Label Critic: DESIGN DATA BEFORE MODELS

Pedro R. A. S. Bassi[1,2,3], *Qilong Wu*[1,4], *Wenxuan Li*[1],
Sergio Decherchi[2], *Andrea Cavalli*[2,3,5], *Alan Yuille*[1], *Zongwei Zhou*[1,*]

[1]Johns Hopkins University [2]Italian Institute of Technology [3]University of Bologna
[4]National University of Singapore [5]École Polytechnique Fédérale de Lausanne

Code & Data: https://github.com/PedroRASB/LabelCritic

ABSTRACT

As medical datasets rapidly expand, creating detailed annotations of different body structures becomes increasingly expensive and time-consuming. We consider that requesting radiologists to create detailed annotations is unnecessarily burdensome and that pre-existing AI models can largely automate this process. Following the spirit *don't use a sledgehammer on a nut*, we find that, rather than creating annotations from scratch, radiologists only have to review and edit errors if the Best-AI Labels have mistakes. To obtain the Best-AI Labels among multiple AI Labels, we developed an automatic tool, called **Label Critic**, that can assess label quality through tireless pairwise comparisons. Extensive experiments demonstrate that, when incorporated with our developed Image-Prompt pairs, pre-existing Large Vision-Language Models (LVLM), trained on natural images and texts, achieve 96.5% accuracy when choosing the best label in a pair-wise comparison, without extra fine-tuning. By transforming the manual annotation task (30–60 min/scan) into an automatic comparison task (15 sec/scan), we effectively reduce the manual efforts required from radiologists by an order of magnitude. When the Best-AI Labels are sufficiently accurate (81% depending on body structures), they will be directly adopted as the gold-standard annotations for the dataset, with lower-quality AI Labels automatically discarded. Label Critic can also check the label quality of a single AI Label with 71.8% accuracy when no alternatives are available for comparison, prompting radiologists to review and edit if the estimated quality is low (19% depending on body structures).

1. INTRODUCTION

Publicly available abdominal CT datasets with per-voxel annotations have experienced rapid growth in recent years. In 2020, datasets like KiTS [1] and LiTS [2] offered a few hundred annotated CT scans. By 2023, datasets such as AbdomenAtlas [3] and FLARE [4] expanded these scans significantly, now exceeding 10,000 annotated scans. This growth is enabled by AI-assisted annotation, where AI performs the initial segmentation and radiologists review and edit errors made by AI [5, 6]. Despite AI assistance, the current scale—now with tens of thousands of annotations per dataset [6]—has made manual detection and editing of label errors increasingly impractical. This raises the question: *Rather than having radiologists detect and edit AI errors, can we—again—use AI to automate these tasks and scale medical datasets?*

Automatic error detection is achievable most label errors in existing datasets because, simply put, *critiquing is easier than creating*. This paper builds on two main insights. **First**, most errors made by AI are easy to detect[1] and do not require the time and expertise of busy, costly radiologists. **Second**, when multiple labels are available[2], comparing them to identify the highest-quality label is even simpler.

We discover that general-purpose Large Vision Language Models (LVLMs), like Llava and GPT-4V [11, 12], trained on massive text-image datasets, can detect errors in medical datasets and compare the label quality among multiple label options *without* additional fine-tuning. We present a new LVLM-based pipeline, **Label Critic**, which can effectively (1) detect a large portion (76.8%) of the obvious label errors in existing medical datasets and (2) select the Best-AI Label by comparing multiple AI Labels.

We show that Label Critic can generalize to over 10,000 CT scans across 89 hospitals with minimal or no training data (≤ 10). It detects 1,441 errors in the datasets, with overall accuracy of 96.5% in detecting label errors and identify the Best-AI Labels in a pair-wise comparison. The success of Label Critic is attributed to our innovative **Input** and **Prompt** designs specialized for 3D CT scans and the integration of

* Correspondence to: Zongwei Zhou (zzhou82@jh.edu)

[1]A common AI error in abdominal CT scans is mislabeling the aortic arch. This error is obvious, as the aorta should appear curved in its top, forming an arch (see Fig. 1). Even non-experts can easily recognize such errors due to the aorta's consistent size, position, and appearance across scans.

