Beyond Instruction Following: Toward Proactive Human–AI Collaboration in Scientific Poster Refinement

Xingda Lyu, Honglin Lu, Xinye Luo, Shiqi Yang

University of Washington

Scan for Paper

arXiv:2609.01813

PROS Workflow

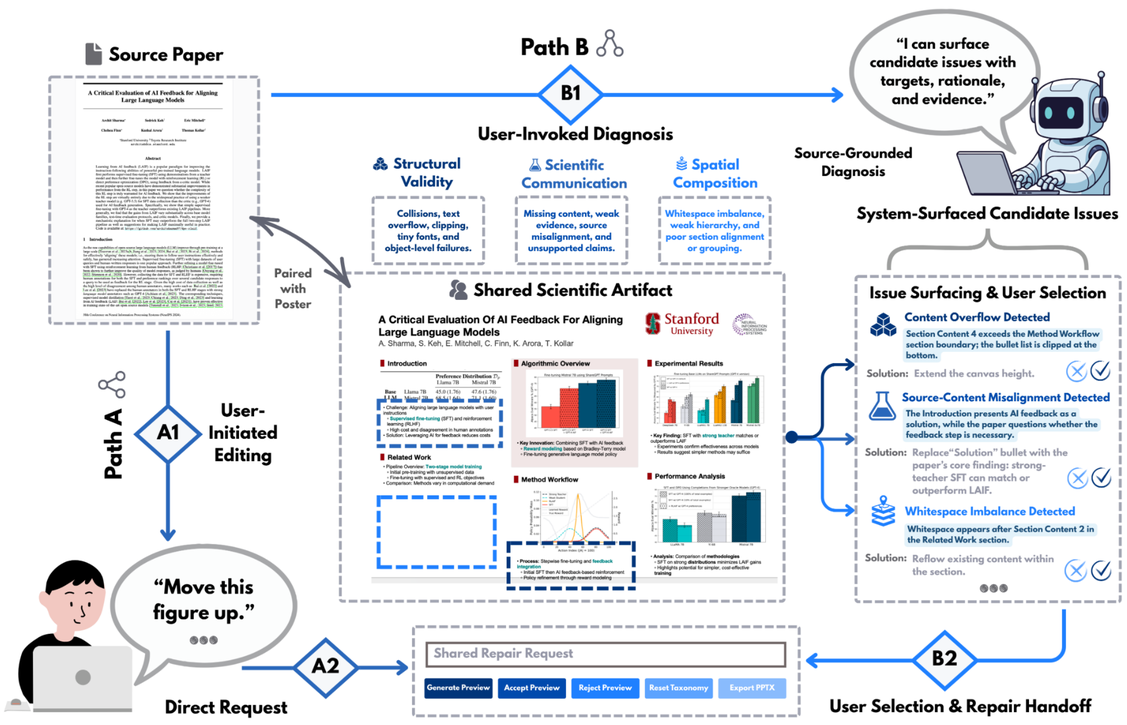


Figure 1. Two entry points: direct edits and source-grounded diagnosis.

Diagnosis and Repair

1. **Structural validity**
   Collisions, occlusion, overflow, and out-of-bounds objects.
2. **Scientific communication**
   Missing results or methods, content overload, and source-coverage mismatch.
3. **Spatial composition**
   Whitespace, alignment, section balance, grouping, and reading order.

Only user-accepted issues enter repair. PROS edits a temporary PPTX and validates the preview; the user accepts or rejects the change.

PROS-Bench

PROS-Bench contains 120 source papers and 320 editable posters. The primary evaluation uses 40 papers, each with Paper2Poster, PosterGen, and expert-authored variants.

