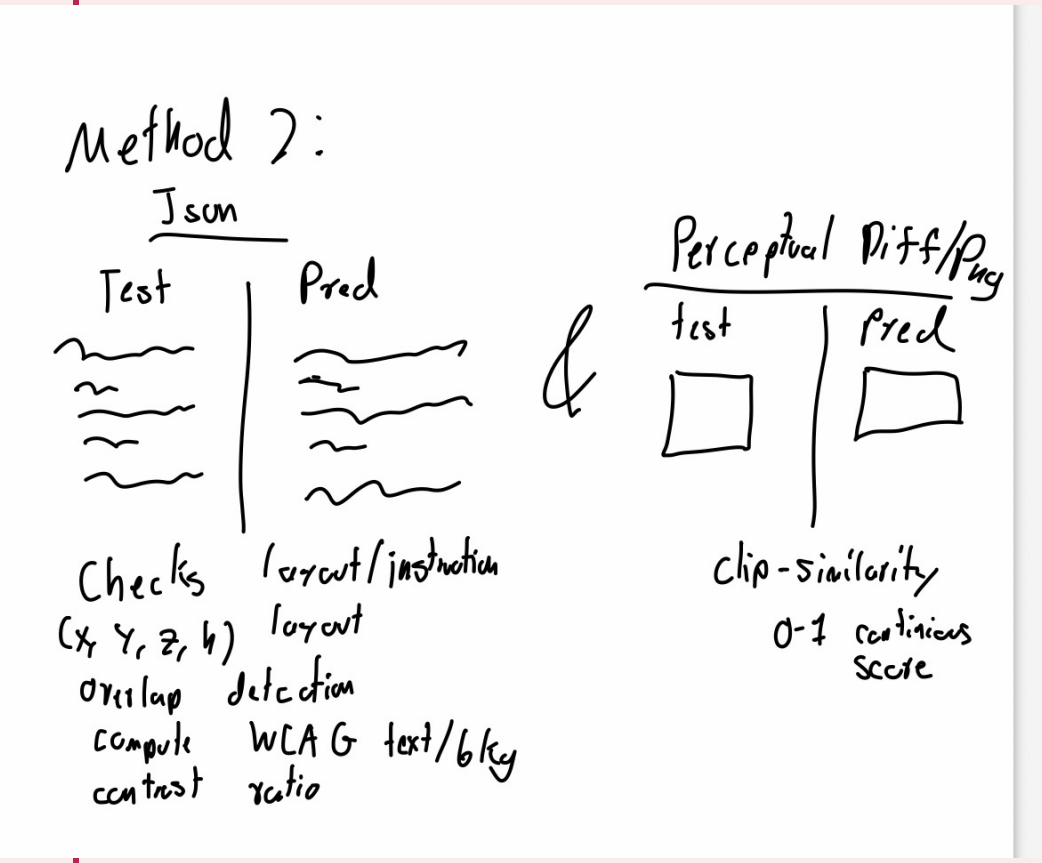
Upgrades?



Concrete upgrades to your benchmark design (task validity)

Negative controls and "impossible" tasks
Include edits that should never succeed, like "merge two master themes into one" or "chang on an embedded video" when that asset doesn't exist. Score any apparent success as a red f the τ-bench style failure mode where trivial or empty actions earn credit. arXiv

Shortcut audits
Seed slides where a model could "pass" without the target skill, for example:

Recoloring text by placing an opaque shape on top.

"Delete slide 3" when slide 3 is already blank or hidden.

"Summarize paragraph" when the paragraph is duplicated elsewhere and the model just del Each task should come with a minimal successful trajectory and a list of invalid bypasses tha for. This is straight from ABC's outcome/task validity framing. arXiv

Difficulty ladder and capability mapping
Tag every instruction with its capability primitive (e.g., text edit, selection by attribute, shap graph mutation, chart data binding, master/layout manipulation) and difficulty tier. Report improvement is interpretable, not just an aggregate number. arXiv