[2]The number of public AI models quickly raises [7]. Medical segmentation benchmarks provide diverse datasets, where participants train different architectures, providing a variety of labels for Label Critic to choose from. E.g., for abdominal organ segmentation in CT, we easily find solutions to the FLARE challenge [8, 9, 10], 11 models trained on AbdomenAtlas are already public, and more will be released after Touchstone Benchmark [7].

prior knowledge about body structures.

First, we design new inputs for LVLMs. Since most LVLMs are designed for 2D inputs, Label Critic uses 2D frontal projections of CT scans with transparent overlays of label projections, ensuring computational efficiency while preserving key volumetric information (§2.2). The projections resemble antero-posterior (AP) X-rays, making them familiar to general-purpose LVLMs.

Second, we design new prompts for LVLMs. They incorporate step-by-step guidance, anatomical descriptions, Dual Confirmation, and variable examples ranging from zero-shot to in-context learning with up to 10 label examples (§2.3). This flexibility enables Label Critic to adapt quickly to new hospitals and segmentation classes, requiring few or no training samples, while avoiding overfitting to specific label error types (§2.3, §2.1). This is the first work to show LVLMs can compare semantic segmentations, using prior knowledge to choose the best AI model for each case and class.

Related Work. Label quality control methods identify potential label errors by flagging uncertainty and inconsistency across AI models [3, 13], but they do not specify which label is better, leaving radiologists to review each flagged case manually. In our dataset, this approach requires manual review of 4,348 labels across two AI models, a time-intensive task. Most existing QC methods are organ-specific (e.g., cardiac or muscle imaging [14, 15]), limiting their scalability to other body structures. There is no prior methods leverage large vision-language models (LVLMs) for label quality control. Our LVLM-based method can significantly reduce manual workload for multi-organ segmentation[3] by comparing and selecting the best labels, discarding incorrect ones, and flagging only the most challenging cases for further manual review, efficiently streamlining the process.

2. METHODOLOGY

As shown in Fig. 1, Label Critic includes projecting the CT scan and labels into 2D, calculating the Dice Similarity Coefficient (DSC) between labels, and prompting a LVLM to select the most accurate label. If the DSC is below a class-specific threshold[4] comparison is skipped, saving computational resources. The DSC check skips comparisons of labels with minor differences, focusing instead on substantial errors detectable through basic anatomical knowledge. When alternative labels or public segmentation models are unavailable, Label Critic assumes a non-comparative approach: it projects the CT with its single label and asks the LVLM to verify its anatomical accuracy, optionally using other CTs and labels as in-context examples.

[3]Spleen, gallbladder, pancreas, postcava, aorta, kidneys, spleen, and liver.

[4]The class-specific thresholds are set at around the dataset's average class DSC minus one standard deviation, identifying outliers. Lower thresholds will lead to more comparisons and a larger computational cost.

2.1. LVLM Architecture and (no) Training

Training AI for label error detection requires a dataset identifying both correct and incorrect labels, but assembling it is challenging: small medical datasets contain few errors, and finding errors in large AI-labeled datasets is labor-intensive–hence the need for automatic error detection. Although synthetic error generation is possible, training on either real or synthetic errors risks shortcut learning: models concentrating on the specific error types in the training dataset and failing to generalize well to unseen types [16]. To address these issues and enable broad adaptability across hospitals, we leverage zero-shot and few-shot learning. Given the limited per-voxel annotations in the training data of large vision-language models, robust out-of-distribution generalization is essential for our pipeline. We experimented with several LVLMs, selecting Qwen2-VL [11]—a large general-purpose model with 70 billion parameters and AWQ quantization for speed—because it can analyze multiple images per prompt, unlike alternatives such as LLaVA-7B [12], LLaVA-Med [17], and M3D-4B [18], which also yielded lower performance (see Tab. 1). Proprietary models like GPT-4V were not considered due to high API costs for processing large volumes of images.

