Enhancing vision-language models for medical imaging: bridging the 3D gap with innovative slice selection

Yuli Wang[1], **Jian Peng**[2], **Yuwei Dai**[1], **Craig Jones**[1], **Haris Sair**[1], **Jinglai Shen**[3], **Nicolas Loizou**[1], **Jing Wu**[2], **Wen-Chi Hsu**[1], **Maliha Imami**[1], **Zhicheng Jiao**[4], **Paul Zhang**[5], **Harrison Bai**[1]
[1]Johns Hopkins University
[2]Second Xiangya Hospital, Central South University
[3]University of Maryland, Baltimore County
[4]Brown University
[5]University of Pennsylvania
{ywang687,ydai55,craigj,hsair1,nloizou,whsu20,mimami1,hbai7}@jhu.edu
{pengjian666@csu,wujing622}@csu.edu
shenj@umbc.edu
zhichengjiao@brown.edu
paul.zhang2@pennmedicine.upenn.edu

Abstract

Recent approaches to vision-language tasks are built on the remarkable capabilities of large vision-language models (VLMs). These models excel in zero-shot and few-shot learning, enabling them to learn new tasks without parameter updates. However, their primary challenge lies in their design, which primarily accommodates 2D input, thus limiting their effectiveness for medical images, particularly radiological images like MRI and CT, which are typically 3D. To bridge the gap between state-of-the-art 2D VLMs and 3D medical image data, we developed an innovative, one-pass, unsupervised representative slice selection method called Vote-MI, which selects representative 2D slices from 3D medical imaging. To evaluate the effectiveness of Vote-MI when implemented with VLMs, we introduce BrainMD, a robust, multimodal dataset comprising 2,453 annotated 3D MRI brain scans with corresponding textual radiology reports and electronic health records. Based on BrainMD, we further develop two benchmarks, BrainMD-select (including the most representative 2D slice of a 3D image) and BrainBench (including various vision-language downstream tasks). Extensive experiments on the BrainMD dataset and its two corresponding benchmarks demonstrate that our representative selection method significantly improves performance in zero-shot and few-shot learning tasks. On average, Vote-MI achieves a 14.6% and 16.6% absolute gain for zero-shot and few-shot learning, respectively, compared to randomly selecting examples. Our studies represent a significant step toward integrating AI in medical imaging to enhance patient care and facilitate medical research. We hope this work will serve as a foundation for data selection as vision-language models are increasingly applied to new tasks. Code and data examples are available at Github: https://github.com/YuliWanghust/BrainMD.

38th Conference on Neural Information Processing Systems (NeurIPS 2024) Track on Datasets and Benchmarks.

https://doi.org/10.52202/079017-3171

1 Introduction

Generalist foundation models, or large vision-language models (VLMs), such as GPT-4V [1], have revolutionized artificial intelligence by leveraging diverse large-scale datasets during pre-training. These models excel across multiple domains, including natural language processing and computer vision [33, 48, 55, 58, 18, 54], positioning them at the forefront of medical imaging advancements. However, a significant limitation of these state-of-the-art (SOTA) models is their restriction to 2D image input. This results in obstacles for their application to medical imaging, particularly with radiological images that are often 3D. To address these challenges, we propose a representative 2D slices selection approach called **Vote-MI**. This one-pass, unsupervised method selects representative 2D slices from 3D images, bridging the gap between SOTA VLMs and medical imaging. By employing Vote-MI, we aim to enhance diagnostic accuracy and automate medical reporting through the application of SOTA VLMs to 3D medical image analysis.

To thoroughly assess the effectiveness of our proposed Vote-MI method, a large, multimodal medical image dataset with paired textual data is essential. Therefore, we introduce **BrainMD**, a comprehensive dataset encompassing seven different types of brain tumors (details in Table 3). BrainMD includes 2,453 annotated 3D MRI brain scans, paired with textual data such as radiology reports, medical records, and demographic information. Based on BrainMD, we developed two benchmarks: **BrainMD-select and BrainBench**. BrainMD-select comprises the most representative 2D slices from the axial, sagittal, and coronal directions of 3D images, annotated by board-certified radiologists. BrainBench is derived from textual data and encompasses various tasks such as disease diagnosis and visual question answering. The dataset and its two benchmarks enable the evaluation of VLMs, ensuring robust model testing and accelerating advancements in AI-driven medical imaging diagnostics. Additionally, BrainMD and its associated benchmarks hold significant potential to benefit the broader research community by facilitating the future development of other 2D/3D VLMs.

Given that VLMs can perform downstream tasks with zero or few task demonstrations [31, 50], thereby eliminating the need for parameter updates, we evaluate the effectiveness of Vote-MI in VLMs through downstream task evaluations, including zero-shot and few-shot learning. Zero-shot testing [57] gauges the model's ability to tackle tasks without prior examples, utilizing its generalization capabilities from training data to novel tasks. Few-shot testing [51], or in-context learning, provides an alternative to traditional supervised tuning. In this study, we explore these capabilities in VLMs using the BrainMD dataset and Vote-MI method, thoroughly comparing model performance across random, Vote-MI, and radiologist-selected slices in zero-shot and few-shot scenarios.

Our experiments, over BrainMD and its two Benchmarks, demonstrate that our representative selective method substantially improves the VLM zero-shot and few-shot testing performance by balancing the diversity and representativeness of selected samples. For instance, Vote-MI achieves an average of 14.6% and 16.6% absolute gain for zero-shot and few-shot learning, respectively, compared to randomly selecting examples. Moreover, the improvement is consistent across different VLMs. Vote-MI representative selection also makes zero-shot learning and few-shot learning learning more stable and reduces the variance. Detailed results are shown in Section 6.

