizing ResNet18 on Test Set A with a training and validation split of seventy to thirty. The network achieved near perfect accuracy on visually distinct modalities (B, C, S), but only 66.7% on P. Conversely, the zero shot vision language model guided routing module achieved superior overall accuracy without requiring a single annotated training sample. Deployed across over 40000 uncurated images, it attained a 99.98 percent overall accuracy derived from the consensus of two physicians who independently analyzed all cases. The system sustained a complete parsing success rate without a single pipeline crash. Among the valid images, comprehensive physician review revealed only three classification errors where Color Doppler cases completely lacking visible color overlays were assigned to B mode.

For quantitative evaluation, we defined algorithmic acceptance as achieving a mean Intersection over Union of at least 0.80 against the delineations of the physician on Test Set B. The finalized spatial heuristic algorithms achieved mean overlap metrics of 91 percent for Color Doppler, 89 percent for Power Doppler, and 96 percent for Spectral Doppler. For B mode images, the measured mean Intersection over Union was 62 percent. The lower mIoU of B is because we prioritized including the entire region to avoid expensive fine annotation, since the kidney occupies a large

proportion of the view, and this coarse extraction is sufficient for patient-level prediction tasks.

MIL Aggregation Comparison. As shown in Table 1, HMMIL achieves the best overall performance among all compared methods. It attains the highest AUC on the training, validation, and test sets with values of 0.841, 0.814, and 0.816, respectively, and also delivers the best accuracy with 0.772, 0.756, and 0.746. Compared with the strongest baseline MambaMIL, HMMIL improves test AUC from 0.777 to 0.816 and test ACC from 0.700 to 0.746, demonstrating more robust generalization. In this experiment, we use HMMIL with B-mode and Color Doppler ultrasound as the input modalities and ShuffleNet as the feature extraction backbone; the rationale for these choices is discussed in Section 3.4 Ablation Study and the Discussion section.

Table 1. Performance comparison with baselines on the training, validation, and test sets.

Methods	Training		Validation		Test	
	AUC	ACC	AUC	ACC	AUC	ACC
Mean Pooling	0.681	0.643	0.576	0.579	0.555	0.551
Max Pooling	0.648	0.583	0.631	0.620	0.588	0.558
ABMIL[7]	0.791	0.690	0.752	0.678	0.763	0.676
Gated ABMIL[8]	0.777	0.712	0.755	0.713	0.758	0.692
MambaMIL[9]	0.803	0.721	0.774	0.721	0.777	0.700
ACMIL[10]	0.785	0.688	0.752	0.683	0.754	0.677
MHIM-MIL[11]	0.791	0.761	0.761	0.708	0.770	0.692
HMMIL(Ours)	**0.841**	**0.772**	**0.814**	**0.756**	**0.816**	**0.746**

3.3 Ablation Study

We evaluated Magos under local deployment on two additional platforms, a Tesla V100 server and an RTX 5060 Ti workstation. The average latency is 10.42 s/img on V100 and 7.14 s/img on RTX 5060 Ti, compared with 3.04 s/img in our development environment. Despite the slowdown, the achieved throughput remains acceptable for our target deployment scenarios. We adopt HMMIL with B and C as the input modalities and ShuffleNet as the feature extraction backbone. Based on this setting, we conduct ablation experiments to quantify the impact of backbone selection, model components, and modality configuration. The results indicate that the choice of backbone materially affects performance, and the intra-modality module constitutes the primary contributor, as its removal yields the largest degradation. The instance-level loss, the inter-modality module, and hard-negative mining further provide consistent improvements. We analyzed attention weights and hard negative scores in representative patients. In positive cases, high-weighted instances consistently corresponded to B and C images. In contrast, S instances had lower attention weights, consistent with our quantitative modal ablation results. In addition, B is the most informative single modality, S alone is insufficient, and multi-modal fusion enhances discriminative capability. Collectively, these findings support the effectiveness of the

proposed HMMIL design and its key components in improving generalization on the validation and test sets. Detailed quantitative results are reported in Table 2.

