The Impact of Image Resolution on Biomedical Multimodal Large Language Models

Liangyu Chen LIANGYUC@STANFORD.EDU
Dept. of Computer Science
Stanford University
Stanford, 94305, CA

James Burgess JBURGESS@STANFORD.EDU
Institute for Computational and Mathematical Engineering (ICME)
Stanford University
Stanford, 94305, CA

Jeffrey Nirschl JNIRSCHL@STANFORD.EDU
Dept. of Pathology
Stanford University
Stanford, 94305, CA

Orr Zohar ORRZOHAR@STANFORD.EDU
Dept. of Electrical Engineering
Stanford University
Stanford, 94305, CA

Serena Yeung-Levy SYYEUNG@STANFORD.EDU
Dept. of Biomedical Data Science
Stanford University
Stanford, 94305, CA

Abstract

Imaging technologies are fundamental to biomedical research and modern medicine, requiring analysis of high-resolution images across various modalities. While multimodal large language models (MLLMs) show promise for biomedical image analysis, most are designed for low-resolution images from general-purpose datasets, risking critical information loss. We investigate how image resolution affects MLLM performance in biomedical applications and demonstrate that: (1) native-resolution training and inference significantly improve performance across multiple tasks, (2) misalignment between training and inference resolutions severely degrades performance, and (3) mixed-resolution training effectively mitigates misalignment and balances computational constraints with performance requirements. Based on these findings, we recommend prioritizing native-resolution inference and mixed-resolution datasets to optimize biomedical MLLMs for transformative impact in scientific research and clinical applications.

1. Introduction

Imaging technologies across a wide spectrum of resolutions are a cornerstone of biomedical research and modern medicine, providing critical insights into biological mechanisms and

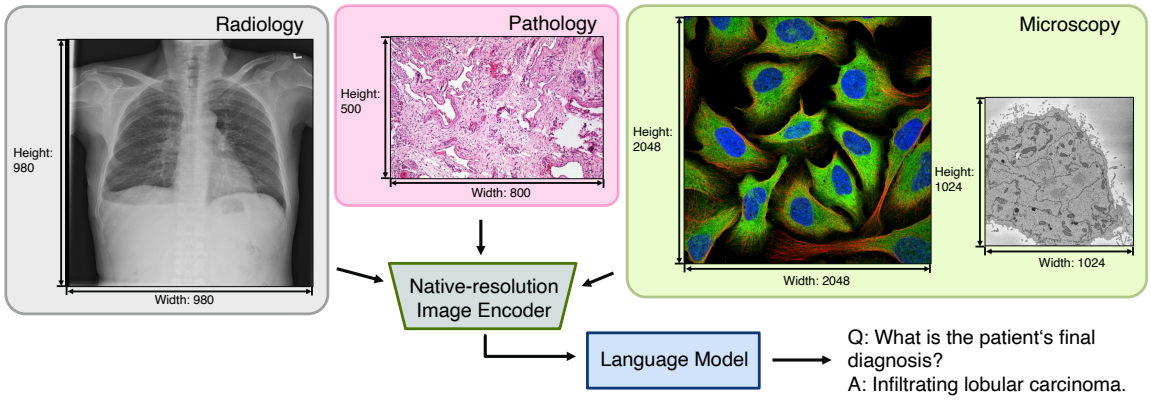

Figure 1: Biomedical MLLMs process images across modalities like radiology, pathology, and microscopy, however standard MLLMs are not designed to handle varying image resolution or high resolutions. This paper argues for native-resolution biomedical MLLMs. The image encoder processes arbitrary-resolution images into tokens dynamically aligned with language tokens, preserving high-frequency details critical for biomedical tasks. The textual output presents an example pathology question-answering task.

enabling advanced diagnostics and disease monitoring (Hussain et al., 2022). Biomedical imaging spans a wide range of scales and modalities, including chest X-rays (Lau et al., 2018), tissue histopathology (Chen et al., 2025), and fluorescence cell microscopy (Caie et al., 2010; Thul et al., 2017), all of which rely on large images that natively have a high-resolution to capture information at multiple levels. For instance, tumors in CT scans need a broad field of view to determine their anatomical location, alongside native-resolution details to classify tumor subtypes (Dunn et al., 2023). Similarly, studying protein function requires a macroscopic view of the entire cell for context and fine-grained resolution to characterize subcellular localization and function (Thul et al., 2017) (Figure 2). Such pathological and microscopic image resolutions range from 2048 by 2048 to more than 12000 by 12000 pixels.

Despite the importance of these native-resolution details, most existing multimodal large language models (MLLMs) are designed to process low-resolution images, such as those commonly found in general-purpose internet datasets (Li et al., 2023a). Adapting these models to native, high-resolution biomedical applications is challenging. Standard approaches often involve downsampling by approximately an order of magnitude to a fixed low resolution – this is necessary to enable image patch processing with fixed-size image encoders (Li et al., 2024; Xie et al., 2024), but it can obscure critical visual information. Other approaches select patches, omitting the global information (Chen et al., 2024a). This raises a key research question:

How can image resolution be effectively leveraged during training and inference to preserve the fine-grained features essential for biomedical applications?

To address this challenge, we investigate the role of resolution fidelity in MLLM performance across tasks where fine-grained visual interpretation is critical. We establish that

downsampling biomedical images during training or inference compromises the integrity of the visual data, limiting the utility of MLLMs in biomedical contexts. We explore strategies to mitigate this issue, by proposing to train biomedical MLLMs using architectures that support *native resolution* image encoding Bai et al. (2023a), thus naturally modeling image modalities with diverse and high resolutions. To adapt these models to biomedical images, we analyze how resolution impacts model performance and propose practical solutions for balancing computational complexity with visual details.