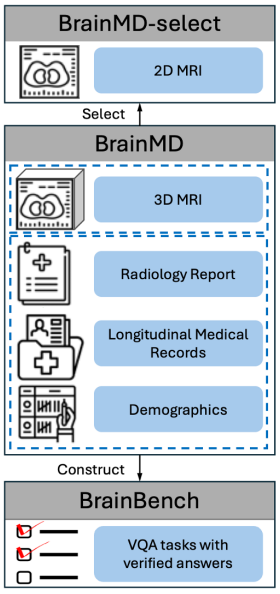
The code of Vote-MI and a few BrainMD examples are open-sourced. Our contributions are summarized as follows:

2 Related Works

2.1 Zero-shot and Few-shot Learning

Zero-shot learning (ZSL) [57] enables models to predict classes they haven't been explicitly trained on by leveraging auxiliary information. This approach addresses the challenge of acquiring labeled data for every class, especially in domains like rare medical diseases. Various ZSL methodologies include attribute-based [24], embedding-based [59], and generative approaches [30]. The versatility of ZSL has significant implications in vision and language processing, enhancing models' ability to generalize across diverse and unseen categories.

Few-shot learning [51] requires only a few annotated examples per test instance, avoiding the need for extensive fine-tuning. Recent research has proposed strategies to enhance few-shot learning, such as meta-training [34, 42], task instructions [26], and task formulation [22]. The selection of few-shot



- **A large-Scale, multimodal longitudinal brain tumor dataset.** BrainMD (as shown in Figure 1) comprises 2453 studies from 561 patients, each including: (1) high-resolution 3D volumetric MRI images, (2) radiologist-selected representative slices in the coronal, sagittal, and axial directions, (3) paired radiology reports, and (4) structured longitudinal medical records. To our knowledge, this is the first dataset that combines 2D/3D medical images with both radiology reports and longitudinal medical records.
- **A novel representative slice selection method.** Vote-MI, a one-pass unsupervised representative slice selective method, is proposed to adapt SOTA 2D VLM for 3D medical image analysis. This method will enable the identification of the most representative 2D slices from complex 3D medical imaging datasets, allowing large VLMs to intake 3D imaging data by capturing the most important information from volumetric data.
- **Two Benchmarks for vision-language model evaluation.** Based on BrainMD, we establish two benchmarks for diagnosing and prognosticating outcomes of brain tumors. We conduct extensive evaluations, including zero-shot and few-shot testing scenarios, to assess VLM performance and capability.

Figure 1: Schematics of BrainMD Dataset and two benchmarks: BrainMD-select and BrainBench. The BrainMD dataset comprises MRI scans, paired with their radiology reports and medical records. BrainMD-select contains Radiologists selected representative slices. We curated diagnostic and prognostic labels based on the radiology reports and medical records to construct BrainBench.

examples is crucial, with studies questioning the necessity of correct labels. Our work introduces representative slice selection within a few-shot learning framework, emphasizing the importance of representative examples on the performance of VLMs.

2.2 Multi-modal Dataset

Publicly available medical datasets continue to drive significant advances in medical AI research [28, 4, 41, 14, 53]. However, very few currently available datasets are large-scale, multi-modal, and extensively labeled, particularly in medical domains that leverage both 2D and 3D medical images (Table 1). The limitations in data availability primarily stem from the inherent challenges associated with the release of medical data. Public sharing of medical data requires rigorous review processes to safeguard sensitive patient information from exposure [27]. Furthermore, the labeling process is often labor-intensive and costly [17, 61, 60].

Among previous contributions, the BRATS dataset [25] stands out as a large-scale work incorporating 3D MRI brain tumor images. However, BRATS lacks paired textual data and selected representative 2D slices. The BMs [38] and TCIA [52] datasets contain 3D MRI images with longitudinal medical record data. However, both have small dataset scales and do not include prognostic task labels or selected representative slices. Our BrainMD dataset addresses these gaps by introducing a large-scale dataset extracted from 2,453 brain tumor cases. It offers multiple modalities and labels, promising to enrich future research in this space.

2.3 Medical Multi-modal Vision Language Model

Recent research [33, 62] highlights the effectiveness of multimodal vision-language models (VLMs) in integrating image and text data for a variety of tasks. These models combine the perceptual capabilities of vision models [40, 47] with the generative power of large language models (LLMs) [43, 11, 13], gaining significant traction, particularly in medical image analysis. Existing medical VLMs [32, 49, 2, 37] often fine-tune publicly available 2D VLMs on medical image and text data to perform tasks such as image-text retrieval, report generation, and visual question answering. Models like LLaVA-Med [32], Med-PaLM-2 [49], and MedFlamingo [37] are derived from LLaVA [35], PaLM-E [12], and Flamingo [2], respectively.

https://doi.org/10.52202/079017-3171

Table 1: BrainMD vs. existing multimodal brain tumor medical image datasets

Dataset	Modalities				Counts		Task Labels	
	Image	Slice	Report	EHR	Patients	Studies	Diagnosis	Prognosis
BRATS [25]	3D MRI	×	×	×	228	many	×	×
Figshare [10]	2D MRI	×	×	×	3,064	3,064	×	×
SARTAJ [7]	2D MRI	×	×	×	3,260	3,260	×	×
BMs [38]	3D MRI	×	×	✓	75	637	2	×
TCIA [52]	3D MRI	×	×	✓	47	156	2	×
BrainMD	3D MRI	✓	✓	✓	561	2,453	2	1

However, these methods face challenges when applied to 3D medical images, such as CT and MRI scans, which contain rich spatial information. The common approach of slice-by-slice analysis is computationally expensive and often inadequate. While models like RadFM [55] support both 2D and 3D images, they primarily focus on text generation tasks like visual question answering (VQA) and generally underperform. More specialized VLMs, such as M3D-LaMed [5], Ct2rep [19], and Merlin [8], are designed specifically for 3D medical image analysis, tackling tasks like report generation and VQA, and pioneering vision question-answering tasks. Despite these advancements, 3D VLMs continue to struggle due to the lack of large, paired 3D image-report datasets and the high computational demands of model training.

