Label Critic: Design Data Before Models

Code & Paper

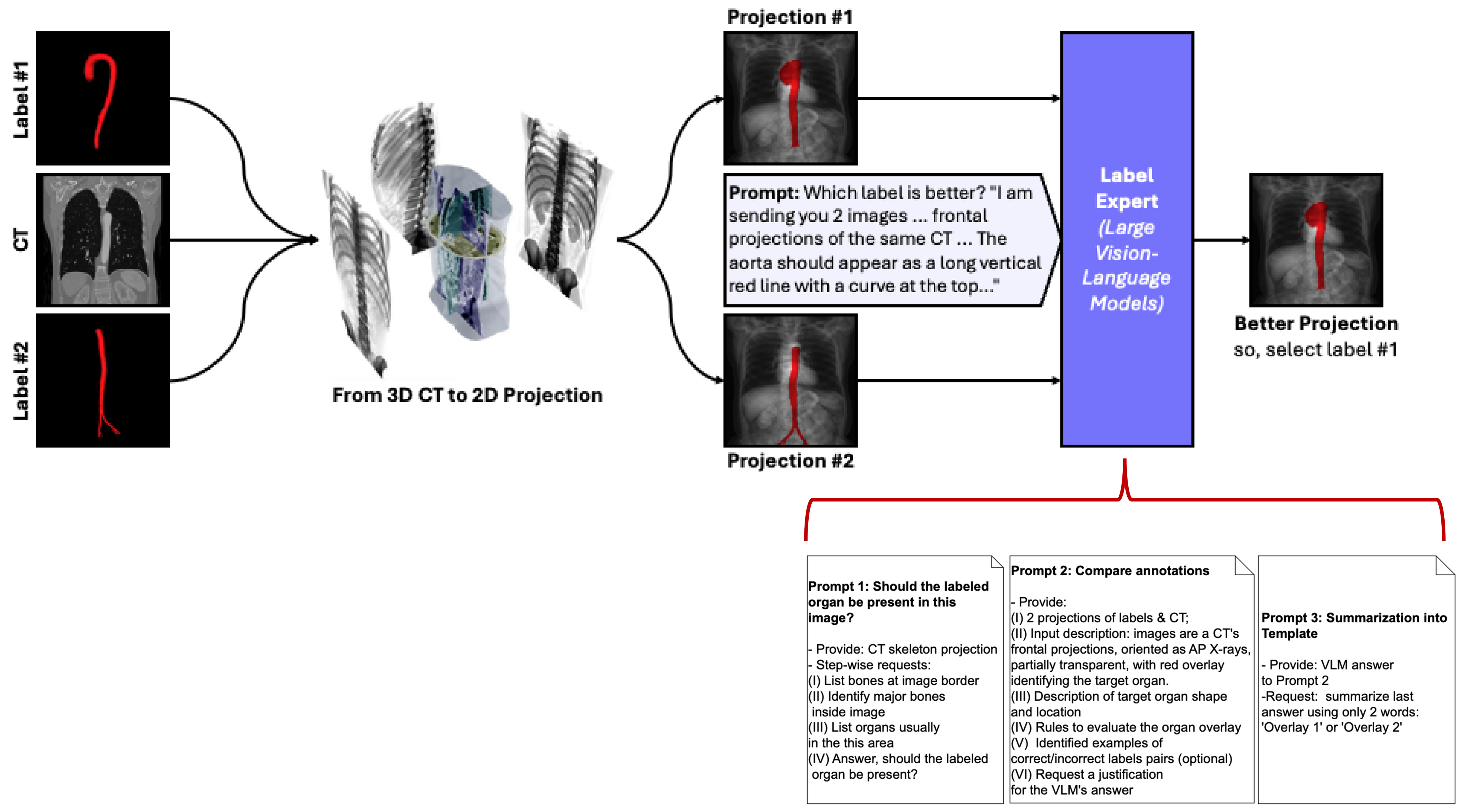
Pedro R. A. S. Bassi[1,2,3], Qilong Wu[1,4], Wenxuan Li[1], Sergio Decherchi[3], Andrea Cavalli[2,3,5], Alan Yuille[1], Zongwei Zhou[1]