More specifically, our experiments demonstrate that when using native-resolution MLLMs during both training and inference significantly improves performance across multiple biomedical tasks, with improvements ranging from 0.54% to 6.8% in accuracy across different modalities. We further reveal that misalignment between training and inference resolutions can severely degrade model performance: accuracy drops by up to 48.7% when using native-resolution training with lower-resolution inference; and accuracy drops up to 43.3% when using lower-resolution training with native-resolution inference. To address the practical challenges of resolution variability in large-scale biomedical datasets, we propose a mixed-resolution training strategy that effectively maintains performance while accommodating computational constraints, achieving results nearly equivalent to aligned native-resolution training and inference with only a 1.0% average performance loss. These findings are further validated through zero-shot inference experiments on popular medical VQA benchmarks, where native-resolution inference improves results by 4.0%.

Based on these findings, we recommend that users of biomedical MLLMs prioritize native-resolution inference when working with models trained with mixed resolutions, and empirically evaluate different inference resolutions when model training details are unknown. For model developers, we advocate implementing balanced mixed-resolution training strategies at the modality level when constructing training datasets, as this approach effectively maintains performance while addressing practical computational constraints. These recommendations aim to optimize the deployment of MLLMs in biomedical applications while preserving the critical fine-grained features necessary for accurate analysis and interpretation.

Generalizable Insights about Machine Learning in the Context of Healthcare

Our work offers several key insights for machine learning applications in healthcare: (i) We demonstrate the critical importance of image resolution fidelity across multiple biomedical imaging modalities (X-rays, histopathology, microscopy), revealing a consistent pattern where native resolution significantly improves model performance – challenging the prevalent downsampling paradigm in medical image analysis; (ii) We identify a substantial performance degradation when training and inference resolutions are misaligned, highlighting the need for consistent resolution strategies throughout the ML pipeline; (iii) We propose a practical mixed-resolution training approach that balances computational constraints with performance requirements, achieving results comparable to fully native-resolution methods; and (iv) We provide actionable recommendations for both model users and developers working with high-resolution biomedical images. These findings extend beyond vision applications to establish a broader principle for healthcare ML: preserving the native information

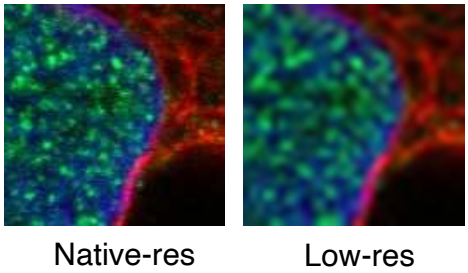

Figure 2: Comparison of image patches at native resolution (left) and low resolution. Downsampling often removes high-frequency details, such as fine textural details in protein staining, that are relevant for accurate interpretation.

density of medical data – regardless of modality – is essential for optimal model performance, especially when fine-grained features contain diagnostically relevant information.

2. Related Work

2.1. Image Resolution in Visual Recognition

Image resolution has been identified as a key attribute of visual recognition. Hao et al. (2023) evaluates the impact of image resolution on object detection. Resolution also affects medical image segmentation performance (Rajaraman et al., 2023). Touvron et al. (2019) raises the train-test resolution discrepancy. They propose an object-resolution-invariant image augmentation technique to counter the effect. Although the augmentation method does not fit modern MLLMs, we identified the resolution discrepancy degrades MLLM performances and proposed solutions.

2.2. High-resolution MLLMs

Most existing MLLMs, such as those utilizing pretrained encoders (Radford et al., 2021; Liu et al., 2023; Alayrac et al., 2022; Li et al., 2023a,c), face limitations in their fixed input resolution, often downscaling images to meet computational constraints. This reduction in resolution leads to a loss of critical fine-grained details that are essential for many specialized tasks in cell biology and pathology, such as small object detection, histopathological analysis (Xu et al., 2024a), and microscopic imaging (Lozano et al., 2024; Burgess et al., 2025).

Several approaches have emerged to address these challenges. Chain-of-Spot (Liu et al., 2024) uses a chain-of-thought prompt templated to localize regions of interest before the user query. In a similar vein, RLogist (Zhao et al., 2023) trains a zoom-in strategy by reinforcement learning to localize pathological tissue of interest. While these enhancement methods can be effective, they often struggle to retain global context, add latency, and fail to improve the employed model. End-to-end models, LLaVA-UHD (Xu et al., 2024b) and LLaVA-OneVision, (Li et al., 2024) employ a multi-crop strategy to handle high-resolution inputs by splitting large images into smaller segments, thus preserving essential local features while managing computational complexity. Qwen-VL (Bai et al., 2023b), PaLI-3

Dataset	Modality	Resolution
Subcellular (Thul et al., 2017)	Immunofluorescence microscopy	2048×2048
Compound (Caie et al., 2010)	Fluorescence microscopy	1280×1024
Cervical (Hussain et al., 2020)	Liquid-based cytology	2048×1536
WSI (Chen et al., 2025)	Whole-slide pathology	$\sim 3000 \times 4000$ (0.25×)
VQA-RAD (Lau et al., 2018)	Radiology	$\sim 1024 \times 1024$
PathVQA (He et al., 2020)	Pathology	$\sim 750 \times 400$
SLAKE (Liu et al., 2021)	Clinical images (multimodal)	$\sim 1024 \times 1024$

Table 1: Datasets. WSI images are resized because of compute constraints. All other images are in native resolutions.