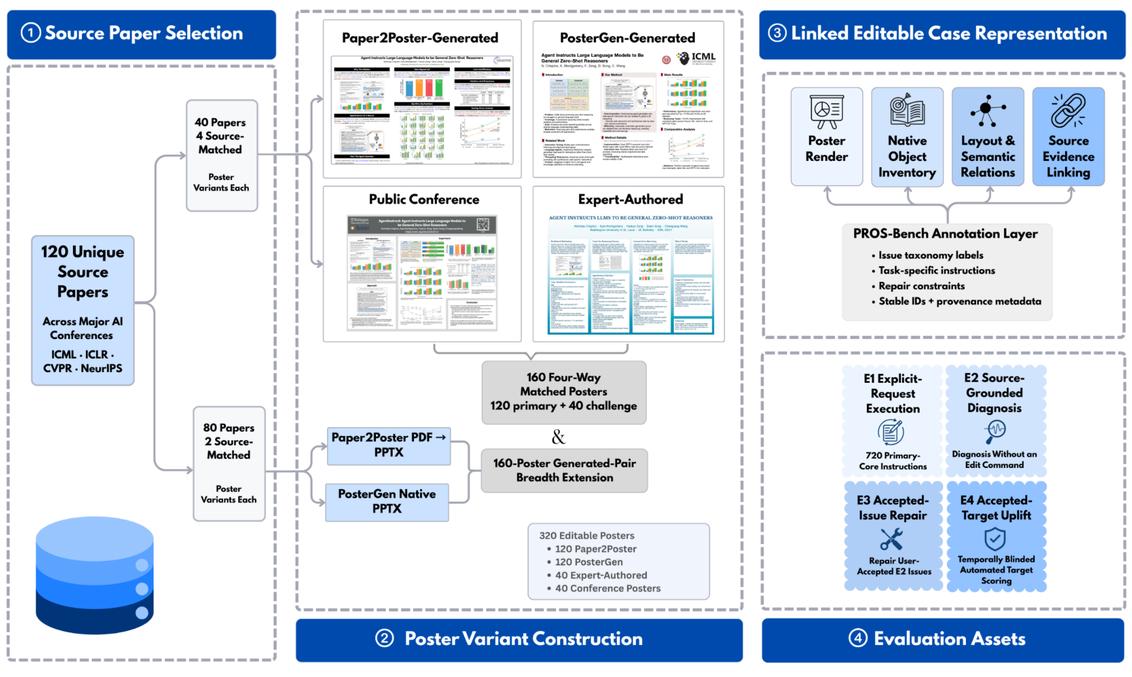


Figure 2. Source papers, poster variants, and evaluation assets.

Evaluation

Measure	Estimate	95% CI	n
E1 Editing	83.3%	[80.3, 86.2]	720
E2 Diagnosis	67.2/100	[64.2, 70.2]	1,100
E3 Repair	87.6%	[84.2, 91.0]	876
E4 Uplift	+22.7 pts	[17.6, 27.5]	865

Table 1. Primary evaluation: 40 papers and 120 posters.

E1/E3: operator-verified success. E2: stage-balanced VLM diagnosis quality. E4: paper-macro target-quality change (0–100), rated by a temporally blinded VLM. CIs are paper-clustered.