[1]Johns Hopkins University, [2]University of Bologna, [3]IIT, [4]NUS, [5]EPFL

ISBI 2025
2025 IEEE International Symposium on Biomedical Imaging
April 14-17, 2025 | Houston, TX, USA

Poster No. 1571091591

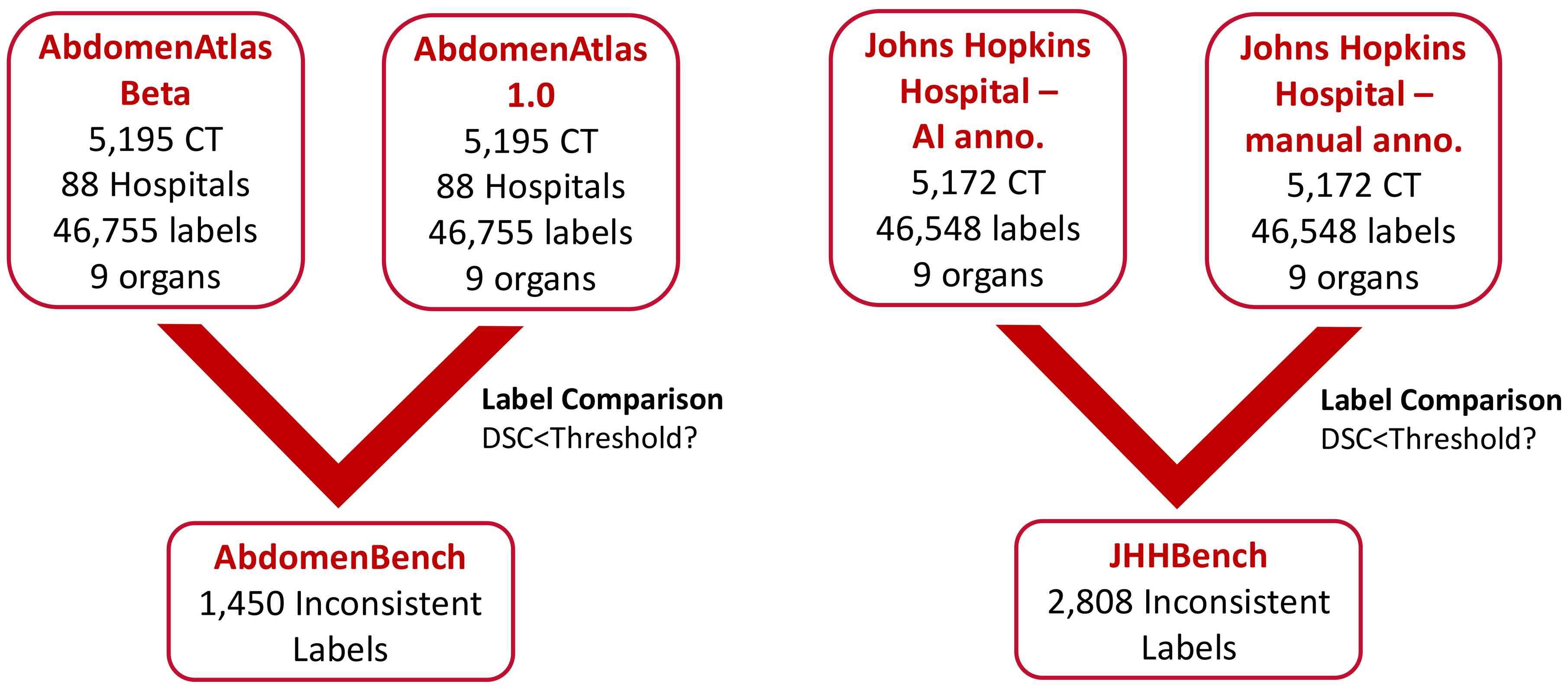
Method: Revising Medical Datasets with LVLMs



Label Critic pipeline for comparing labels.