(Chen et al., 2023b), and PaLI-X (Chen et al., 2023a) attempt to gradually scale the input resolution of their pretrained encoders, but these approaches often still need to reduce image size, potentially overlooking important visual details at training. Qwen2-VL (Wang et al., 2024) introduces naive dynamic resolution support, allowing the model to flexibly adapt to varying image sizes by employing 2D-RoPE (Heo et al., 2024) and post-ViT token compression to limit memory usage and maintain efficiency. We use Qwen2-VL as the base model to adapt to the various resolutions of biomedical images.

2.3. Biomedical Applications of MLLMs

Integrating MLLMs in biomedical applications has shown promising advancements, particularly in enhancing the interpretative capabilities across various imaging domains. Models like LLaVA-Med (Li et al., 2023b) and BiomedGPT (Zhang et al., 2023) have pioneered efforts to merge medical imaging with scientific textual data, effectively supporting tasks such as diagnosis, visual question answering, and medical report generation. These models build on general-purpose LLMs by introducing specialized biomedical instruction-following datasets (Li et al., 2023b; Xie et al., 2024), which enhance their ability to understand domain-specific visual and textual cues.

However, many existing biomedical MLLMs are constrained by limited input resolution. For example, Visual Med-Alpaca (Shu et al., 2023) and LLaVA-Med (Li et al., 2023b) employed CLIP-based image encoders, which are restricted by their native low-resolution capabilities. Such limitations hinder these models' ability to capture detailed biomedical imaging signals – biomedical MLLM training data has diverse resolutionsLozano et al. (2025) , and tasks often require understanding of microanatomical structures. Recently, Dragonfly (Chen et al., 2024a) and Llama3-Med (Chen et al., 2024b) leveraged multi-resolution branches with pretrained vision encoders to support high-resolution biomedical images. However, they still lack flexibility in resolution or fail to scale to very high resolutions (millions of pixels) that are essential to many biomedical applications. Both methods applied hierarchical resolution branches to process the thumbnail image and high-resolution patches, which causes computation overhead by processing redundant visual information. Moreover, none of the prior works studied the impact of inference resolution on performance.

3. Experiments

Our experiments focus on classification and Visual Question Answering (VQA) tasks, where an MLLM is provided with an image and a text-based question and must generate an accurate answer. We evaluate performance across seven tasks from three representative biomedical imaging modalities (radiology, pathology, microscopy), chosen for their reliance on native-resolution data and their importance in both research and clinical practice. For clarity in the figures and tables, we denote these datasets as "Subcellular", "Compound", "Cervical", "WSI", "VQA-RAD", "PathVQA", "SLAKE" (Figure 2.2). Leveraging the QwenVL-2 architecture (Wang et al., 2024), which dynamically processes images of varying resolutions by splitting them into fixed-size patches (Figure 1), we preserve rich visual details with reasonable computational resources.

Through this study, we aim to highlight the critical importance of native-resolution image processing in the design and application of biomedical MLLMs, while offering practical recommendations to optimize their performance. Our findings highlight the critical role of native-resolution images in advancing biomedical MLLMs. First, we demonstrate that native-resolution training significantly improves performance across multiple biomedical classification and visual question-answering tasks. Second, we establish that alignment between training and inference resolutions is crucial, as misalignment leads to substantial performance degradation. To address the practical challenge posed by resolution variability in large-scale biomedical datasets, we propose mixed-resolution training, which effectively mitigates misalignment issues while preserving the benefits of native-resolution inference. Based on these insights, we recommend that future biomedical MLLMs prioritize native-resolution inference and that training datasets incorporate a balanced mix of resolutions to maximize performance and generalizability. These principles are essential for optimizing MLLMs to meet the demands of fine-grained biomedical image analysis.

3.1. Data

To comprehensively evaluate the performance of multimodal large language models (MLLMs) in biomedical applications, we conducted experiments on seven diverse datasets that span critical imaging modalities and biomedical tasks (Figure 2.2).

- Subcellular, Thul et al. (2017) This dataset contains immunofluorescence microscopy images of human cells, annotated for subcellular protein localization . It provides native, high-resolution images essential for studying protein function and cellular context. We measure the cell line classification accuracy on this dataset.

- Compound, Caie et al. (2010) This dataset focuses on fluorescence microscopy-based high-content screening for compound profiling. Images capture cellular responses to chemical perturbations, requiring fine-grained details for accurate classification. We measure the compound profiling classification accuracy on this dataset.

- Cervical, Hussain et al. (2020) This dataset includes native-resolution images from liquid-based cytology, annotated for pre-cancerous and cervical cancer lesions. The dataset is critical for evaluating diagnostic capabilities in cervical cytology. We measure the lesion classification accuracy on this dataset.

- WSI, Chen et al. (2025) This dataset comprises whole-slide pathology images annotated for tasks such as tumor classification and diagnostic visual question answering.

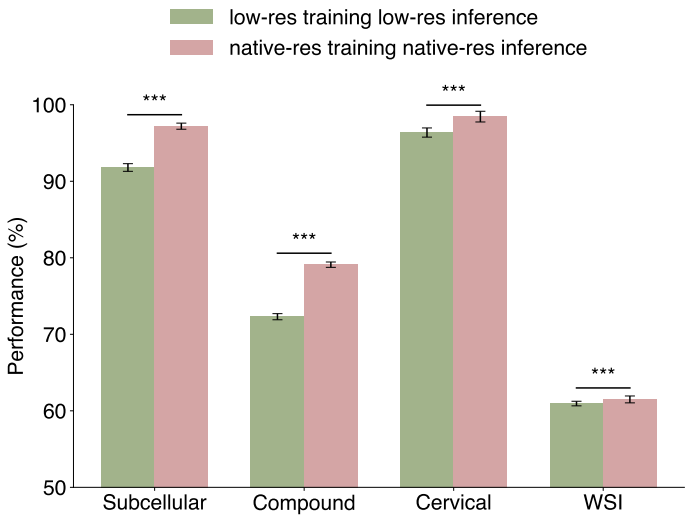

Figure 3: Performance comparison of native-resolution and low-resolution training and inference. Native-resolution training and inference achieve superior results, emphasizing the importance of resolution fidelity.