Example Repair

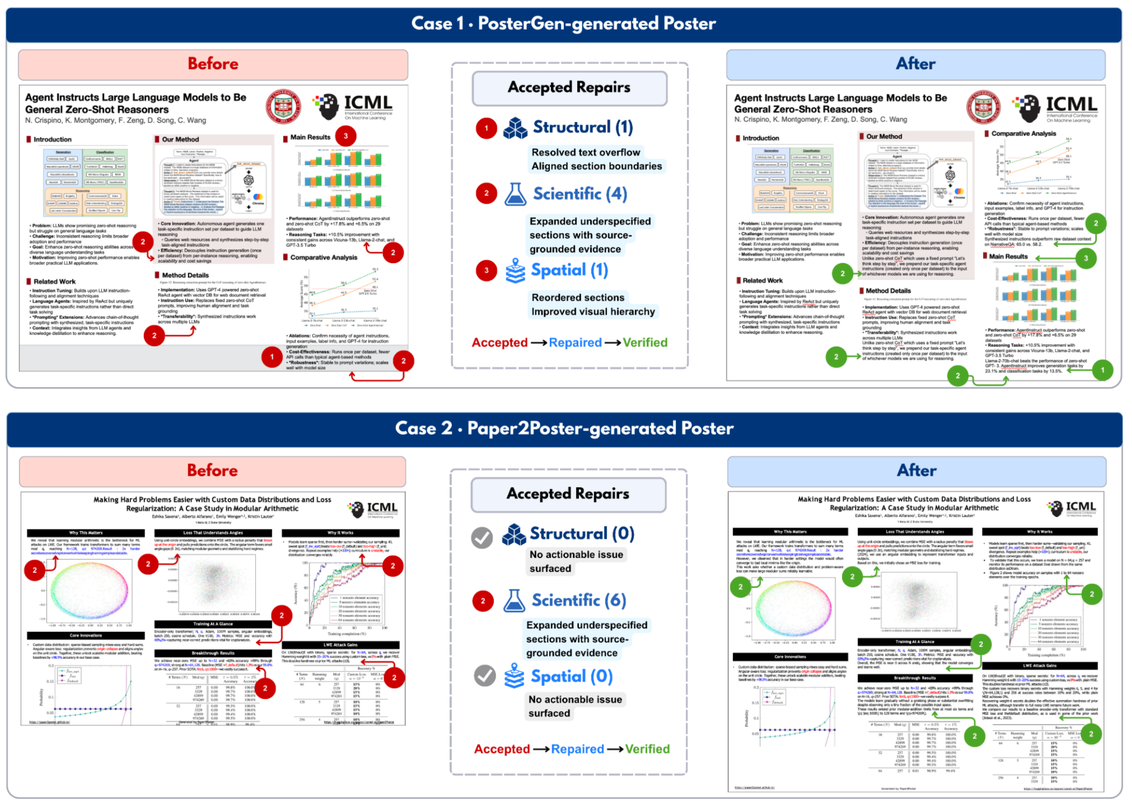


Figure 3. PosterGen and Paper2Poster posters before and after accepted repairs.

Outcomes Across Papers

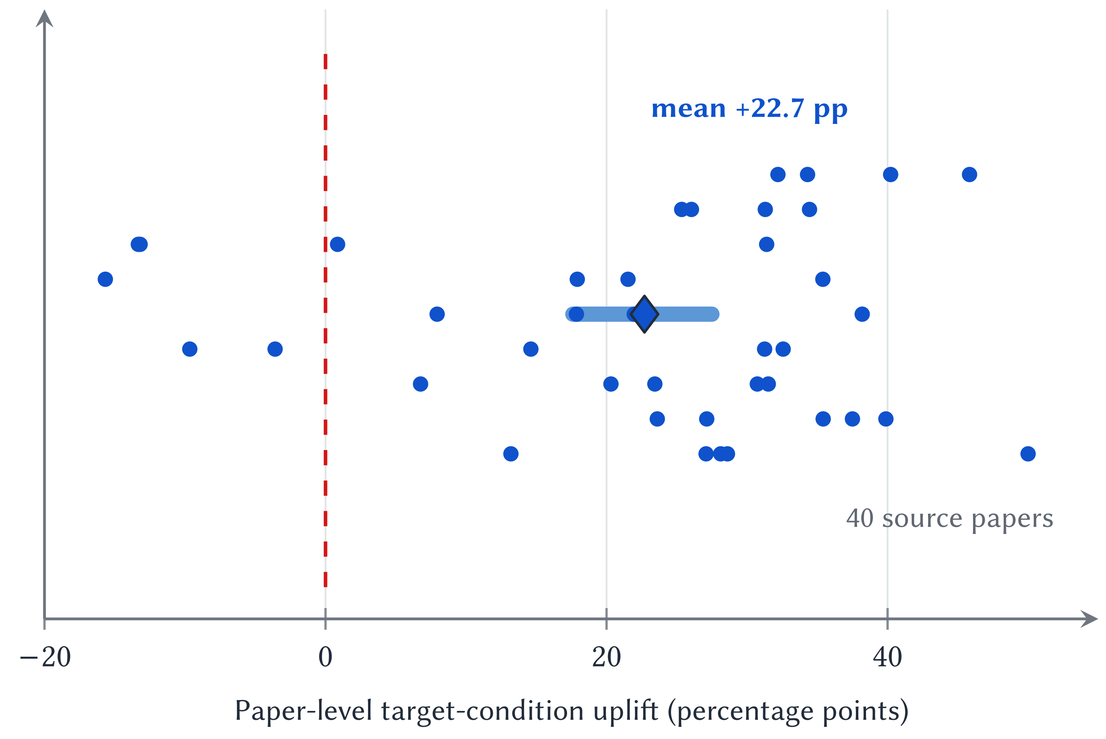


Figure 4. Paper-level uplift. Diamond: mean; bar: 95% CI.

Of 865 assessable targets, 67.1% improved, 18.2% tied, and 14.8% declined in final-state ratings.

xlyu18@uw.edu

Human-AI Complementarity Workshop