AutoDesign: Meta-Harness Optimization for Long-Horizon Agentic Design

Yaxin Luo, Haobin Jiang, Jialv Zou, Xu Huang, Wenhao Yan, Haodong Li, Zhengrong Yue, Jing Li, Xiaofu Chen, Xiaohan Zhao, Jiacheng Liu, Jiacheng Cui, Zhiqiang Shen, Xiaotong Li
Meituan, MBZUAI, Huazhong University of Science and Technology, Peking University, Tsinghua University, The Chinese University of Hong Kong, Shanghai Jiao Tong University

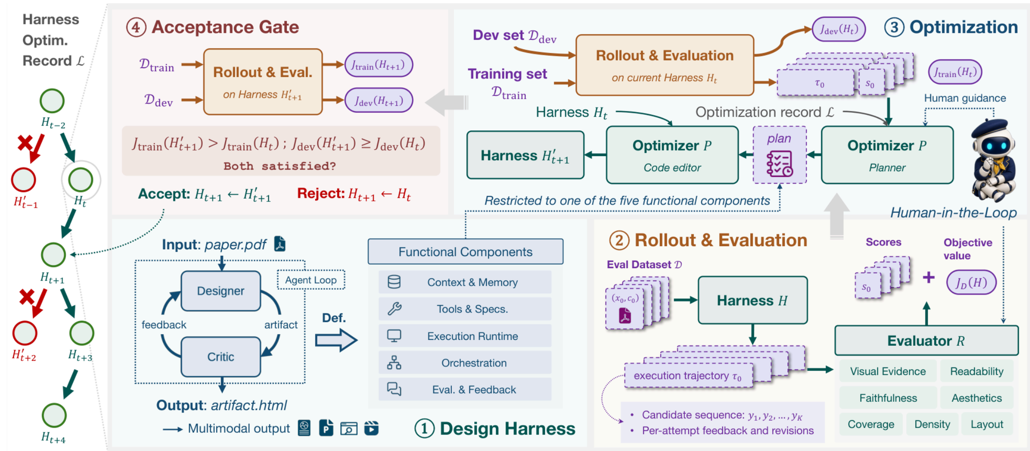
1. The Motivation

Problem: Multimodal design needs persistent, human-aligned production capabilities rather than isolated output edits. A useful system must preserve source evidence, communicate dense content, and remain usable after rendering.

Gap: Existing paper-to-poster benchmarks often inspect isolated layout, extraction, faithfulness, or visual quality; AutoDesign frames the task as joint evidence preservation, dense communication, and rendered usability.

Output: The generated poster remains editable for localized revision instead of becoming a static image.

Need: Design quality has to combine source faithfulness, dense communication, and reliable execution.



Revision: Given a paper, DesignHarness selects relevant evidence, creates an editable artifact, checks rule-based constraints and visual quality, then keeps the best valid candidate. The localized revision path supports direct use without discarding the paper-linked structure.

2. Existing Methods

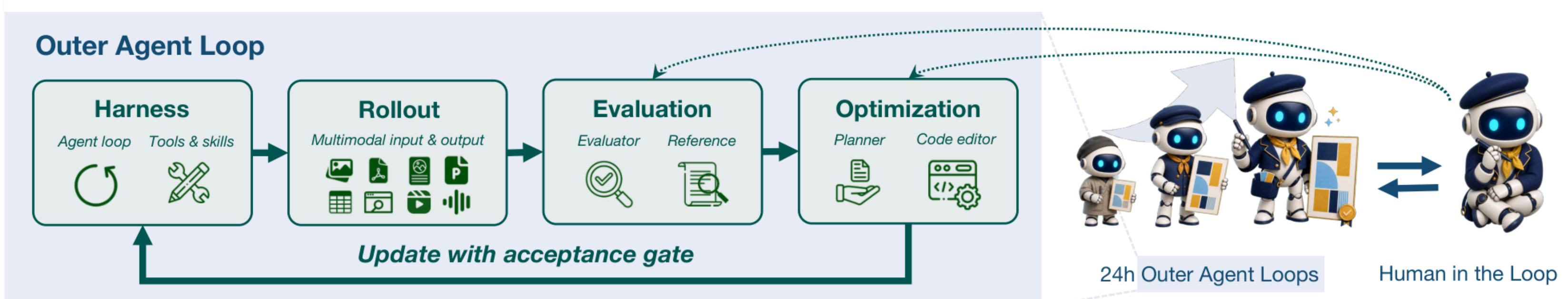
Structure: The inner loop revises artifacts; the outer loop revises the harness.