1. Frontally **project** the CT scan and overlay it with the projections of two candidate labels (red), y1 and y2, creating two images
2. Verify **DSC** between the 2 projections, skip the comparison if DSC is too high—avoiding comparing overly similar labels
3. Ask a LVLM (Qwen2-VL) to **compare the labels** and choose the best
4. Dual Confirmation: LVLM can confidently choose the best label, or flag difficult cases for human review

Massive CT Datasets: AbdomenAtlas & JHH



Results

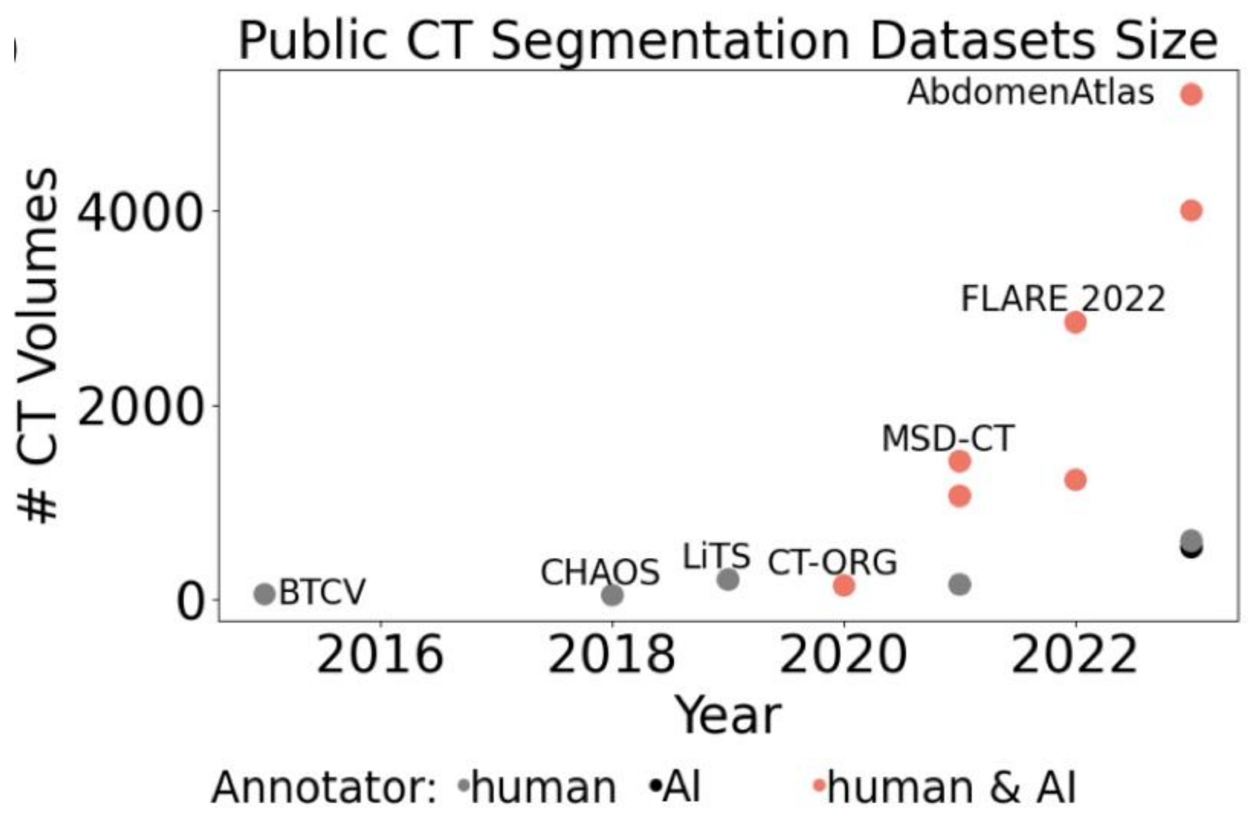
Table 1. Label Critic excels in two datasets. We report Accuracy as the proportion of labels correctly evaluated out of the total evaluated. Each class contains an equal number of correct and incorrect labels. The LVLM used here is Qwen2-VL [29]; we also tested Llava [16], Llava-Med [26], and M3D [27], but these alternatives performed poorly, with average Accuracies of 54.1%, 50.2%, and 49.4%, respectively, for error detection on AtlasBench.