These large, native-resolution images capture essential morphological details across broad tissue areas. Because the original resolution is too high to fit in the memory of a single server, we downsample it to be around 25% of the original resolution, which is a bout 3000*4000. We measure the accuracy of the closed-form question-answering on this dataset.

- VQA-RAD, Lau et al. (2018) VQA-RAD is a medical VQA benchmark dataset containing radiology images paired with clinically relevant questions. The questions cover diverse topics, such as anatomical structures and disease diagnosis, requiring both visual and textual understanding. We measure the closed-set question-answering accuracy on this dataset.

- PathVQA, He et al. (2020) PathVQA is a VQA dataset based on pathology images, designed to test a model's ability to answer questions about cellular and tissue-level features. The dataset emphasizes the need for native-resolution image processing. We measure the closed-set question-answering accuracy on this dataset.

- SLAKE, Liu et al. (2021) SLAKE is another medical VQA dataset that focuses on multimodal reasoning over clinical images and associated textual data. Its questions demand fine-grained visual interpretation alongside contextual understanding. We measure the closed-set question-answering accuracy on the English subset of this dataset.

3.2. Experiment Settings

We use Qwen2-VL 2B (Wang et al., 2024) as a base model unless otherwise specified, with no architectural change. Qwen2-VL was chosen for two reasons: (1) it handles arbitrary image resolution by mapping image patches into a dynamic number of image tokens that merge with language tokens, balancing performance and compute efficiency; (2) it was the

Table 2: Resolution alignment in training and inference for Qwen2-VL 2B. Mixed-resolution training enables models to adapt effectively to both low-, native-, and mixed-resolution inference.

Training	Inference	Subcellular	Compound	Cervical	WSI	VQA-RAD	PathVQA	SLAKE
No	Low-res	1.80	8.60	21.35	14.55	39.25	16.96	49.09
	Native-res	3.85	9.10	22.63	18.36	39.65	18.19	48.71
	Mixed-res	2.86	8.84	21.94	16.45	39.43	17.52	48.92
Low-res	Low-res	91.80	72.30	96.37	60.95	30.93	20.81	56.63
	Native-res	64.35	29.00	82.90	58.91	29.58	19.20	52.48
	Mixed-res	77.68	50.45	89.46	59.85	30.32	20.04	54.57
Native-res	Low-res	48.50	46.30	80.83	54.45	38.54	18.65	54.74
	Native-res	97.20	79.10	98.45	61.49	42.49	24.88	60.23
	Mixed-res	72.85	62.70	89.64	58.06	40.56	21.75	57.48
Mixed-res	Low-res	78.55	58.80	89.24	56.93	39.35	20.45	55.98
	Native-res	95.20	78.00	97.93	61.34	41.92	22.94	58.03
	Mixed-res	90.48	72.65	95.31	60.24	41.08	22.15	57.36

Table 3: Resolution alignment in training and inference for InternVL2.5 2B. Mixed-resolution training enables models to adapt effectively to both low-, native-, and mixed-resolution inference.

Training	Inference	Subcellular	Compound	Cervical	WSI	VQA-RAD	PathVQA	SLAKE
No	Low-res	2.15	9.45	23.86	16.12	41.75	18.45	52.18
	Native-res	4.32	10.28	25.19	20.43	42.06	19.84	51.92
	Mixed-res	3.29	9.82	24.53	18.25	41.89	19.14	52.05
Low-res	Low-res	94.38	75.64	97.25	63.47	32.85	22.54	59.21
	Native-res	68.74	31.56	84.68	61.35	31.42	20.86	55.13
	Mixed-res	81.53	53.20	90.92	62.41	32.15	21.68	57.32
Native-res	Low-res	52.64	49.85	83.47	57.28	40.96	20.34	57.42
	Native-res	98.45	82.36	99.12	64.23	45.18	26.73	63.15
	Mixed-res	75.32	66.15	91.26	60.74	42.95	23.47	60.18
Mixed-res	Low-res	82.16	62.47	92.38	59.84	41.87	22.18	58.76
	Native-res	97.35	81.24	98.64	64.05	44.63	24.85	60.94
	Mixed-res	93.67	76.83	97.02	63.28	43.95	24.08	60.28

best-performing MLLM in general domains at the time of experiments. The model has a simple architecture orchestrating a native-resolution image encoder and a language model, as illustrated in Figure 2 of Wang et al. (2024).

In a common vision-language modeling approach, the Qwen2-VL model employs a language-pretrained large language model for further vision-language pretraining and instruction tuning. The vision-language training data details are not revealed.

Our finetuning on biomedical data was done with controlled hyperparameters. The total batch size is 128. The micro-batch size and gradient accumulation steps vary to control the

Table 4: Resolution alignment in training and inference for LLaVA-OneVision 2B. Mixed-resolution training enables models to adapt effectively to both low-, native-, and mixed-resolution inference.