2.2. Projections and Overlays

The usual 2D representation of CT volumes consists in their 2D slices. However, slices only show a small portion of the scan, and multiple slices would be needed to represent an annotation. Conversely, projections show through the entire body, conveying the entire CT and annotation in a single image. They cannot capture all possible label errors, such as holes inside annotations. However, they are a cost-effective solution: transformers' computational cost increases quadratically with input length, hindering the use of many CT slices as input. Antero-posterior projections of CT scans resemble AP X-rays, making them familiar and interpretable to pre-trained LVLMs. E.g., asked to describe the projections in Fig. 1, GPT-4V says *"a frontal X-ray-like projection"* or "*a frontal projection of a CT scan*". Alg. 1 describes the projection procedure. The algorithm is designed for simplicity and computational efficiency. Instead, advanced X-ray simulation methods could be employed to project CT scans, but this would increase Label Critic's computational cost.

Algorithm 1 2D Projection of a 3D CT Scan

1: **Threshold:** Limit HU values to $[-500, 1500]$—a window that makes projections more X-ray-like.
2: **Project in 2D:** Sum over the antero-posterior axis.
3: **Normalize:** Apply $x_i = (x_i - x_{\min})/(x_{\max} - x_{\min})$.
4: **Resize & RGB:** Resize to 512 p on the longest image side, keeping aspect ratio, and replicate in 3 channels (RGB).

To project the label and overlay it over the CT projection, we first repeat Alg. 1 steps 1 and 2 for the label. Then, we zero