Update path: Across design tasks, the outer loop evaluates rollout outputs, proposes one component change, and accepts or rejects the candidate update. Critic feedback drives artifact revision inside the current harness before that system-level decision.

Contrast: TextGrad, DSPy, and GEPA optimize components or declarative systems, while STOP, GPTSwarm, ADAS, and AFlow search code- or graph-represented workflows. AutoDesign instead updates a reusable design harness from rollout records and evaluation outcomes.

Optimization unit	Representative systems	Scope
Components or declarations	TextGrad, DSPy, GEPA	Program-level optimization
Code or workflow graphs	STOP, GPTSwarm, ADAS, AFlow	Agent procedure search
Design harness	AutoDesign	Cross-task system evolution

3. The Method



Learning cycle: Rollout, evaluation, one-component update proposal, and acceptance define each iteration; human guidance can redirect the search.

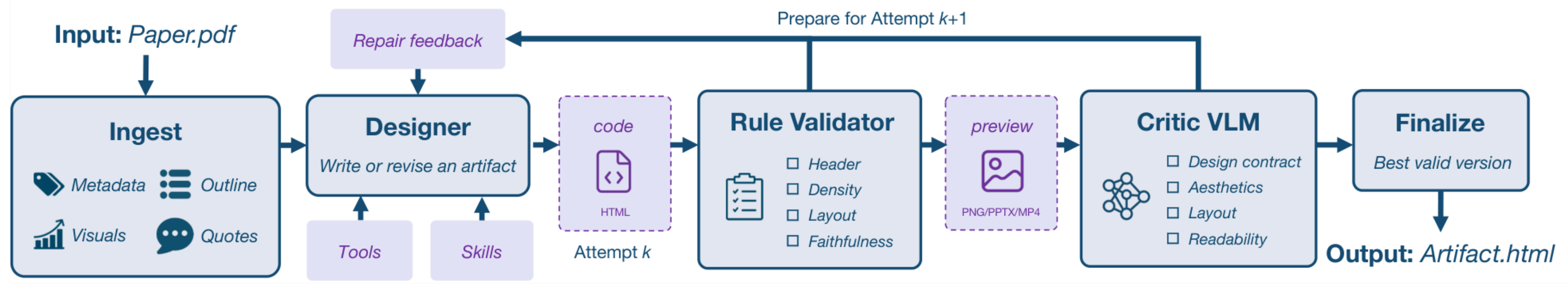
Acceptance gate: Promote updates only with training gain and preserved development performance.

Bounded update: Rollout trajectories, evaluation scores, and the optimization record inform one candidate harness change; that change remains inside one selected functional component.

Model boundary: The underlying model parameters remain fixed while the surrounding operating system is optimized.

Control	Paper-grounded policy
Update unit	One selected component
Functional scope	Five components
Promotion	Training gain + development retention
Search control	Optional human redirection

4. How It Works



Feedback: A designer revises an editable artifact using a rule-based validator and a model-based visual critic, then finalizes the best valid candidate. Paper ingestion precedes localized revision and selection of the best valid candidate.

Context: Metadata, section outlines, key passages, and paper visuals are selected once and retained across refinement, enabling localized code edits rather than regeneration of the whole artifact.