3 Vote-MI: Representative slice selection method

In addressing the methodological challenges of transitioning from 2D to 3D medical images, particularly when employing SOTA VLMs, we propose the efficient, unsupervised Vote-MI method. This approach aims to efficiently identify highly diverse and representative 2D slices from 3D medical images in just one pass. Our method includes two main parts: 1) An unsupervised feature extraction process that operates directly on raw, unannotated images, and 2) A new criterion for assessing image diversity and representativeness during the selection process.

As shown in Figure 2, the representative selection pipeline consists of three major components: a) A patch-wise Variational Autoencoder (VAE) [39, 36] that serves as the unsupervised feature extraction network, effectively transposing each image sample into a low-dimensional feature descriptor; b) The Vote-MI algorithm, which identifies and selects a diverse and representative subset of images from the pool of unannotated data; and c) The VLMs, which are used downstream in various diagnostic and prognostic tasks. More details about the representative selection pipeline, including feature extraction network, Vote-MI representative slices selection, and other representative selection methods, are described in Appendix D.

4 Cohort Definition and Dataset Composition

Our study, approved by the Johns Hopkins University (Appendix A), identified 2,453 cases involving MRI brain tumor scans from 2010 to 2020. The cohort of MRI brain tumors was identified through a protocol involving random sampling, data cleaning, and inclusion criteria, resulting in a final cohort of 2,453 cases from 561 distinct patients (see Appendix C.1 for more details). Each of these images is de-identified. Summary statistics of the demographic characteristics of our final cohort are available in Table 2. Based on this cohort, we release the following as the BrainMD dataset:

- MRI images: The imaging slices for the BrainMD cohort in DICOM format.
- Radiology Report: The "Findings" and "Impression" section of the corresponding radiologist reports for all cases in the BrainMD cohort.
- Data From Medical Records: De-identified structured data from longitudinal records for each patient in our cohort, including diagnoses, procedures, lab results, and demographics.

A detailed description of the formatting and licensing details of BrainMD is in the Appendix C.2.

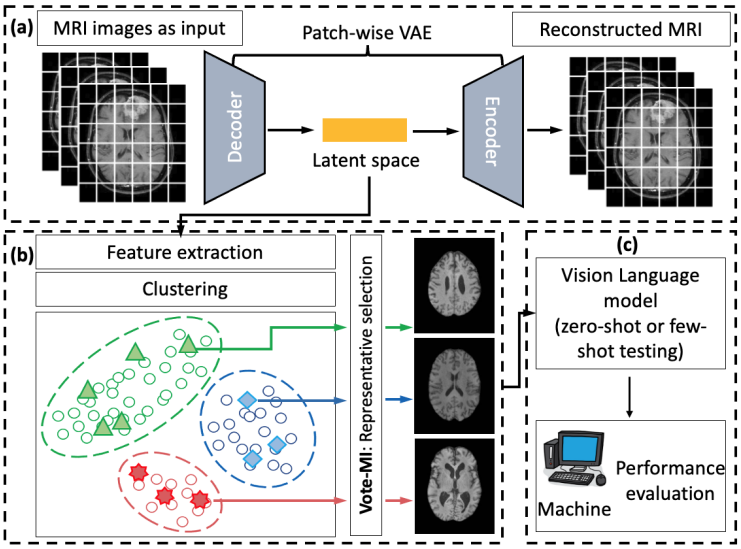


Figure 2: The workflow of our one-pass representative selection framework on brain MRI; (a) 2D patch-wise VAE for feature extraction; (b) Vote-MI: clustering-based representative selection; (C) downstream task evaluation using VLMs with zero-shot and few-shot scenarios.

Table 2: Demographic statistics of the proposed BrainMD dataset.

Demographics Statistics		
Attributes		**All**
	Cases	2,453
	Patients	561
Gender	Female	245 (43.6%)
	Male	237 (42.2%)
	Unknown	79 (14.2%)
Age	0-18	38 (6.8%)
	18-55	123 (21.9%)
	>55	387 (69.0%)
	Unknown	13 (2.3%)
Race	White	393 (70.0%)
	Asian	24 (4.2%)
	Black	29 (5.2%)
	Unknown	115 (20.6%)

5 Benchmark

To evaluate our proposed Vote-MI method and further demonstrate its effectiveness when applied to VLM downstream tasks, we developed two benchmarks: BrainMD-select and BrainBench. BrainMD-select, illustrated in Figure 1, is a 2D dataset created to evaluate the selection efficacy of Vote-MI. This dataset comprises the most representative 2D slices annotated by radiologists from the 3D BrainMD dataset. Furthermore, we introduce BrainBench, a benchmark designed to assess the performance of our representative slice selection method within SOTA VLMs, particularly under zero-shot and few-shot scenarios. BrainBench encompasses various tasks such as disease diagnosis, visual question answering, and even report generation.

These benchmarks, detailed in Appendix C.2, involve varied task formulations to allow comprehensive evaluations across different scenarios. Our experiments, as outlined in section 5.1 and 5.2, test the

https://doi.org/10.52202/079017-3171

efficacy of the Vote-MI selection method and its adaptability within VLMs under zero- or few-shot scenarios.

5.1 The effectiveness studies on the Vote-MI

To evaluate the Vote-MI method's effectiveness, we compare its output against four baselines: Uncertainty [23], Core-set [16], K-center [21], and Random sampling. Uncertainty selects instances where its confidence is lowest, indicating high uncertainty; Core set chooses data points furthest from the major data cluster, thus adding the most informative instances; K-center selects k data points as centers such that the maximum distance from any data point to its nearest center is minimized. We assess the performance of Vote-MI and these baselines using our BrainMD-select dataset, where radiologist selection is the gold standard. To ensure statistical validity, each method is run three times. An ablation study is also conducted to determine the contribution of each component in the Vote-MI method, which is shown in Appendix D.3.