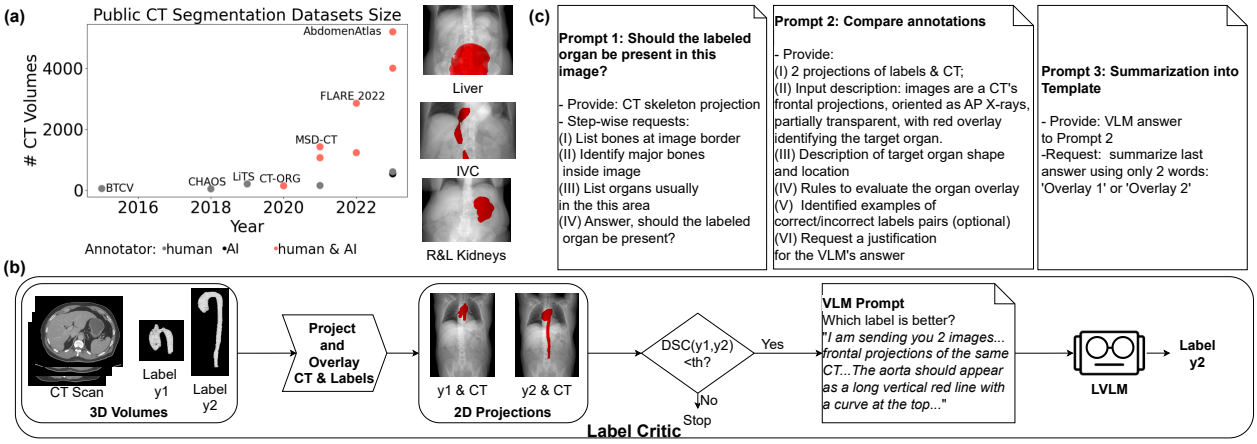

Fig. 1. **(a) Public CT datasets** with per-voxel labels are rapidly expanding, largely due to AI-assisted labeling. However, AI often makes obvious errors, exampled in the liver, IVC, and kidneys, highlighting the need for efficient, automated error detection. **(b) Label Critic pipeline for comparing labels.** (I) Frontally *project* (§2.2) the CT scan and overlay it with the projections of two candidate labels, y_1 and y_2 (red), creating two images; (II) *verify the dice* score (DSC) between the 2 label projections, skip the comparison if DSC is above a class-specific threshold—avoiding comparing very similar labels; (III) ask a *LVLM* (§2.1) to compare the labels and choose the most correct. If y_1 is a dataset label we are evaluating, we consider it wrong if the LVLM prefers y_2, the output of an alternative public segmentation model. **(c) 3-Step Prompt Design.** *Prompt 1* asks if the target organ should be in the CT, providing a skeleton projection as reference. If the LVLM says no, we select an empty label (if available) or flag the case for review. Otherwise, *Prompt 2* asks the LVLM to compare two label overlays using class-aware prompts with anatomical guidance, optional in-context learning, and complexity based on the LVLM's background knowledge of each class (§2.3). *Prompt 3* asks the LVLM to summarizes its previous answer. Summarization provides an easily processable binary answer, but allows detailed justifications and step-by-step reasoning in earlier steps.

the blue and green channels of the CT projection where the label projection is not 0, creating a semi-transparent red overlay that doesn't obscure the CT. Also, we create skeleton projections to help the LVLM identifying missing or misplaced labels (see §2.3). To create them, we use a window of [400, 2000] in Alg. 1 and enhance the projection's contrast with CLAHE (grid 8, clip 5) and gamma adjustment ($\gamma = 0.6$). For M3D, we provide 3D CTs with labels overlaid in black (lowest CT HU value), as preliminary tests showed this color outperformed white or gray overlays, possibly due to its more natural look inside CTs.

2.3. *Prompt! Prompt! Prompt!*

Prompt design impacts accuracy (§3). In the large (N=5,195) AtlasBench CT dataset (§3), we iteratively created a prompt, ran Label Critic, analyzed wrong LVLM answers, and improved the prompt accordingly. This process led to a standardized 3-step prompt, detailed in Fig. 1. Step 2, which requests the LVLM to compare two labels, is class dependent. Prompt complexity, strictness and number of in-context examples (from 0-10) depends on the LVLM background knowledge of the class: liver, spleen, kidneys and pancreas are classes the LVLM is more familiar with, allowing more complex and less strict prompts, with abstract shape references (e.g., "wedge-like") and multiple anatomical landmarks (e.g.,"below the diaphragm"); VLMs are less familiar to aorta and postcava, and our prompts used simple anatomical descriptions and strict guidance, focusing on linear shape, extension, and continuity; stomach and gallbladder have less well-defined shape, and our prompt focus on label location and gross shape errors. For stomach we use in-context learning, providing one example of correct label.

We repeat Prompts 2 and 3 (Fig. 1), inverting the image order in the LVLM input, and we check if its answers are consistent across the repetitions. This procedure, dubbed Dual Confirmation, reveals unreliable LVLM answers for minute label errors or cases where both labels are wrong. Also, we observed the LVLM itself can reject these comparisons, saying both labels are bad or similar. If Dual Confirmation finds inconsistent answers or the VLM rejects comparisons, we remove the case from the dataset, flagging for human review. To detect errors without label comparison, we skip the dice check and Dual Confirmation and modify prompt 2 to ask the LVLM to evaluate a single label, giving it examples of other CTs and correct/incorrect labels. We also created class-agnostic 3-step prompts, readily applicable to new classes, by removing class information in Prompt 2. Prompts were summarized for Llava, Llava-med and M3D, due to smaller context length. All prompts are available in our code.

Table 1. **Label Critic excels in two datasets.** We report Accuracy as the proportion of labels correctly evaluated out of the total evaluated. Each class contains an equal number of correct and incorrect labels. The LVLM used here is Qwen2-VL [19]; we also tested Llava [12], Llava-Med [17], and M3D [18], but these alternatives performed poorly, with average Accuracies of 54.1%, 50.2%, and 49.4%, respectively, for error detection on AtlasBench.