Table 2. Ablation Study of HMMIL: Backbone, Module, and Modality Analysis (w/o: without; Intra: intra-modality module; Inter: inter-modality module; Linst: instance-level loss; Hard-N: hard-negative mining; B: B-mode; C: Color Doppler; S: Spectral Doppler).

Methods		Training		Validation		Test	
		AUC	ACC	AUC	ACC	AUC	ACC
Backbone analysis	Effientnetb0	0.745	0.684	0.637	0.613	0.595	0.566
	ResNet18	0.626	0.604	0.597	0.584	0.579	0.569
	Mobilenetv2	0.824	0.740	0.797	0.726	0.797	0.714
Module analysis	w/o Intra	0.800	0.733	0.761	0.717	0.730	0.667
	w/o Inter	0.770	0.704	0.721	0.690	0.750	0.709
	w/o Linst	0.796	0.675	0.767	0.682	0.764	0.641
	w/o Hard-N	0.810	0.731	0.760	0.692	0.770	0.699
Modality analysis	B+S	0.806	0.747	0.773	0.720	0.793	0.747
	B+C+S	0.814	0.736	0.770	0.714	0.808	0.750
	Only B	0.808	0.731	0.778	0.717	0.797	0.751
	Only C	0.798	0.734	0.756	0.704	0.781	0.706
	Only S	0.573	0.548	0.580	0.578	0.535	0.527
HMMIL(Ours)		**0.841**	**0.772**	**0.814**	**0.756**	**0.816**	**0.746**

4 Discussion

Our proposed FARSIGHT framework addresses this by providing an end-to-end solution: MAGOS offers zero-fine-tuning data curation, while HMMIL delivers patient-level prediction by handling irregular clinical data structures. This integrated approach bypasses the primary bottleneck of costly human intervention, enabling scalable deployment. From a translational perspective, FARSIGHT represents a preliminary framework for hospital environments, designed to be compatible with existing PACS workflows. It requires no changes to physician workflows, yet dramatically reduces the massive time and cost associated with building multi-modal retrospective cohorts by automating data governance and risk prediction. Regarding diagnostic utility for CAI, our experimental results indicated that C and B images provided significant diagnostic information, whereas Spectral Doppler images offered limited utility when analyzed purely visually without explicit scale information. This observation, that a purely visual approach to S waveforms proved less effective for extracting meaningful diagnostic information, contrasts with clinical practice where physicians routinely assess renal perfusion by analyzing waveform trends and numerical values.

While FARSIGHT demonstrates strong performance on our dataset, effectively addressing real-world challenges, certain limitations define our immediate path forward. Its current validation necessitates rigorous multi-center external testing and comprehensive evaluation on publicly available datasets to fully confirm Magos's

design robustness and generalizability. Furthermore, FARSIGHT's ultimate vision includes leveraging VLM capabilities for image-to-text generation to enhance task interpretability; however, current models like Qwen3-VL struggle with complex medical images due to their general domain training. Moving forward, a primary focus will be validating FARSIGHT's performance on multi-center datasets and benchmark public archives. Concurrently, we will fine-tune small-parameter, domain-specific VLMs or train novel foundation models tailored for medical imaging. This crucial step will enable FARSIGHT to generate insightful, interpretable text explanations, thereby completing its logical loop from data handling to comprehensive, explainable medical image analysis.

Acknowledgments. This work was supported by the National Key Research and Development Program of China (2023YFF1204600), National Natural Science Foundation of China (82441010, 82272029, U25C2048, 92259303, 62027901), and Beijing Natural Science Foundation (Z250003, 20250484964).