5.2 Zero-shot and Few-shot Learning Tasks

We compare the performance of the VLMs, specifically Flamingo [2], Med-Flamingo [37], and Med-PaLM-2 [44] (due to our computation limitation) in a zero-shot setting using our custom benchmark, BrainBench (Figure 1). The models are evaluated on the following 2 diagnosis tasks:

- "Presence of cancer in the image (yes/no)? (**W/o cancer**)"
- "Name brain cancer types? (**Cancer types**)"

We adopt a bifurcated approach for few-shot learning (Figure 3). Initially, a representative subset is curated, selecting a finite collection of samples for labeling ahead of evaluation. Subsequently, for each test sample, pertinent examples are collated from this curated set, a step termed random prompt retrieval. The total labeling effort is demarcated by the number of samples curated and annotated in the preliminary phase. The second phase is constrained by the VLM's input capacity. Within these bounds, Vote-MI is recommended for its strategic selection of varied and indicative samples for selection. The model's performance is then assessed utilizing the BrainBench benchmark over two tasks as delineated in zero-shot learning and a new prognostic task as follows:

- "Describe the cancer status? (**Cancer status**)"

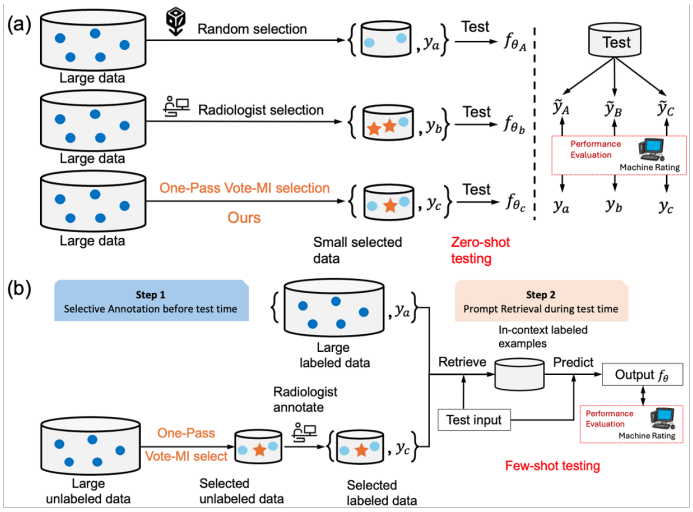


Figure 3: Schematic framework of downstream tasks for (a) zero-shot learning and (b) few-shot learning. In (a), we compare the one-pass Vote-MI representative 2D slice selection from 3D imaging with random and radiologist selections. For (b), the framework for few-shot learning utilizes Vote-MI, with alternatives being random or radiologist selections for the downstream task.

The statistical distribution of task labels for ground-truth of each sub-category ("W/o cancer", "Cancer types", and "Cancer status") is summarized in Table 3.

Table 3: Statistical analysis of three diagnostic and prognostic task labels in the BrainMD dataset, encompassing 2,453 cases and 561 patients. LGG: Low-Grade Gliomas; GBM: Glioblastoma Multiforme; CPTs: Choroid plexus tumors

Task Labels Statistics			
Types	**Question**	**Answers**	**All**
			Cases 2,453 Patients 561
Diagnosis	**W/o cancer**	Yes No Uncertain	1714 (69.9%) 381 (15.5%) 357 (14.6%)
Diagnosis	**Cancer type**	GBM Glioma LGG Pineal tumors Medulloblastoma CPTs Gangliocytoma	1326 (54.0%) 715 (29.1%) 242 (9.8%) 26 (1.1%) 72 (2.9%) 51 (2.0%) 20 (0.81%)
Prognosis	**Cancer status**	Improving Progressing No change	38 (6.8%) 123 (21.9%) 387 (69.0%)

5.3 Measuring Accuracy and Stability

Accuracy: These three tasks have predefined answer choices. Thus, we utilize accuracy (denoted as "ACC"), measuring the proportion of correctly identified cases to evaluate the VLM downstream tasks' performance.

Stability: Given a set of raw data, our Vote-MI representative slice selection method is not deterministic, with certain randomness. To assess the stability of Vote-MI and its impact on VLM performance, we conduct each experiment three times and average the results. Despite its non-deterministic nature, Vote-MI consistently enhances stability compared to other selection methods for both prognostic and diagnostic tasks.

6 Results and Analysis

6.1 Effectiveness studies of Vote-MI

Table 4 summarizes the slice selection accuracy of different methods on the BrainMD dataset and BrainMD-select benchmark. The accuracies are calculated with a slice number error tolerance of $\pm$ 5, given the nature of brain tumor images where multiple representative slices can capture the characteristics of the tumors. The Vote-MI method achieved the highest accuracy at 59.4% and the lowest variance $\pm$ 4.2%, which is statistically significantly better than the other baseline methods.

Table 4: Performance comparison of different selection methods on BrainMD dataset.

	Uncertainty	**Core-set**	**K-center**	**Random**	**Vote-MI**
Accuracy	52.2 ($\pm$ 4.9)	53.5 ($\pm$ 7.0)	47.6 ($\pm$ 5.4)	28.2 ($\pm$ 8.6)	59.4 ($\pm$ 4.2)

https://doi.org/10.52202/079017-3171

6.2 Zero-shot Learning Results

As shown in Figure 4 are our results from zero-shot learning over the BrainMD dataset with two downstream tasks including: 1) with or without cancer (binary classification) and 2) identifying cancer type (vision question answering). Over all datasets, the Vote-MI representative selective method outperforms the random baseline by a large margin (23.5% absolute gain on average in w/o cancer and 14.4% absolute gain on average in cancer type) under the zero-shot scenario.