		AtlasBench (error detection)								
prompt	in-context	aorta	gallbladder	kidneys	liver	pancreas	postcava	spleen	stomach	average
class-agnostic	0-shot	51.0 (530/1040)	50.0 (59/118)	84.9 (107/126)	55.6 (10/18)	63.2 (72/114)	0.0 (0/2)	40.0 (8/20)	66.7 (8/12)	54.8 (794/1450)
class-aware	0-shot	58.7 (610/1040)	50.8 (60/118)	89.7 (113/126)	83.3 (15/18)	85.1 (97/114)	50.0 (1/2)	80.0 (16/20)	50.0 (6/12)	63.3 (918/1450)
	1-shot	63.9 (665/1040)	50.8 (60/118)	83.3 (105/126)	83.3 (15/18)	76.3 (87/114)	100.0 (2/2)	70.0 (14/20)	50.0 (6/12)	65.8 (954/1450)
	10-shot	72.2 (751/1040)	50.8 (60/118)	77.0 (97/126)	83.3 (15/18)	80.7 (92/114)	100.0 (2/2)	75.0 (15/20)	75.0 (9/12)	71.8 (1041/1450)
		AtlasBench (label comparison)								
class-agnostic	0-shot	78.7 (546/694)	68.0 (34/50)	95.7 (90/94)	100.0 (14/14)	97.1 (68/70)	- (0/0)	100.0 (12/12)	100.0 (2/2)	81.8 (766/936)
class-aware	0-shot	**96.5** (440/456)	**74.4** (58/78)	**96.4** (106/110)	**100.0** (12/12)	**92.2** (94/102)	- (0/0)	**100.0** (12/12)	**66.7** (4/6)	**93.6** (726/776)
		JHHBench (label comparison)								
class-aware	0-shot	**98.4** (1234/1254)	**92.9** (340/366)	**85.7** (12/14)	**100.0** (62/62)	**100.0** (22/22)	**100.0** (346/346)	**100.0** (18/18)	**93.8** (122/130)	**97.5** (2156/2212)

3. RESULTS AND DISCUSSION

We created two datasets, AtlasBench and JHHBench, to evaluate Label Critic. They contain errors from real public and private datasets, including mistakes in AI and human labels. As a ground truth, labels in AtlasBench and JHHBench were manually deemed correct or incorrect. Both dataset have labels for eight abdominal organs[3].

AtlasBench: The public AbdomenAtlas dataset [6], annotated by AI-assisted radiologists, includes 5,195 abdominal CT volumes from 88 hospitals worldwide. We used Label Critic to compare an intermediate development version of AbdomenAtlas (Beta) to the current release (1.0). Label Critic's DSC check (Fig. 1) selected 1,450 labels with low DSC, finding labels that were updated from Beta to 1.0, potentially due to errors. We dubbed this subset AtlasBench. We have released it as the first public dataset specifically for benchmarking error detection and label comparison methods.

JHHBench: JHH consists of 5,172 CT volumes from Johns Hopkins Hospital, annotated manually by radiologists. To construct JHHBench, we compared these with pseudo-annotations from a public nnU-Net ResEncL trained on AbdomenAtlas during the Touchstone Benchmark [7]. Here, the Label Critic's DSC check selected 2,808 low-DSC labels.

Label Critic was accurate and generalized to many types of label errors. For pair-wise comparison, it correctly chose the best label 97.5% of the time in JHHBench, and 93.5% in AtlasBench (Tab. 1). Most labels deemed wrong were AI segmentation errors, but Label Critic found 188 errors in human-made labels–133 aorta errors due to the aortic arch being out of the annotator's region of interest, and 55 label corruptions, like missing slices. We never trained the LVLMs in Label Critic, and its prompts were developed in AtlasBench considering AI errors only. Thus, finding errors in human annotations means strong out-of-distribution (OOD) generalization.

AtlasBench results show Label Critic's superior adaptability and accurate label comparisons. While in-context learning (10-shot) improved non-comparative error detection, it remained less accurate than the comparative Label Critic. Therefore, for detecting errors in a dataset, Label Critic's comparison of dataset labels to outputs from public segmentation models is preferable. Additionally, Label Critic's class-tailored prompts outperformed class-agnostic ones, but even the latter achieved 81.8% accuracy, demonstrating Label Critic's ability to adapt effectively to new classes.