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

References

1. Dong, H., Kang, Z., Yin, W., Liang, L., Chao, C., Jiao, R.: Scalable vision language model training via high quality data curation. In: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL), pp. 33272-33293 (2025)
2. Udandarao, V., Burg, M.F., Albanie, S., Bethge, M.: Visual data-type understanding does not emerge from scaling vision-language models. arXiv preprint arXiv:2310.08577 (2023)
3. Zhao, Y., Zhong, E., Yuan, C., Li, Y., Zhao, M., Li, C., Hu, J., Liu, W., Liu, C.: Med-VLM: Enhancing medical image segmentation accuracy through vision-language model. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 7283-7293 (2025)
4. Jiang, S., Wang, Y., Song, S., Zhang, Y., Meng, Z., Lei, B., Wu, J., Sun, J., Liu, Z.: Omniv-med: Scaling medical vision-language model for universal visual understanding. arXiv preprint arXiv:2504.14692 (2025)
5. Jeong, D.P., Garg, S., Lipton, Z.C., Oberst, M.: Medical adaptation of large language and vision-language models: Are we making progress? In: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 12143-12170 (2024)
6. Silva-Rodríguez, J., Shakeri, F., Bahig, H., Dolz, J., Ben Ayed, I.: Few-shot, now for real: Medical VLMs adaptation without balanced sets or validation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), LNCS, pp. 237-247. Springer (2025)
7. Ilse, M., Tomczak, J., Welling, M.: Attention-based deep multiple instance learning. In: Proceedings of the International Conference on Machine Learning (ICML), pp. 2127-2136 (2018)

8. Andersson, A., Koriakina, N., Sladoje, N., Lindblad, J.: End-to-end multiple instance learning with gradient accumulation. In: 2022 IEEE International Conference on Big Data (Big Data), pp. 2742-2746 (2022)
9. Yang, S., Wang, Y., Chen, H.: Mambamil: Enhancing long sequence modeling with sequence reordering in computational pathology. In: International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), LNCS, pp. 296-306. Springer (2024)
10. Zhang, Y., Li, H., Sun, Y., Zheng, S., Zhu, C., Yang, L.: Attention-challenging multiple instance learning for whole slide image classification. In: European Conference on Computer Vision (ECCV), LNCS, pp. 125-143. Springer (2024)
11. Tang, W., Huang, S., Zhang, X., Zhou, F., Zhang, Y., Liu, B.: Multiple instance learning framework with masked hard instance mining for whole slide image classification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 4078-4087 (2023)
12. Lu, Y., Li, H., Cong, X., Zhang, Z., Wu, Y., Lin, Y., Liu, Z., Liu, F., Sun, M.: Learning to generate structured output with schema reinforcement learning. In: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL), pp. 4905-4918 (2025)
13. Guo, Z., Saluja, R., Yao, T., Liu, Q., Huo, Y., Liechty, B., Pisapia, D.J., Ikemura, K., Sabuncu, M.R., Yang, Y., et al.: Glo-vlms: Leveraging vision-language models for fine-grained diseased glomerulus classification. arXiv preprint arXiv:2508.15960 (2025)
14. Bai, S., Cai, Y., Chen, R., et al.: Qwen3-VL technical report. arXiv preprint arXiv:2511.21631 (2025)
15. Liu, L., Chen, L., Zhang, Z., Tian, M., Cui, H., Li, R., Liu, Z., Ju, Q., Li, Q., Zhou, H.-Y.: SkinFlow: Efficient information transmission for open dermatological diagnosis via dynamic visual encoding and staged RL. arXiv preprint arXiv:2601.09136 (2026)
16. Nankivell, B.J., Chapman, J.R., Gruenewald, S.M.: Detection of chronic allograft nephropathy by quantitative Doppler imaging. Transplantation 74(1), 90-96 (2002)
17. Wang, Q., Zhou, Y., Huang, J., Liu, Z., Li, L., Xu, W., Cheng, J.-Z.: Hierarchical attention-based multiple instance learning network for patient-level lung cancer diagnosis. In: 2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), pp. 1156-1160 (2020)
18. Li, Z., Jiang, Y., Lu, M., Li, R., Xia, Y.: Survival prediction via hierarchical multimodal co-attention transformer: A computational histology-radiology solution. IEEE Transactions on Medical Imaging 42(9), 2678-2689 (2023)
19. Zhang, X., Zhou, X., Lin, M., Sun, J.: ShuffleNet: An extremely efficient convolutional neural network for mobile devices. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6848–6856 (2018).