$$H^* = \arg\max_H J(H)$$

5. PosterBench Evaluation

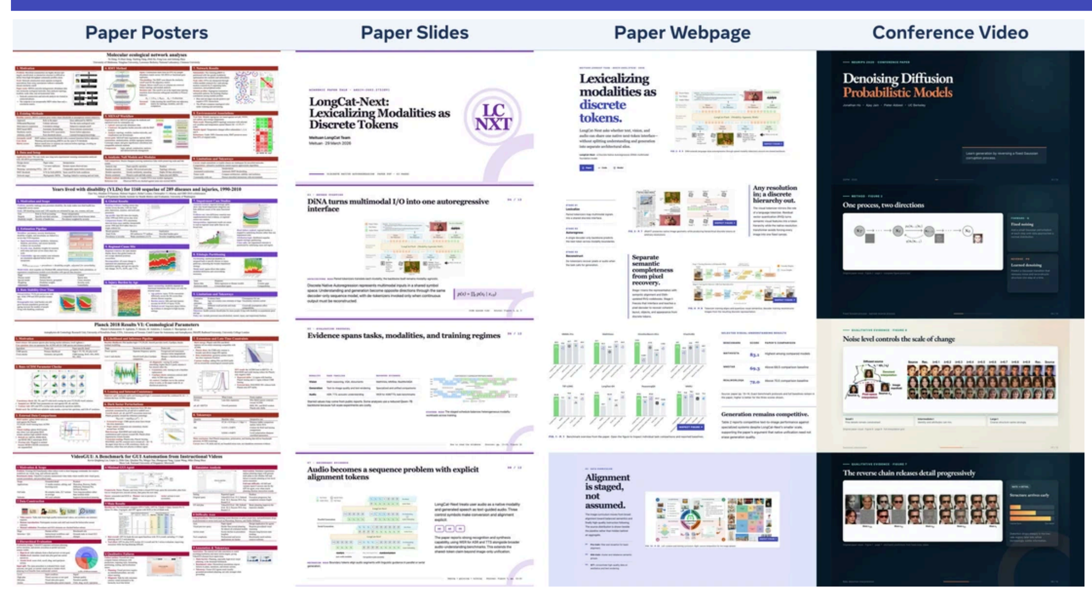
System	Design Harness	Coding Agent	Model	Score	Faith.	Cover.	Density	Vis. Ev.	Layout	Read.	Aesth.
Design Agent											
AutoDesign	**DesignHarness**	Claude Code	Claude 4.8	78.32	9.35	9.40	8.41	5.97	8.55	8.17	5.59
AutoDesign	**DesignHarness**	Codex	GPT 5.5	77.97	9.57	9.35	8.63	5.47	7.75	8.08	7.16
Claude Design	Claude Design	Claude Code	Claude 4.8	70.87	9.22	9.90	6.48	7.62	8.08	5.96	7.36
OpenDesign	OpenDesign	Claude Code	Claude 4.8	69.45	9.17	9.37	7.51	6.12	7.08	6.19	7.01
OpenDesign	OpenDesign	Codex	GPT 5.5	62.17	9.24	8.49	6.91	5.05	7.24	5.76	6.00
Coding Agent											
Codex	—	Codex	GPT 5.5	73.37	9.64	9.71	7.69	7.43	8.03	6.08	6.60
Claude Code	—	Claude Code	Claude 4.8	70.01	9.25	9.88	5.71	7.01	9.46	6.68	6.53
Doubao	—	Claude Code	Seed 2.1	61.14	8.93	8.96	5.95	5.97	6.51	5.63	4.61
GLM	—	Claude Code	GLM 5.2	52.22	8.60	7.44	5.23	5.29	7.03	4.90	5.15
Kimi	—	Claude Code	Kimi K2.7	51.46	8.33	7.15	4.72	5.62	7.26	4.93	4.81
DeepSeek	—	Claude Code	DeepSeekV4-Pro	46.01	8.23	6.68	4.10	3.82	6.95	4.41	3.38
Human-Crafted Workflow											
PosterGen	—	—	Claude 4.8	56.71	8.84	8.25	4.31	5.62	8.61	5.36	5.28
Any2Poster	—	—	Claude 4.8	49.09	8.26	5.44	4.18	3.65	9.59	4.47	3.10
Paper2Poster	—	—	Claude 4.8	44.61	6.35	2.35	8.36	2.16	3.69	4.87	1.69

Benchmark comparison: AutoDesign with DesignHarness leads the Design Agent group with 78.32 using Claude Code / Claude 4.8 and 77.97 using Codex / GPT-5.5, ahead of Claude Design and OpenDesign.

Dimension readout: The table reports the overall score together with faithfulness, coverage, density, visual evidence, layout, readability, and aesthetics, making the gain traceable across PosterBench criteria.

Benchmark scope: The comparison includes Design Agent, Coding Agent, and Human-Crafted Workflow baselines under the same PosterBench evaluation protocol.

