Visual Prompt Discovery via Semantic Exploration

Jaechang Kim[1,2], Yotaro Shimose[1], Zhao Wang[1], Kuang-Da Wang[1,3], Jungseul Ok[2], Shingo Takamatsu[1]
[1]SONY, [2]POSTECH, [3]NYCU

ECCV 2026 | MALMÖ | SEPT 8–12

TL;DR: Research agents can discover effective methods through experiments

About me

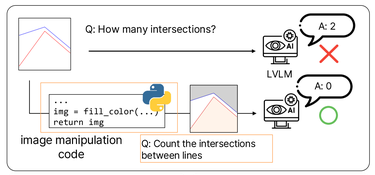
Project page

Motivation

Why visual prompt discovery?
- Effects are hard to predict → experiments are necessary
- The design space is too vast → needs to be efficient

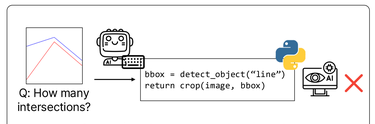
Manual discovery: extensive human trial-and-error
Zero-shot generation: high per-sample cost, model-specific failure

Can an agent efficiently discover a reusable visual prompt from a small development set?



Effective visual prompt | Manual discovery

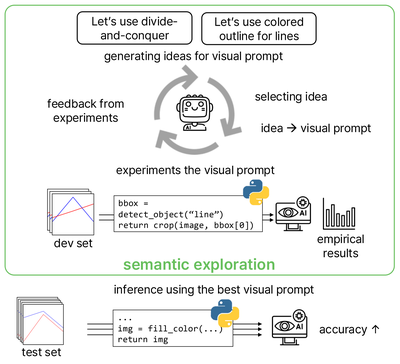
(a) Visual prompt (code + text prompt)



sub-optimal per-sample visual prompt

Suboptimal visual prompt | High inference cost

(b) Zero-shot visual prompt generation



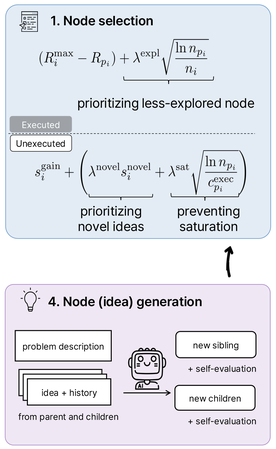
Effective visual prompt | Automatic discovery | Low inference cost | Efficient exploration

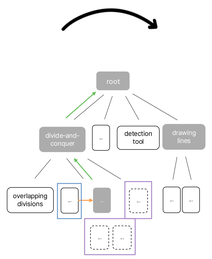
(c) Task-wise visual prompt exploration (ours)

Method (SEmantic Visual prompt EXploration)

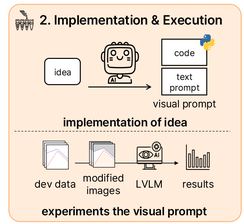
We structure the research process around three principles:
- **Abstract**: reason over hypotheses rather than low-level code
- **Explore & refine**: balance between novel directions and promising branches
- **Learn**: extract actionable insights from experiments

Why not vanilla UCT?
the number of LLM generated ideas is infinite
→ NUCT estimate the score of a node without execution



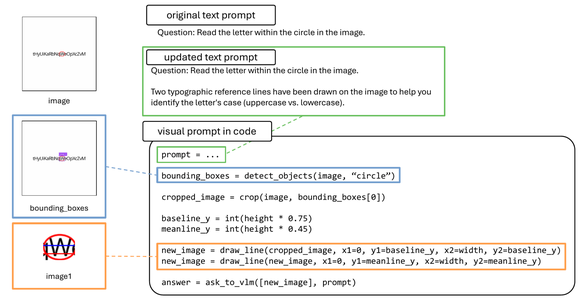


Node components
Abstract idea
Self-evaluation ($s^{\text{gain}}, s^{\text{novel}}$)
Visual prompt
Experiment History (reward, insight)