Training	Inference	Subcellular	Compound	Cervical	WSI	VQA-RAD	PathVQA	SLAKE
No	Low-res	1.45	7.23	18.76	12.38	34.85	14.27	43.65
	Native-res	3.12	7.84	19.85	15.67	35.42	15.63	43.18
	Mixed-res	2.34	7.58	19.30	14.06	35.15	14.94	43.47
Low-res	Low-res	82.75	65.48	89.75	54.32	27.84	18.26	50.15
	Native-res	58.63	24.75	76.53	52.45	26.73	16.85	47.29
	Mixed-res	70.42	45.13	83.16	53.40	27.31	17.63	48.74
Native-res	Low-res	43.27	40.65	74.38	48.93	34.62	16.48	49.35
	Native-res	88.64	71.85	91.26	55.74	38.75	21.95	54.87
	Mixed-res	65.85	56.24	82.91	52.35	36.58	19.16	52.13
Mixed-res	Low-res	70.86	51.74	82.65	50.87	35.48	18.17	50.76
	Native-res	87.35	70.94	90.84	55.28	37.94	20.65	53.25
	Mixed-res	83.64	65.32	88.75	54.16	37.31	19.94	52.48

Table 5: Resolution alignment in training and inference for Qwen2-VL 7B. Mixed-resolution training enables models to adapt effectively to both low-, native-, and mixed-resolution inference.

Training	Inference	Subcellular	Compound	Cervical	WSI	VQA-RAD	PathVQA	SLAKE
No	Low-res	2.46	10.24	23.59	16.82	42.57	19.34	52.64
	Native-res	4.93	11.35	25.41	21.52	43.24	20.85	52.18
	Mixed-res	3.78	10.93	24.53	19.37	42.95	20.12	52.43
Low-res	Low-res	96.37	78.54	98.25	65.42	33.85	32.36	61.27
	Native-res	69.28	33.76	85.62	63.45	32.38	30.82	56.84
	Mixed-res	82.48	56.32	92.14	64.37	33.13	31.65	59.05
Native-res	Low-res	53.42	52.37	84.76	59.78	41.95	35.82	58.93
	Native-res	98.45	85.23	99.35	67.26	46.78	40.47	64.85
	Mixed-res	76.32	68.94	92.18	63.52	44.32	38.24	61.92
Mixed-res	Low-res	83.62	64.85	92.75	61.34	42.85	33.75	60.38
	Native-res	97.83	83.45	98.83	66.85	45.95	38.67	62.73
	Mixed-res	93.47	78.38	97.25	65.24	44.86	37.23	61.85

total batch size. The learning rate is 1e-5 with a cosine scheduler. The weight decay is 0.1. The experiments are performed with NVIDIA A100 and A6000 graphic processing units.

All the models are fully finetuned with bf16 and DeepSpeed Zero2. The training time ranged from 10 to 120 A100 GPU hours, depending on the resolution and image quantities. We compared full finetuning (38.54%, training for 8 A100 hours) with LoRA (30.22%, training for 4.8 A100 hours) on VQA-RAD when we started the experiments, native-res training low-res inference. We choose full finetuning for the rest of the experiments because it performs better with manageable computing expenditure.

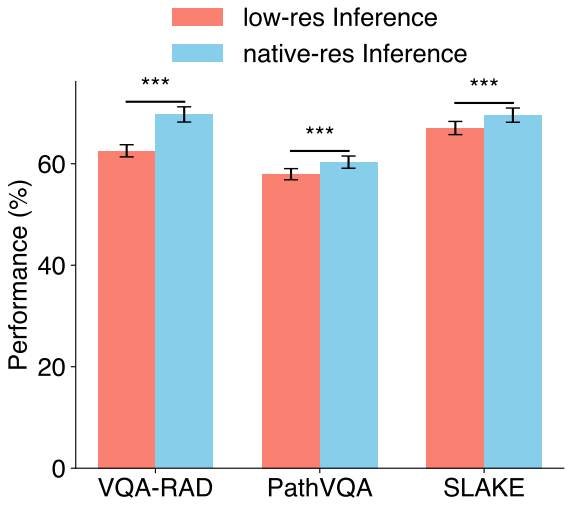

Figure 4: Native-resolution inference improves off-the-shelf models. Qwen2-VL 7B shows consistent gains on radiological and pathological question-answering tasks with native-resolution input.

3.3. Impact of Native Resolution on Model Performance

Our first finding establishes that performing training and inference with native-resolution images improves performance on biomedical VQA tasks. We fine-tune the base MLLM in native or low resolution, and performs inference with the same resolution. The results are in Figure 3. Native-resolution training improves immunofluorescence microscopy cell (Thul et al., 2017) classification accuracy by 5.4%, $P < 0.001$; fluorescence microscopy compound-profiling (Caie et al., 2010) classification accuracy by 6.8%, $P < 0.001$; pre-cancerous and cervical cancer lesion (Hussain et al., 2020) classification accuracy by 2.08%, $P < 0.001$.; whole-slide pathology (Chen et al., 2025) visual question answering accuracy by 0.54%, $P < 0.001$. These results support the intuition that image processing tasks requiring fine details need native-resolution images, suggesting that biomedical MLLMs should incorporate native-resolution images.

3.4. Resolution Alignment Between Training and Inference

Our second finding reveals that misalignment between training and inference resolutions can substantially degrade performance in biomedical classification/VQA tasks. When we use a model trained with native resolution but perform inference with lower resolution, we observe severe performance degradation: -48.7%, -32.8%, -17.6% and -7.0% across tasks (Table 2). Similarly, training with lower resolution and testing on native resolution leads to significant performance losses of -27.4%, -43.3%, -13.5% and -2.0%. Notably, these misaligned configurations perform worse than consistently using lower resolution for both training and inference, indicating that resolution alignment between training and testing is more crucial than the actual resolution used for inference. Decreasing inference resolution

leads to more severe performance than the other way around. We attribute it to the task difficulty gap between training and inference. Training with native resolution and inference with low resolution imposes a more difficult task that the model didn't learn at training.