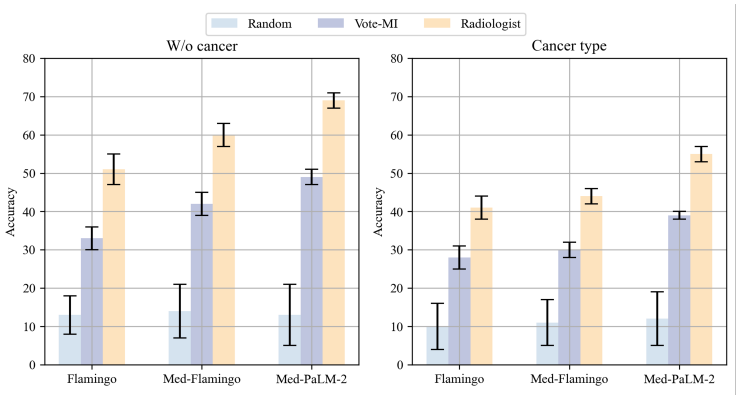


Figure 4: Performance comparisons of various VLMs under two disease diagnostics downstream task settings. Compared to random selection, the Vote-MI representative selection method consistently improves the performance with a large margin of zero-shot learning with pre-trained VLMs.

6.3 Few-shot Learning Results

Table 5: Few-shot learning results with random, radiologist, and Vote-MI selection methods on the BrainMD dataset, with a selection budget of 2, 4, or 8. Across the board, representative selection with Vote-MI substantially outperforms the random selection baseline for few-shot learning. Further, Vote-MI largely reduces the variance over three trials (see the results in Appendix D.3), making few-shot learning more stable. Δ means absolute gain between Random and Vote-MI.

Size	**Method**	**Tasks**		
$\lvert\mathcal{L}\rvert$	Selection	W/o cancer	Cancer types	Cancer status
2	Random	21.3 ($\pm$ 5.6)	18.3 ($\pm$ 7.1)	19.6 ($\pm$ 4.0)
2	Vote-MI	43.5 ($\pm$ 2.3)	35.3 ($\pm$ 4.5)	37.7 ($\pm$ 3.8)
2	Radiologist	64.5 ($\pm$ 3.1)	55.3 ($\pm$ 4.2)	48.4 ($\pm$ 3.9)
2	Δ Absolute gain	+22.2	+17.0	+18.1
4	Random	28.1 ($\pm$ 5.1)	24.5 ($\pm$ 6.2)	26.7 ($\pm$ 5.8)
4	Vote-MI	50.7 ($\pm$ 2.8)	43.0 ($\pm$ 3.9)	42.7 ($\pm$ 4.0)
4	Radiologist	68.2 ($\pm$ 3.4)	60.2 ($\pm$ 4.5)	52.7 ($\pm$ 4.3)
4	Δ Absolute gain	+22.6	+19.5	+16.0
8	Random	30.1 ($\pm$ 6.1)	28.7 ($\pm$ 6.5)	30.4 ($\pm$ 6.3)
8	Vote-MI	55.3 ($\pm$ 3.1)	46.0 ($\pm$ 4.2)	46.7 ($\pm$ 4.1)
8	Radiologist	70.1 ($\pm$ 3.5)	62.3 ($\pm$ 4.3)	56.9 ($\pm$ 4.4)
8	Δ Absolute gain	+15.2	+18.3	+16.3

In this study, we perform an extensive analysis of few-shot learning to provide further guidance, examining representative slice selection from multiple dimensions: varying VLMs, different downstream tasks, and selection sizes. Our findings from the BrainMD dataset, detailed in Table 5, show results for selection budgets ranging from 2, 4 to 8. This range accommodates the input limits of VLMs, allowing full integration of examples into prompts without additional sampling. Across all tasks and VLMs, the Vote-MI method for selecting representative slices significantly outperforms a random baseline for all selection sizes, with a standout 16.6% average absolute gain when the set

size is 8 (see Figure 5). Notably, using just two Vote-MI selected examples achieves better outcomes than eight randomly chosen ones across all tasks, highlighting the effectiveness of strategic example selection in few-shot learning.

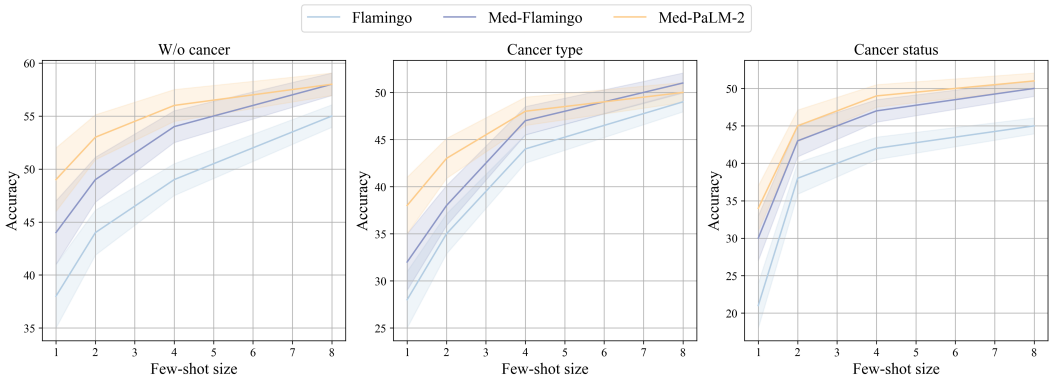


Figure 5: Few-shot testing comparison among three different downstream tasks. Vote-MI representative slice selection improves the performance of zero-shot learning with pre-trained VLMs.

6.4 Impact of Vote-MI on Stability

As shown above, our extensive experiments demonstrated that representative selection yields higher accuracy across all downstream tasks and is crucial for the VLMs' success in both zero-shot and few-shot learning. However, Vote-MI is not a deterministic representative selection method and includes certain randomness, conditioned on a set of unlabeled image samples and a selection budget. Thus, we explored the stability of the method. From our experiments, we observed a reduction in variance for both zero-shot and few-shot learning across all downstream tasks. Therefore, the variance of Vote-MI arises solely from how the unlabeled samples are collected, significantly improving the robustness of zero-shot and few-shot learning. We, therefore, recommend that researchers and practitioners use representative slice selection methods (e.g., our Vote-MI method) to better benefit from the zero-shot and few-shot learning capabilities of VLMs with increased stability.