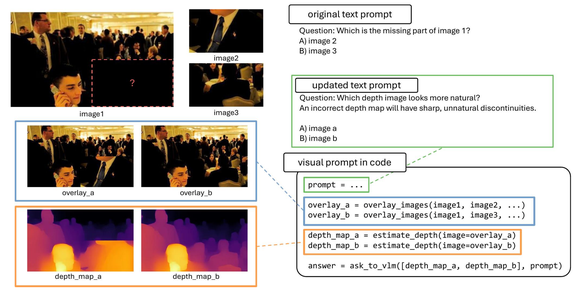


Examples

The agent refines implementation details
detect target → crop → add typographic reference lines
→ distinguish upper/lower case



The agent discovers an unintended use of an existing tool
Overlay candidate pieces → estimate depth
→ use depth discontinuities as a consistency signal



Visual prompts are backbone-specific
divide-and-conquer is effective for Claude, but not for GPT

	Gemini	Claude	GPT
few-shot	90.5	82.9	59.1
divide and conquer	87.9	87.8	31.0
crop	62.3	35.0	57.1

Results

30-image development set, 50 iterations, gemini 2.5

78.9% accuracy
vs 71.6% for naive

1.38k inference tokens
vs 15.62k for SketchPad

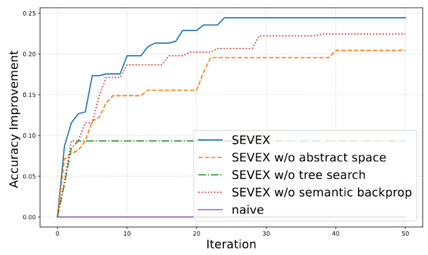
85k tokens / iter
vs 738k for SketchPad+APE

4.4%p dev-test gap
vs 10.1%p of SketchPad+APE

Task	Naive		SketchPad		SEVEX(ours)	
	accuracy	inf. cost	accuracy	inf. cost	accuracy	inf. cost
LineIntersections	73.0	981	33.3	16730	**90.5**	991
CircledLetter	78.8	1323	82.0	13902	**83.7**	1337
SubwayMap	60.0	1333	58.1	16482	**62.8**	1585
OverlappingShapes	50.4	2429	16.3	24833	**52.4**	2612
Average-BlindTest	65.6	1517	47.4	13987	**72.4**	1631
Jigsaw	75.8	869	70.8	14896	**95.8**	682
Depth	83.0	1349	**85.1**	14158	**85.1**	1457
Spatial	81.4	1525	85.8	15818	**86.7**	1555
SemanticCorr.	55.0	712	58.7	9477	**63.3**	1358
VisualCorr.	87.3	641	**90.9**	14291	89.4	801
Average-Blink	76.5	1019	78.3	13728	**84.1**	1171
Average	71.6	1240	64.6	15621	**78.9**	1375

Task	SketchPad+APE				SEVEX(ours)			
	dev acc.	test acc.	expl. cost	inf. cost	dev acc.	test acc.	expl. cost	inf. cost
LineIntersections	30.0	16.4	797k	11005	93.3	90.5	39k	991
CircledLetter	93.3	73.7	980k	17647	85.6	83.7	79k	1337
SubwayMap	50.5	42.5	759k	21471	80.0	62.8	165k	1585
OverlappingShapes	33.3	22.9	818k	20672	56.7	52.4	86k	2612
Average-BlindTest	51.7	38.9	838k	17699	78.9	72.4	92k	1631
Jigsaw	76.7	66.4	761k	18176	96.7	95.8	75k	682
Depth	76.7	85.9	794k	18867	85.6	85.1	87k	1457
Spatial	90.0	76.2	589k	19074	91.1	86.7	83k	1555
SemanticCorr.	76.7	52.6	385k	12354	72.2	63.3	74k	1358
VisualCorr.	83.3	83.1	761k	11937	88.9	89.4	78k	801
Average-Blink	80.7	72.8	658k	16082	86.9	84.1	80k	1171
Average	67.8	57.7	738k	16800	83.3	78.9	85k	1375

Removing any component slows or weakens discovery



Conclusion

- Agents can automate trial-and-error in method discovery
- Effective research agents need efficient exploration
- SEVEX points toward scalable human-AI collaboration for method discovery