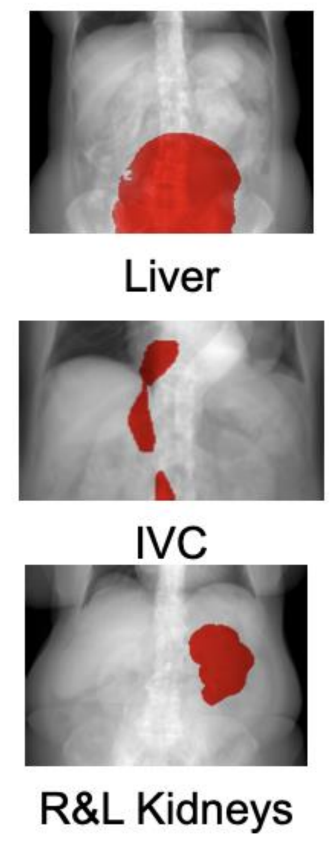
prompt	in-context	aorta	gallbladder	kidneys	liver	pancreas	postcava	spleen	stomach	average
		AtlasBench (error detection)								
class-agnostic	0-shot	51.0 (530/1040)	50.0 (59/118)	84.9 (107/126)	55.6 (10/18)	63.2 (72/114)	0.0 (0/2)	40.0 (8/20)	66.7 (8/12)	54.8 (794/1450)
class-aware	0-shot	58.7 (610/1040)	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)

- Label Comparison accuracy:
  97.5% on JHHBench, **93.6%** on AtlasBench
- Label Comparison accuracy w/ class-agnostic prompt: **81.8%**
- Label Error Detection accuracy: **71.8%**

Problem: Label Errors

- Segmentation datasets are quickly growing thanks to AI annotations, which can contain **label errors**
- Many errors are easy to identify, but large-scale revision is too **time consuming** for radiologists





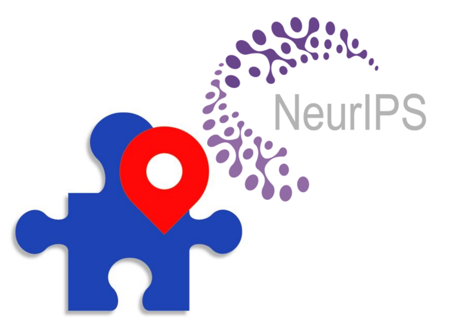
Objectives

- Use LVLMs to find and correct errors in medical segmentation annotations
- Reduce radiologist workload for revising medical segmentation datasets
- Choose the best label across between different medical segmentation labels

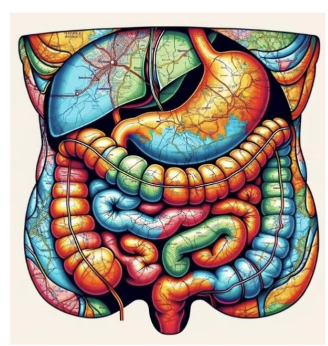
Conclusion

- **High accuracy** in label comparison (93.5% – 97.5%)
- Zero-shot LVLMs can accurately **correct** errors and send complicate cases for human review
- Label Critic **generalizes** to diverse label error types
- Label **comparisons** are more accurate than error detection with a single label
- We present **class-agnostic prompts** for easy deployment on new datasets and classes
- **General LVLMs** surpassed medical LVLMs in label error detection

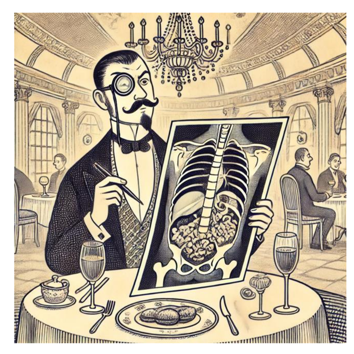
More of the BodyMaps Project



Touchstone Benchmark
5,185 training CTs
6,933 OOD test CTs
84 hospitals
14 teams



AbdomenAtlas 3.0 Dataset
9,262 CTs, 9,262 **Reports**
2,947 kidney, liver & pancreas **tumors**



Label Critic code & prompts available below

Johns Hopkins University | IIT Istituto Italiano di Tecnologia | NUS National University of Singapore | Alma Mater Studiorum Università di Bologna