7 Limitation and Conclusion

7.1 Limitations

First, BrainMD contains data from only a single site, and the Vote-MI representative selection model or other future models trained on BrainMD may not generalize to other patient populations. Second, although labels are assigned based on large language model output and manually reviewed by radiologists, there still might be inaccuracies in some cases. Finally, although the effectiveness of Vote-MI improved significantly compared to random and other slice selection methods, the downstream task performance is still statistically significantly worse than the radiologist's selection. We hope our paper can serve as a baseline and inspire further research on methods for representative selection, bridging the gap between 2D vision language models and 3D medical image data.

Future efforts will focus on optimizing the representative selection framework to further improve accuracy. This includes researching potentially better feature extraction networks. Given the relatively high homogeneity within tumor lesions, generative-based (e.g., diffusion probabilistic models [45, 56]) or contrastive-based unsupervised learning methods [9, 20] may be more effective and accurate in doing the feature extraction. Additionally, new criteria for assessing image diversity and representativeness are needed. Beyond the mutual information metrics used in our paper, graph-based metrics [46, 6] and confidence-based scoring [29] could potentially enhance selection accuracy and, consequently, the VLM's downstream task performance.

https://doi.org/10.52202/079017-3171

7.2 Conclusion

There are three main contributions to this work. First, we present BrainMD, a large-scale medical dataset with multiple modalities, comprising health records of 2,453 high-quality MRIs from 561 patients, complete with radiology reports, and structured data from medical records. Second, we propose a novel one-pass unsupervised representative slice selection method, Vote-MI, to select representative 2D slices from 3D volumetric data, bridging the gap between current vision-language models and their application to medical images. Third, we use this dataset to create two benchmarks, BrainMD-select and BrainBench. Using these benchmarks, we conducted in-depth studies on our proposed Vote-MI method. In terms of task performance, Vote-MI significantly improves performance across three diverse tasks. In conclusion, this work has laid the foundation for future research into representative slice selection methods for analyzing 3D medical imaging data with VLMs that only take 2D input. By openly sharing BrainMD, we hope to spark new advances in this critical area of healthcare.

Acknowledgments and Disclosure of Funding

This publication was made possible by the Johns Hopkins Institute for Clinical and Translational Research (ICTR), which is funded in part by Grant Number 1UM1TR004926-01 from the National Center for Advancing Translational Sciences (NCATS) a component of the National Institutes of Health (NIH), and NIH Roadmap for Medical Research. Its contents are solely the responsibility of the authors and do not necessarily represent the official view of the Johns Hopkins ICTR, NCATS or NIH.

https://doi.org/10.52202/079017-3171

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[2] 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.

[3] Christophe Andrieu and Johannes Thoms. A tutorial on adaptive mcmc. *Statistics and computing*, 18:343–373, 2008.

[4] Anmol Arora, Joseph E Alderman, Joanne Palmer, Shaswath Ganapathi, Elinor Laws, Melissa D McCradden, Lauren Oakden-Rayner, Stephen R Pfohl, Marzyeh Ghassemi, Francis McKay, et al. The value of standards for health datasets in artificial intelligence-based applications. *Nature Medicine*, 29(11):2929–2938, 2023.

[5] Fan Bai, Yuxin Du, Tiejun Huang, Max Q-H Meng, and Bo Zhao. M3d: Advancing 3d medical image analysis with multi-modal large language models. *arXiv preprint arXiv:2404.00578*, 2024.

[6] Slobodan Beliga, Ana Meštrović, and Sanda Martinčić-Ipšić. An overview of graph-based keyword extraction methods and approaches. *Journal of information and organizational sciences*, 39(1):1–20, 2015.

[7] Sartaj Bhuvaji, Ankita Kadam, Prajakta Bhumkar, Sameer Dedge, and Swati Kanchan. Brain tumor classification (mri), 2020.

[8] Louis Blankemeier, Joseph Paul Cohen, Ashwin Kumar, Dave Van Veen, Syed Jamal Safdar Gardezi, Magdalini Paschali, Zhihong Chen, Jean-Benoit Delbrouck, Eduardo Reis, Cesar Truyts, et al. Merlin: A vision language foundation model for 3d computed tomography. *arXiv preprint arXiv:2406.06512*, 2024.

[9] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

[10] Jun Cheng. Brain tumor dataset. https://doi.org/10.6084/m9.figshare.1512427.v5, 2017. DOI: https://doi.org/10.6084/m9.figshare.1512427.v5.

[11] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113, 2023.

[12] Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al. Palm-e: An embodied multimodal language model. *arXiv preprint arXiv:2303.03378*, 2023.

[13] Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. Glm: General language model pretraining with autoregressive blank infilling. *arXiv preprint arXiv:2103.10360*, 2021.

[14] Grant Duffy, Paul P Cheng, Neal Yuan, Bryan He, Alan C Kwan, Matthew J Shun-Shin, Kevin M Alexander, Joseph Ebinger, Matthew P Lungren, Florian Rader, et al. High-throughput precision phenotyping of left ventricular hypertrophy with cardiovascular deep learning. *JAMA cardiology*, 7(4):386–395, 2022.

[15] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *kdd*, volume 96, pages 226–231, 1996.

[16] Dan Feldman. Introduction to core-sets: an updated survey. *arXiv preprint arXiv:2011.09384*, 2020.

[17] Mingchen Gao, Junzhou Huang, Xiaolei Huang, Shaoting Zhang, and Dimitris N Metaxas. Simplified labeling process for medical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2012: 15th International Conference, Nice, France, October 1-5, 2012, Proceedings, Part II 15*, pages 387–394. Springer, 2012.

[18] Yuan Gu, Mingyue Wang, Yishu Gong, Xin Li, Ziyang Wang, Yuli Wang, Song Jiang, Dan Zhang, and Chen Li. Unveiling breast cancer risk profiles: a survival clustering analysis empowered by an online web application. *Future Oncology*, 19(40):2651–2667, 2023.