A large, general-purpose LVLM (Qwen2-VL 70B [11]) surpassed smaller LVLMs (Llava-7B [12]) and medical LVLMs (Llava-Med [17] and M3D [18]). Qwen2-VL's advantages come from its larger size and longer context length (32,768 vs. 8,000 tokens), which improve reasoning, handling instructions, and processing larger prompts. In contrast, Llava-Med and M3D, LVLMs fine-tuned on smaller medical datasets (100K medical volumes for M3D vs. 400M images for CLIP), reached about chance-level accuracy in error detection (Tab. 1). Thus, they struggled to generalize to Label Critic's out-of-distribution tasks [20], as label error detection and comparison are uncommon tasks in their training data.

4. CONCLUSION

Label Critic was highly effective for detecting and comparing label errors in organ segmentation, choosing the best label with a high accuracy of 97.5% on JHHBench and 93.5% on AtlasBench. This study is the first to show LVLMs can compare segmentation outputs and automatically select the Best-AI label for each sample, finding and discarding errors, and minimizing manual label revision. Label Critic generalizes to many error types (in AI and human labels), and easily adapts to new hospitals and classes. Thus, it can help dataset creators improve label quality in massive medical datasets, and model creators improve data before training AI. In future work, we plan to extend Label Critic to the tumor class.

Acknowledgments. This work was supported by the Lustgarten Foundation for Pancreatic Cancer Research, the McGovern Foundation, and Istituto Italiano di Tecnologia (73010, Arnesano, LE, Italy; 16163, Genova, GE, Italy).

5. REFERENCES

[1] Nicholas Heller, Niranjan Sathianathen, Arveen Kalapara, Edward Walczak, Keenan Moore, Heather Kaluzniak, Joel Rosenberg, Paul Blake, Zachary Rengel, Makinna Oestreich, et al., "The kits19 challenge data: 300 kidney tumor cases with clinical context, ct semantic segmentations, and surgical outcomes," *arXiv preprint arXiv:1904.00445*, 2019.

[2] Patrick Bilic, Patrick Ferdinand Christ, Eugene Vorontsov, Grzegorz Chlebus, Hao Chen, Qi Dou, Chi-Wing Fu, Xiao Han, Pheng-Ann Heng, Jürgen Hesser, et al., "The liver tumor segmentation benchmark (lits)," *arXiv preprint arXiv:1901.04056*, 2019.

[3] Wenxuan Li, Alan Yuille, and Zongwei Zhou, "How well do supervised models transfer to 3d image segmentation?," in *International Conference on Learning Representations*, 2024.

[4] Jun Ma, Yao Zhang, Song Gu, Xingle An, Zhihe Wang, Cheng Ge, Congcong Wang, Fan Zhang, Yu Wang, Yinan Xu, et al., "Fast and low-gpu-memory abdomen ct organ segmentation: the flare challenge," *Medical Image Analysis*, vol. 82, pp. 102616, 2022.

[5] Jakob Wasserthal, Hanns-Christian Breit, Manfred T Meyer, Maurice Pradella, Daniel Hinck, Alexander W Sauter, Tobias Heye, Daniel T Boll, Joshy Cyriac, Shan Yang, et al., "Totalsegmentator: robust segmentation of 104 anatomic structures in ct images," *Radiology: Artificial Intelligence*, vol. 5, no. 5, 2023.

[6] Wenxuan Li, Chongyu Qu, Xiaoxi Chen, Pedro RAS Bassi, Yijia Shi, Yuxiang Lai, Qian Yu, Huimin Xue, Yixiong Chen, Xiaorui Lin, et al., "Abdomenatlas: A large-scale, detailed-annotated, & multi-center dataset for efficient transfer learning and open algorithmic benchmarking," *Medical Image Analysis*, p. 103285, 2024.