6. Future Directions



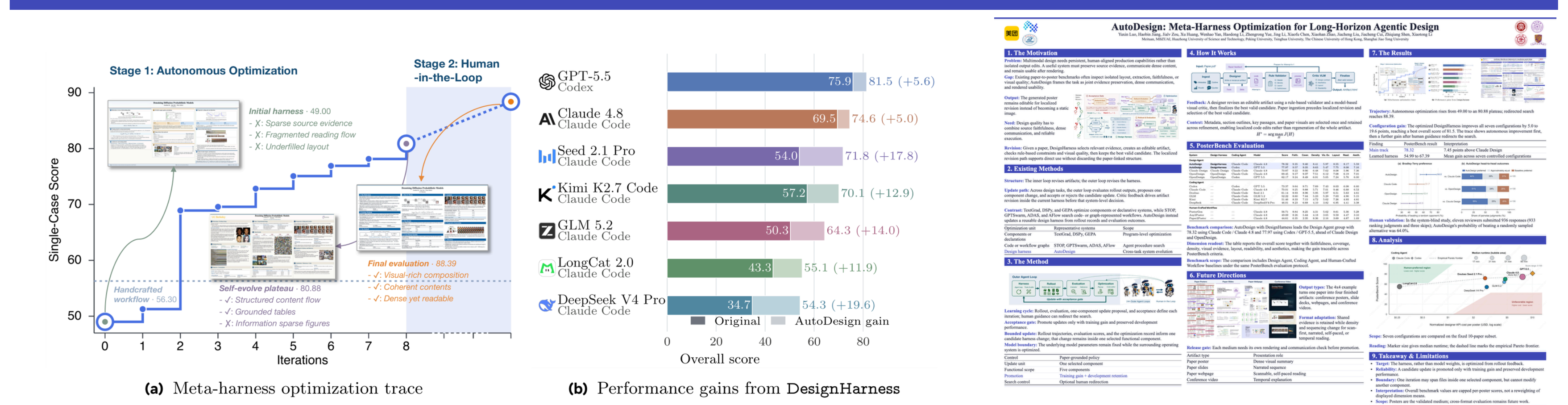
Output types: The 4x4 example turns one paper into four finished artifacts: conference posters, slide decks, webpages, and conference videos.

Format adaptation: Shared evidence is retained while density and sequencing change for scan-first, narrated, self-paced, or temporal reading.

Release gate: Each medium needs its own rendering and communication check before promotion.

Artifact type	Presentation role
Paper poster	Dense visual summary
Paper slides	Narrated sequence
Paper webpage	Scannable, self-paced reading
Conference video	Temporal explanation

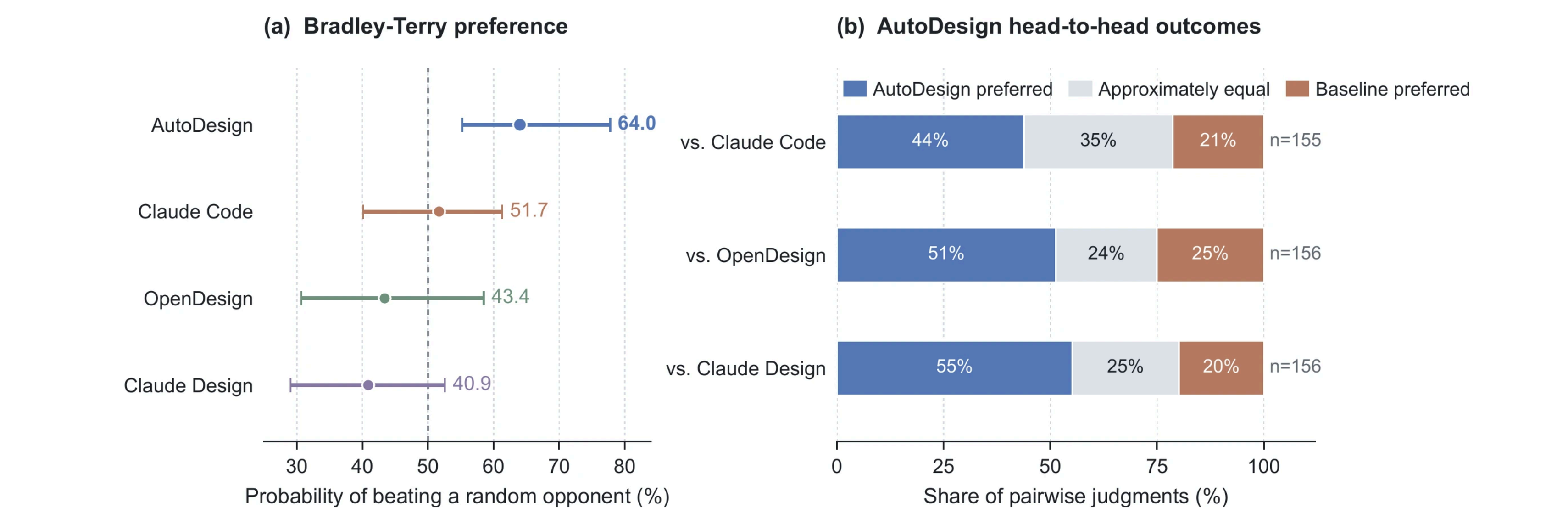
7. The Results



Trajectory: Autonomous optimization rises from 49.00 to an 80.88 plateau; redirected search reaches 88.39.