[19] Ibrahim Ethem Hamamci, Sezgin Er, and Bjoern Menze. Ct2rep: Automated radiology report generation for 3d medical imaging. *arXiv preprint arXiv:2403.06801*, 2024.

[20] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020.

[21] Dorit S Hochbaum and David B Shmoys. A best possible heuristic for the k-center problem. *Mathematics of operations research*, 10(2):180–184, 1985.

[22] Stephen James, Michael Bloesch, and Andrew J Davison. Task-embedded control networks for few-shot imitation learning. In *Conference on robot learning*, pages 783–795. PMLR, 2018.

[23] Hans Janssen. Monte-carlo based uncertainty analysis: Sampling efficiency and sampling convergence. *Reliability Engineering & System Safety*, 109:123–132, 2013.

[24] Pichai Kankuekul, Aram Kawewong, Sirinart Tangruamsub, and Osamu Hasegawa. Online incremental attribute-based zero-shot learning. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3657–3664. IEEE, 2012.

[25] Anahita Fathi Kazerooni, Nastaran Khalili, Xinyang Liu, Debanjan Haldar, Zhifan Jiang, Syed Muhammed Anwar, Jake Albrecht, Maruf Adewole, Udunna Anazodo, Hannah Anderson, et al. The brain tumor segmentation (brats) challenge 2023: Focus on pediatrics (cbtn-connect-dipgr-asnr-miccai brats-peds). *ArXiv*, 2023.

[26] Brenden M Lake, Tal Linzen, and Marco Baroni. Human few-shot learning of compositional instructions. *arXiv preprint arXiv:1901.04587*, 2019.

[27] Philippe Lambin, Ralph TH Leijenaar, Timo M Deist, Jurgen Peerlings, Evelyn EC De Jong, Janita Van Timmeren, Sebastian Sanduleanu, Ruben THM Larue, Aniek JG Even, Arthur Jochems, et al. Radiomics: the bridge between medical imaging and personalized medicine. *Nature reviews Clinical oncology*, 14(12):749–762, 2017.

[28] Oran Lang, Doron Yaya-Stupp, Ilana Traynis, Heather Cole-Lewis, Chloe R Bennett, Courtney R Lyles, Charles Lau, Michal Irani, Christopher Semturs, Dale R Webster, et al. Using generative ai to investigate medical imagery models and datasets. *EBioMedicine*, 102, 2024.

[29] David D Lewis. A sequential algorithm for training text classifiers: Corrigendum and additional data. In *Acm Sigir Forum*, volume 29, pages 13–19. ACM New York, NY, USA, 1995.

[30] Alexander C Li, Mihir Prabhudesai, Shivam Duggal, Ellis Brown, and Deepak Pathak. Your diffusion model is secretly a zero-shot classifier. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2206–2217, 2023.

[31] Chunyuan Li, Zhe Gan, Zhengyuan Yang, Jianwei Yang, Linjie Li, Lijuan Wang, Jianfeng Gao, et al. Multimodal foundation models: From specialists to general-purpose assistants. *Foundations and Trends® in Computer Graphics and Vision*, 16(1-2):1–214, 2024.

[32] 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.

[33] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023.

[34] Xiaoxu Li, Zhuo Sun, Jing-Hao Xue, and Zhanyu Ma. A concise review of recent few-shot meta-learning methods. *Neurocomputing*, 456:463–468, 2021.

[35] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024.

[36] Shuai Lu, Weihang Zhang, Jia Guo, Hanruo Liu, Huiqi Li, and Ningli Wang. Patchcl-ae: Anomaly detection for medical images using patch-wise contrastive learning-based auto-encoder. *Computerized Medical Imaging and Graphics*, 114:102366, 2024.

[37] Michael Moor, Qian Huang, Shirley Wu, Michihiro Yasunaga, Yash Dalmia, Jure Leskovec, Cyril Zakka, Eduardo Pontes Reis, and Pranav Rajpurkar. Med-flamingo: a multimodal medical few-shot learner. In *Machine Learning for Health (ML4H)*, pages 353–367. PMLR, 2023.

 https://doi.org/10.52202/079017-3171

[38] Beatriz Ocaña-Tienda, Julián Pérez-Beteta, José D Villanueva-García, José A Romero-Rosales, David Molina-García, Yannick Suter, Beatriz Asenjo, David Albillo, Ana Ortiz de Mendivil, Luis A Pérez-Romasanta, et al. A comprehensive dataset of annotated brain metastasis mr images with clinical and radiomic data. *Scientific data*, 10(1):208, 2023.

[39] Lucas Pinheiro Cinelli, Matheus Araújo Marins, Eduardo Antúnio Barros da Silva, and Sérgio Lima Netto. Variational autoencoder. In *Variational Methods for Machine Learning with Applications to Deep Networks*, pages 111–149. Springer, 2021.

[40] 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*, pages 8748–8763. PMLR, 2021.

[41] Pranav Rajpurkar, Jeremy Irvin, Aarti Bagul, Daisy Ding, Tony Duan, Hershel Mehta, Brandon Yang, Kaylie Zhu, Dillon Laird, Robyn L Ball, et al. Mura: Large dataset for abnormality detection in musculoskeletal radiographs. *arXiv preprint arXiv:1712.06957*, 2017.

[42] Mengye Ren, Eleni Triantafillou, Sachin Ravi, Jake Snell, Kevin Swersky, Joshua B Tenenbaum, Hugo Larochelle, and Richard S Zemel. Meta-learning for semi-supervised few-shot classification. *arXiv preprint arXiv:1803.00676*, 2018.

[43] Konstantinos I Roumeliotis and Nikolaos D Tselikas. Chatgpt and open-ai models: A preliminary review. *Future Internet*, 15(6):192, 2023.