However, these findings present a practical challenge: biomedical MLLMs typically train on large datasets from various sources with different image resolutions, making strict training-inference resolution alignment impractical. To address this challenge, we investigated mixed-resolution training, where 50% of training samples use lower resolution. Our results in Table 2 show that mixed-resolution training effectively mitigates the problems of misaligned train-test resolutions. We observe consistent results on other models like InternVL2.5 2B (Table 3), LLaVA-OneVision 2B (Table 4), and larger models like Qwen2-VL 7B (Table 5). When using high-resolution inference with mixed-resolution training, performance nearly matches that of aligned native-resolution training and inference, with only a 1.0% average performance loss. Furthermore, with mixed-resolution training, native-resolution inference consistently outperforms lower-resolution inference by an average margin of 12.2%. Notably, the compute cost of MLLM increases proportionally with the number of pixels. Thus mixed-resolution training also balances between compute efficiency and performance requirements. For model developers, we recommend implementing balanced mixed-resolution training strategies, as our results show that while training-inference misalignment can be catastrophic, 50-50 mixed-resolution training effectively maintains performance. We suggest implementing this balance at the level of each imaging modality when constructing training datasets.

3.5. Inference Strategy

Based on the previous findings, we offer two practical recommendations. For users inferencing with biomedical MLLMs, we recommend using native-resolution inference when working with models known to be trained with mixed resolutions. If the model's training resolution details are unknown, users should empirically evaluate different inference resolutions to determine the optimal setting. We validate this approach through zero-shot inference experiments with a pretrained Qwen2-VL 7B model on standard medical VQA benchmarks: VQA-RAD (Lau et al., 2018), PathVQA (He et al., 2020), and SLAKE (Liu et al., 2021). These experiments confirm that inference resolution significantly affects performance when the training resolution of a massive pre-trained model is unknown, with native resolution improving results by 4.0% on average (Figure 4).

3.6. Resizing

In order to ablate the effect of image resolution, we compare images upsampled to the native resolution from low-resolution images, as shown in Figure 5. Our analysis of different resampling techniques reveals that advanced resampling methods can partially mitigate the performance degradation caused by resolution reduction, though they cannot fully restore native-resolution performance. Lanczos resampling consistently outperforms other methods across all tasks, achieving 93.1% accuracy on subcellular classification compared to the native-resolution performance of 97.2%, and 75.2% on compound classification versus 79.1% with native resolution. Bicubic resampling shows moderate improvements over bilinear resampling, with performance gains of 2.5% and 3.3% on subcellular and compound

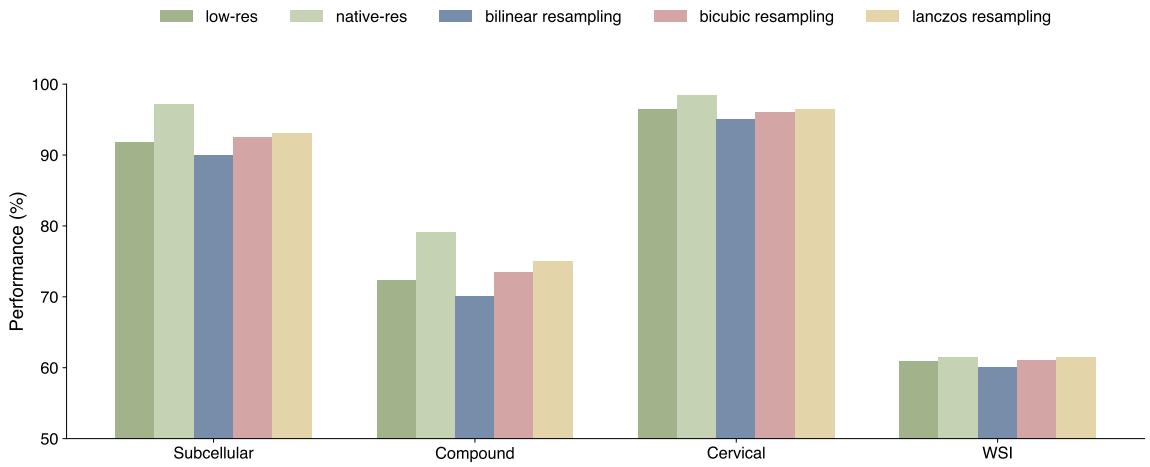

Figure 5: Comparing low-resolution, high-resolution, and images resampled from low-resolution images.

classification tasks, respectively. Notably, the effectiveness of resampling methods varies across different imaging modalities, with the greatest benefits observed in tasks requiring fine-grained feature preservation, such as subcellular and compound analysis, while showing minimal impact on whole-slide imaging tasks where the performance gap between resampled and native resolution remains relatively constant.

Based on our experimental results, we recommend a hierarchical approach to resolution handling in biomedical MLLMs. When data access permits, native resolution should be maintained as it consistently delivers superior performance across all tasks. In resource-constrained environments where a user only has access to low-resolution images, Lanczos resampling should be prioritized as the preferred downsampling method, as it recovers 95.7% of native-resolution performance on average across all tasks.

4. Discussion

In this study, we present two key findings about image resolution in biomedical MLLMs, followed by practical recommendations for implementation. First, we demonstrate that native-resolution images improve performance across multiple biomedical visual question-answering (VQA) tasks. Second, we show the critical importance of alignment between training and inference resolutions. These findings lead us to important practical considerations for both users and developers of biomedical MLLMs.