[7] Pedro RAS Bassi, Wenxuan Li, Yucheng Tang, Fabian Isensee, Zifu Wang, Jieneng Chen, Yu-Cheng Chou, Yannick Kirchhoff, Maximilian Rokuss, Ziyan Huang, Jin Ye, Junjun He, Tassilo Wald, Constantin Ulrich, Michael Baumgartner, Saikat Roy, Klaus H. Maier-Hein, Paul Jaeger, Yiwen Ye, Yutong Xie, Jianpeng Zhang, Ziyang Chen, Yong Xia, Zhaohu Xing, Lei Zhu, Yousef Sadegheih, Afshin Bozorgpour, Pratibha Kumari, Reza Azad, Dorit Merhof, Pengcheng Shi, Ting Ma, Yuxin Du, Fan Bai, Tiejun Huang, Bo Zhao, Haonan Wang, Xiaomeng Li, Hanxue Gu, Haoyu Dong, Jichen Yang, Maciej A. Mazurowski, Saumya Gupta, Linshan Wu, Jiaxin Zhuang, Hao Chen, Holger Roth, Daguang Xu, Matthew B. Blaschko, Sergio Decherchi, Andrea Cavalli, Alan L. Yuille, and Zongwei Zhou, "Touchstone benchmark: Are we on the right way for evaluating ai algorithms for medical segmentation?," *Conference on Neural Information Processing Systems*, 2024.

[8] Jie Liu, Yixiao Zhang, Jie-Neng Chen, Junfei Xiao, Yongyi Lu, Bennett A Landman, Yixuan Yuan, Alan Yuille, Yucheng Tang, and Zongwei Zhou, "Clip-driven universal model for organ segmentation and tumor detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 21152–21164.

[9] Ziyan Huang, Haoyu Wang, Jin Ye, Jingqi Niu, Can Tu, Yuncheng Yang, Shiyi Du, Zhongying Deng, Lixu Gu, and Junjun He, "Revisiting nnu-net for iterative pseudo labeling and efficient sliding window inference," in *Fast and Low-Resource Semi-supervised Abdominal Organ Segmentation: MICCAI 2022 Challenge, FLARE 2022, Held in Conjunction with MICCAI 2022, Singapore, September 22, 2022, Proceedings*, pp. 178–189. Springer, 2023.

[10] Jie Liu, Alan Yuille, Yucheng Tang, and Zongwei Zhou, "Clip-driven universal model for partially labeled organ and pan-cancer segmentation," in *MICCAI 2023 FLARE Challenge*, 2023.

[11] 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.

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

[13] Maria-Florina Balcan, Andrei Broder, and Tong Zhang, "Margin based active learning," in *International Conference on Computational Learning Theory*. Springer, 2007, pp. 35–50.

[14] Giacomo Tarroni, Ozan Oktay, Wenjia Bai, Andreas Schuh, Hideaki Suzuki, Jonathan Passerat-Palmbach, Antonio De Marvao, Declan P O'Regan, Stuart Cook, Ben Glocker, et al., "Learning-based quality control for cardiac mr images," *IEEE transactions on medical imaging*, vol. 38, no. 5, pp. 1127–1138, 2018.

[15] Fahim Ahmed Zaman, Lichun Zhang, Honghai Zhang, Milan Sonka, and Xiaodong Wu, "Segmentation quality assessment by automated detection of erroneous surface regions in medical images," *Computers in biology and medicine*, vol. 164, pp. 107324, 2023.

[16] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix Wichmann, "Shortcut learning in deep neural networks," *Nature Machine Intelligence*, vol. 2, pp. 665–673, 11 2020.

[17] 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," *arXiv preprint arXiv:2306.00890*, 2023.

[18] 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," 2024.

[19] 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*, 2023.

[20] Rheeya Uppaal, Junjie Hu, and Yixuan Li, "Is fine-tuning needed? pre-trained language models are near perfect for out-of-domain detection," 2023.