[44] Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Kevin Clark, Stephen Pfohl, Heather Cole-Lewis, Darlene Neal, et al. Towards expert-level medical question answering with large language models. *arXiv preprint arXiv:2305.09617*, 2023.

[45] Yang Song, Liyue Shen, Lei Xing, and Stefano Ermon. Solving inverse problems in medical imaging with score-based generative models. *arXiv preprint arXiv:2111.08005*, 2021.

[46] Hongjin Su, Jungo Kasai, Chen Henry Wu, Weijia Shi, Tianlu Wang, Jiayi Xin, Rui Zhang, Mari Ostendorf, Luke Zettlemoyer, Noah A Smith, et al. Selective annotation makes language models better few-shot learners. *arXiv preprint arXiv:2209.01975*, 2022.

[47] Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. Eva-clip: Improved training techniques for clip at scale. *arXiv preprint arXiv:2303.15389*, 2023.

[48] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[49] Tao Tu, Shekoofeh Azizi, Danny Driess, Mike Schaekermann, Mohamed Amin, Pi-Chuan Chang, Andrew Carroll, Charles Lau, Ryutaro Tanno, Ira Ktena, et al. Towards generalist biomedical ai. *NEJM AI*, 1(3):AIoa2300138, 2024.

[50] Chenguang Wang, Ruoxi Jia, Xin Liu, and Dawn Song. Benchmarking zero-shot robustness of multimodal foundation models: A pilot study. *arXiv preprint arXiv:2403.10499*, 2024.

[51] Yaqing Wang, Quanming Yao, James T Kwok, and Lionel M Ni. Generalizing from a few examples: A survey on few-shot learning. *ACM computing surveys (csur)*, 53(3):1–34, 2020.

[52] Yibin Wang, William Neil Duggar, David Michael Caballero, Toms Vengaloor Thomas, Neha Adari, Eswara Kumar Mundra, and Haifeng Wang. A brain mri dataset and baseline evaluations for tumor recurrence prediction after gamma knife radiotherapy. *Scientific Data*, 10(1):785, 2023.

[53] Yuli Wang, Wen-Chi Hsu, Victoria Shi, Gigin Lin, Cheng Ting Lin, Xue Feng, and Harrison Bai. Cardcros: A dataset and benchmark for enhancing cardiovascular artery segmentation through disconnected components repair and open curve snake. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 179–189. Springer, 2024.

[54] Yuli Wang and Ji Yi. Deep learning-based image registration method: with application to scanning laser ophthalmoscopy (slo) longitudinal images. In *Medical Imaging 2023: Image Processing*, volume 12464, pages 614–618. SPIE, 2023.

[55] Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. Towards generalist foundation model for radiology. *arXiv preprint arXiv:2308.02463*, 2023.

[56] Junde Wu, Rao Fu, Huihui Fang, Yu Zhang, Yehui Yang, Haoyi Xiong, Huiying Liu, and Yanwu Xu. Medsegdiff: Medical image segmentation with diffusion probabilistic model. In *Medical Imaging with Deep Learning*, pages 1623–1639. PMLR, 2024.

[57] Yongqin Xian, Bernt Schiele, and Zeynep Akata. Zero-shot learning-the good, the bad and the ugly. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4582–4591, 2017.

[58] Lin Yang, Shawn Xu, Andrew Sellergren, Timo Kohlberger, Yuchen Zhou, Ira Ktena, Atilla Kiraly, Faruk Ahmed, Farhad Hormozdiari, Tiam Jaroensri, et al. Advancing multimodal medical capabilities of gemini. *arXiv preprint arXiv:2405.03162*, 2024.

[59] Li Zhang, Tao Xiang, and Shaogang Gong. Learning a deep embedding model for zero-shot learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2021–2030, 2017.

[60] Linmei Zhao, Maliha Imami, Yuli Wang, Yitao Mao, Wen-Chi Hsu, Ruohua Chen, Esther Mena, Yang Li, Jingyi Tang, Jing Wu, et al. Deep learning-based lesion characterization and outcome prediction of prostate cancer on [18 f] dcfpyl psma imaging. 2024.

[61] Peng Zhou, Zheng Liu, Hemmings Wu, Yuli Wang, Yong Lei, and Shiva Abbaszadeh. Automatically detecting bregma and lambda points in rodent skull anatomy images. *PloS one*, 15(12):e0244378, 2020.

[62] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023.

 https://doi.org/10.52202/079017-3171

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: In abstract e claim that we developed Vote-MI, an innovative, unsupervised method for selecting representative 2D slices from 3D medical imaging, and introduced BrainMD, a multimodal dataset to evaluate its effectiveness with VLMs. Using BrainMD, we also developed two benchmarks: BrainMD-select and BrainBench. All these three items accurately reflect the paper's contributions and scope.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: A section discussing the work's limitations from both methodological and clinical perspectives is included in the paper.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [No]

   Justification: This paper focuses on the application, where we deliver a multimodal dataset and a method for representative slice selection; therefore, there are no theoretical assumptions or proofs.

   Guidelines:

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The reproduction details are summarized in the main paper, appendix, and open-source code.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: All the code in this paper is open-source, and the data is accessible upon request to the senior authors.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: All the training and testing details are specified in the main paper, appendix, and open-source code.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

Justification: All results are accompanied by error bars, confidence intervals, or statistical significance tests.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We indicate required experiments compute resources and information in the paper.

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: Code of Ethics is confirmed.

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The positive societal impacts are discussed in the paper, and at the current stage, the authors do not believe there are any negative societal impacts.

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [Yes]

    Justification: The dataset and benchmark developed in this paper have been manually de-identified, with all patient-related information removed.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All data, code, and models are properly cited or explicitly mentioned.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The assets in the paper are well documented and are the documentation provided.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [Yes]

    Justification: Required information is included in the supplementary materials.

    Guidelines:

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

https://doi.org/10.52202/079017-3171

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [Yes]

Justification: Both IRB approval and Data Use agreement are obtained for the data we used in this paper.