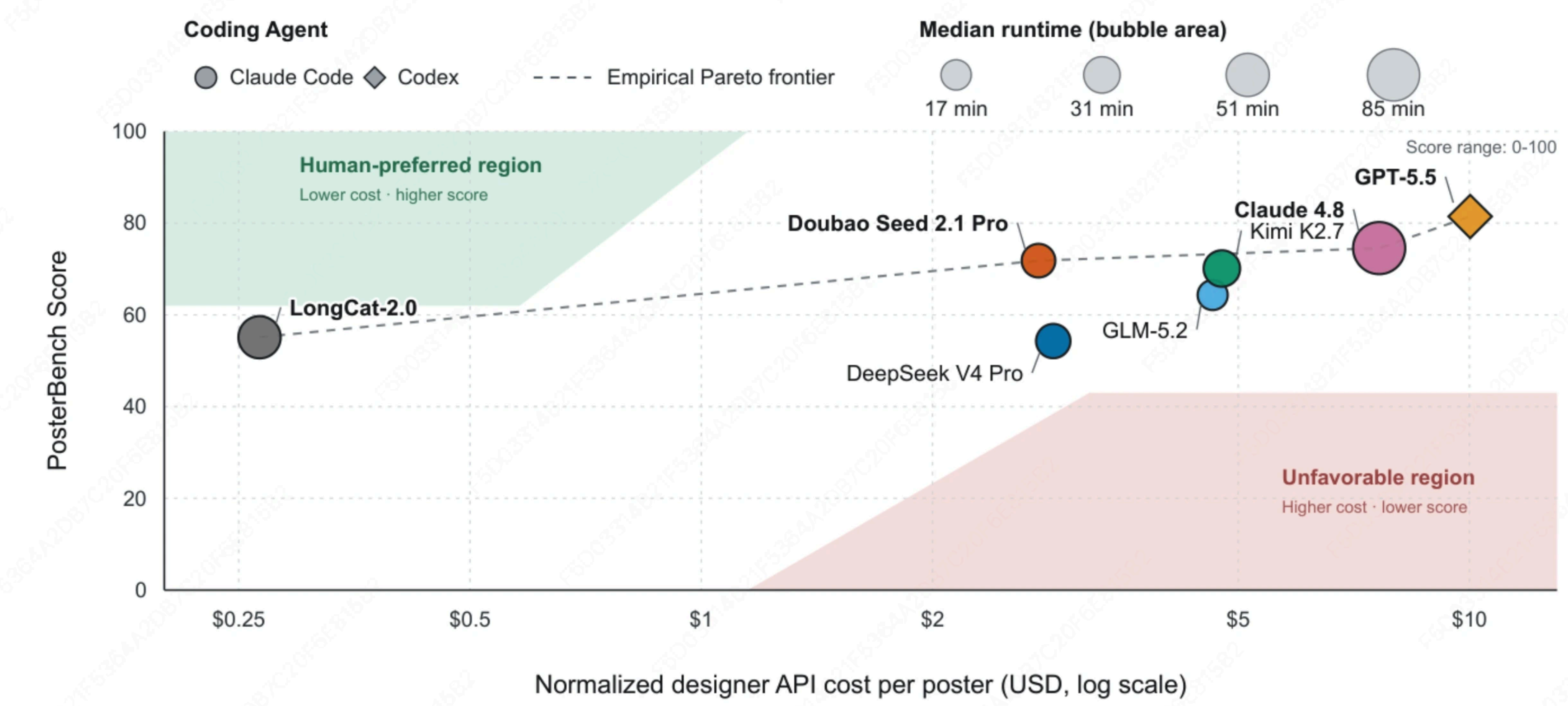
Configuration gain: The optimized DesignHarness improves all seven configurations by 5.0 to 19.6 points, reaching a best overall score of 81.5. The trace shows autonomous improvement first, then a further gain after human guidance redirects the search.

Finding	PosterBench result	Interpretation
Main track	78.32	7.45 points above Claude Design
Learned harness	54.99 to 67.39	Mean gain across seven controlled configurations



Human validation: In the system-blind study, eleven reviewers submitted 936 responses (933 ranking judgments and three skips); AutoDesign's probability of beating a randomly sampled alternative was 64.0%.

8. Analysis



Scope: Seven configurations are compared on the fixed 10-paper subset.

Reading: Marker size gives median runtime; the dashed line marks the empirical Pareto frontier.

9. Takeaway & Limitations

- **Target:** Optimize the harness, not model weights.
- **Promotion:** Keep only updates with training gain and no development regression.
- **Scope:** Update one selected component; PosterBench currently validates posters.