Broader Impact The findings from this study have significant implications for the deployment of MLLMs in healthcare settings, potentially improving diagnostic accuracy and research outcomes across various biomedical imaging modalities. By establishing best practices for resolution handling in biomedical MLLMs, our work could accelerate the development of more reliable artificial intelligence systems for medical image analysis, potentially

leading to earlier disease detection and more accurate diagnoses. However, the computational resources required for native-resolution processing may limit accessibility to well-resourced institutions, potentially exacerbating healthcare disparities. Additionally, improved performance of these systems could lead to over-reliance on automated analysis, emphasizing the importance of maintaining human oversight and using these tools as aids rather than replacements for clinical expertise.

Limitations and Future Work While our study demonstrates the importance of resolution fidelity in biomedical MLLMs, several limitations should be acknowledged. First, our experiments primarily focus on specific imaging modalities and may not generalize to all types of biomedical imaging. The computational demands of native-resolution processing also present practical constraints for real-time applications and resource-limited settings. Furthermore, our mixed-resolution training strategy, while effective, may not be optimal for all scenarios, and the ideal ratio of resolution mixing might vary across different applications and imaging modalities. Future work should explore more efficient architectures for handling multi-resolution inputs and investigate adaptive resolution selection mechanisms based on task-specific requirements and computational constraints.

Acknowledgement We thank Xiaohan Wang, Alejandro Lozano, Anita Rau, and Sanket Gupte for their thoughtful discussions.

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736, 2022.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023a.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 2023b.

James Burgess, Jeffrey J Nirschl, Laura Bravo-Sánchez, Alejandro Lozano, Sanket Rajan Gupte, Jesus G Galaz-Montoya, Yuhui Zhang, Yuchang Su, Disha Bhowmik, Zachary Coman, et al. Microvqa: A multimodal reasoning benchmark for microscopy-based scientific research. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 19552–19564, 2025.

Peter D Caie, Rebecca E Walls, Alexandra Ingleston-Orme, Sandeep Daya, Tom Houslay, Rob Eagle, Mark E Roberts, and Neil O Carragher. High-content phenotypic profiling of drug response signatures across distinct cancer cells. *Molecular cancer therapeutics*, 9(6): 1913–1926, 2010.

Kezhen Chen, Rahul Thapa, Rahul Chalamala, Ben Athiwaratkun, Shuaiwen Leon Song, and James Zou. Dragonfly: Multi-resolution zoom supercharges large visual-language model. *arXiv preprint arXiv:2406.00977*, 2024a.

Pingyi Chen, Chenglu Zhu, Sunyi Zheng, Honglin Li, and Lin Yang. Wsi-vqa: Interpreting whole slide images by generative visual question answering. In *European Conference on Computer Vision*, pages 401–417. Springer, 2025.

Xi Chen, Josip Djolonga, Piotr Padlewski, Basil Mustafa, Soravit Changpinyo, Jialin Wu, Carlos Riquelme Ruiz, Sebastian Goodman, Xiao Wang, Yi Tay, et al. Pali-x: On scaling up a multilingual vision and language model. *arXiv preprint arXiv:2305.18565*, 2023a.

Xi Chen, Xiao Wang, Lucas Beyer, Alexander Kolesnikov, Jialin Wu, Paul Voigtlaender, Basil Mustafa, Sebastian Goodman, Ibrahim Alabdulmohsin, Piotr Padlewski, Daniel Salz, Xi Xiong, Daniel Vlasic, Filip Pavetic, Keran Rong, Tianli Yu, Daniel Keysers, Xiaohua Zhai, and Radu Soricut. Pali-3 vision language models: Smaller, faster, stronger, 2023b.

Zekai Chen, Arda Pekis, and Kevin Brown. Advancing high resolution vision-language models in biomedicine. *arXiv preprint arXiv:2406.09454*, 2024b.

Bryce Dunn, Mariaelena Pierobon, and Qi Wei. Automated classification of lung cancer subtypes using deep learning and ct-scan based radiomic analysis. *Bioengineering*, 10(6): 690, 2023.

Yu Hao, Haoyang Pei, Yixuan Lyu, Zhongzheng Yuan, John-Ross Rizzo, Yao Wang, and Yi Fang. Understanding the impact of image quality and distance of objects to object detection performance. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 11436–11442. IEEE, 2023.

Xuehai He, Yichen Zhang, Luntian Mou, Eric Xing, and Pengtao Xie. Pathvqa: 30000+ questions for medical visual question answering. *arXiv preprint arXiv:2003.10286*, 2020.

Byeongho Heo, Song Park, Dongyoon Han, and Sangdoo Yun. Rotary position embedding for vision transformer. In *European Conference on Computer Vision*, pages 289–305. Springer, 2024.

Elima Hussain, Lipi B Mahanta, Himakshi Borah, and Chandana Ray Das. Liquid based-cytology pap smear dataset for automated multi-class diagnosis of pre-cancerous and cervical cancer lesions. *Data in brief*, 30:105589, 2020.

Shah Hussain, Iqra Mubeen, Niamat Ullah, Syed Shahab Ud Din Shah, Bakhtawar Abduljalil Khan, Muhammad Zahoor, Riaz Ullah, Farhat Ali Khan, and Mujeeb A Sultan. Modern diagnostic imaging technique applications and risk factors in the medical field: a review. *BioMed research international*, 2022(1):5164970, 2022.

Jason J Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. A dataset of clinically generated visual questions and answers about radiology images. *Scientific data*, 5(1):1–10, 2018.

Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Jingkang Yang, and Ziwei Liu. Otter: A multi-modal model with in-context instruction tuning. *arXiv preprint arXiv:2305.03726*, 2023a.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *NeurIPS*, abs/2306.00890, 2023b.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36, 2024.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023c.

Bo Liu, Li-Ming Zhan, Li Xu, Lin Ma, Yan Yang, and Xiao-Ming Wu. Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pages 1650–1654. IEEE, 2021.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning, 2023.

Zuyan Liu, Yuhao Dong, Yongming Rao, Jie Zhou, and Jiwen Lu. Chain-of-spot: Interactive reasoning improves large vision-language models. *arXiv preprint arXiv:2403.12966*, 2024.

Alejandro Lozano, Jeffrey Nirschl, James Burgess, Sanket Rajan Gupte, Yuhui Zhang, Alyssa Unell, and Serena Yeung-Levy. {\mu}-bench: A vision-language benchmark for microscopy understanding. *arXiv preprint arXiv:2407.01791*, 2024.

Alejandro Lozano, Min Woo Sun, James Burgess, Liangyu Chen, Jeffrey J Nirschl, Jeffrey Gu, Ivan Lopez, Josiah Aklilu, Anita Rau, Austin Wolfgang Katzer, et al. Biomedica: An open biomedical image-caption archive, dataset, and vision-language models derived from scientific literature. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 19724–19735, 2025.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, pages 8748–8763. PMLR, 2021.

Sivaramakrishnan Rajaraman, Feng Yang, Ghada Zamzmi, Zhiyun Xue, and Sameer Antani. Assessing the impact of image resolution on deep learning for tb lesion segmentation on frontal chest x-rays. *Diagnostics*, 13(4):747, 2023.

Chang Shu, Baian Chen, Fangyu Liu, Zihao Fu, Ehsan Shareghi, and Nigel Collier. Visual med-alpaca: A parameter-efficient biomedical llm with visual capabilities. 2023.

Peter J Thul, Lovisa Åkesson, Mikaela Wiking, Diana Mahdessian, Aikaterini Geladaki, Hammou Ait Blal, Tove Alm, Anna Asplund, Lars Björk, Lisa M Breckels, et al. A subcellular map of the human proteome. *Science*, 356(6340):eaal3321, 2017.

Hugo Touvron, Andrea Vedaldi, Matthijs Douze, and Hervé Jégou. Fixing the train-test resolution discrepancy. *Advances in neural information processing systems*, 32, 2019.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

Yunfei Xie, Ce Zhou, Lang Gao, Juncheng Wu, Xianhang Li, Hong-Yu Zhou, Sheng Liu, Lei Xing, James Zou, Cihang Xie, et al. Medtrinity-25m: A large-scale multimodal dataset with multigranular annotations for medicine. *arXiv preprint arXiv:2408.02900*, 2024.

Hanwen Xu, Naoto Usuyama, Jaspreet Bagga, Sheng Zhang, Rajesh Rao, Tristan Naumann, Cliff Wong, Zelalem Gero, Javier González, Yu Gu, et al. A whole-slide foundation model for digital pathology from real-world data. *Nature*, pages 1–8, 2024a.

Ruyi Xu, Yuan Yao, Zonghao Guo, Junbo Cui, Zanlin Ni, Chunjiang Ge, Tat-Seng Chua, Zhiyuan Liu, Maosong Sun, and Gao Huang. Llava-uhd: an lmm perceiving any aspect ratio and high-resolution images. *arXiv preprint arXiv:2403.11703*, 2024b.

Kai Zhang, Jun Yu, Zhiling Yan, Yixin Liu, Eashan Adhikarla, Sunyang Fu, Xun Chen, Chen Chen, Yuyin Zhou, Xiang Li, Lifang He, Brian D. Davison, Quanzheng Li, Yong Chen, Hongfang Liu, and Lichao Sun. Biomedgpt: A unified and generalist biomedical generative pre-trained transformer for vision, language, and multimodal tasks, 2023.

Boxuan Zhao, Jun Zhang, Deheng Ye, Jian Cao, Xiao Han, Qiang Fu, and Wei Yang. Rlogist: fast observation strategy on whole-slide images with deep reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 3570–3578, 2023.

Appendix A. Data and Code Availability

All the datasets are available online from the original sources:

- Subcellular, Thul et al. (2017) Immunofluorescence microscopy images of human cells, annotated for subcellular protein localization.

- Compound, Caie et al. (2010) Fluorescence microscopy-based high-content screening for compound profiling.

- Cervical, Hussain et al. (2020) Native-resolution images from liquid-based cytology, annotated for pre-cancerous and cervical cancer lesions.

- WSI, Chen et al. (2025) Whole-slide pathology images annotated for tasks such as tumor classification and diagnostic visual question answering.

- VQA-RAD, Lau et al. (2018) A medical VQA benchmark dataset containing radiology images paired with clinically relevant questions.

- PathVQA, He et al. (2020) A VQA dataset based on pathology images, designed to test a model's ability to answer questions about cellular and tissue-level features.

- SLAKE, Liu et al. (2021) A medical VQA dataset, focusing on multimodal reasoning over clinical images and associated textual data.

The third-party code for training is available on GitHub (https://github.com/modelscope/ms-swift). The evaluation code is available on (https://github.com/cliangyu/med_eval).

Institutional Review Board (IRB) This study uses public datasets under the data license. The original authors previously de-identified any patient-derived images in compliance with applicable privacy laws and institutional guidelines. The Institutional Review Board guidelines were reviewed, and the public use of deidentified images does not constitute human subjects research.