Critique

1. **Tested only clean public datasets** – In the paper it's mentioned that only the work has been done on a few public datasets like -

- GSTRIDE gait/fall data,
- a palliative-care anxiety/hope dataset,
- and a combined colonoscopy polyp dataset.

These are relatively tidy research datasets, not messy, multi-hospital EHR data.

Improvement - Real hospital data is much uglier: missing values everywhere, inconsistent coding, weird column names, multiple hospitals, different scanners, etc. A system that works on 2–3 clean datasets might struggle badly in a real hospital. Add experiments on realistic, messy EHR data.

- "This is the model I plan to use - do you agree?"
- "This is how I classified your columns - any corrections?"

2. **No true multimodal fusion** - The framework supports tabular pipelines and image pipelines, but they are effectively separate tracks that happen to live in one system. Tabular data goes through GBM/deep models; images go through DETR; there is no model that jointly uses *both* for the same patient.

Improvement - Many real clinical tasks need multimodal reasoning: EHR tables + imaging + maybe notes. Treating each modality in isolation means you miss interactions (e.g., polyp image + patient risk factors). Why not implement multimodal branch into the framework.

3. **Implement human-in-the-loop design -** The framework is designed to be fully automatic, but the paper doesn't explain how clinicians or data scientists can review what the agents did, fix mistakes (like wrong feature type or wrong model), or feedback improvements. In medicine, that's risky people need to double-check, override bad decisions, and understand why the system acted a certain way.

Improvement - Add a simple review screen at key steps:

- "Here is the model I chose; do you agree?"
- "Here is how I classified your columns; any corrections?"

Let human feedback be stored and reused so the system becomes better tuned to each site.