A HUMAN-INSPIRED FRAMEWORK FOR CONTINUOUS SPATIAL REASONING WITH TANGRAM PUZZLES

Yikun Zong *
Department of Engineering
University of Cambridge, United Kingdom
{yz977}@cam.ac.uk

Cheston Tan
Centre for Frontier AI Research
A*STAR, Singapore
{cheston-tan}@a-star.edu.sg

ABSTRACT

Humans excel at spatial reasoning tasks like Tangram puzzle assembly through cognitive processes involving mental rotation, iterative refinement, and visual feedback. In contrast, Vision–Language Models (VLMs) struggle with continuous geometric reasoning despite their success on discrete benchmarks. This paper bridges **human cognition and AI reasoning** by introducing a framework that explicitly models human spatial cognitive capabilities and incorporates them into AI reasoning processes. We propose a **human-inspired test-time refinement framework** that mimics how humans iteratively correct spatial predictions through feedback-guided adjustments, combining in-context learning with reward-guided feedback loops. Experiments reveal that current VLMs achieve only 0.41 IoU on single-piece tasks (dropping to 0.23 on two-piece composition), far below human performance; humans can complete Tangram tasks even in childhood, demonstrating significantly high continuous spatial reasoning ability (Bohning & Althouse, 1997). Our **human-inspired verifier–refiner agent** applies reward-guided refinement loops that model human iterative correction processes, achieving IoU improvements from 0.63 to 0.932 *without any model retraining*, demonstrating that incorporating explicit models of human cognitive capabilities can substantially enhance AI reasoning in continuous spatial domains. Our work is available at this anonymous link https://anonymous.4open.science/r/TangramVLM-F582/.

1 INTRODUCTION

Humans excel at spatial reasoning tasks like Tangram puzzle assembly through cognitive processes involving mental rotation, iterative refinement, and visual feedback (Shepard & Metzler, 1971). Despite their success on discrete benchmarks (Radford et al., 2021), current Vision–Language Models (VLMs) struggle in continuous geometric reasoning, revealing a fundamental gap between human cognitive capabilities and AI reasoning. Building on cognitive science research showing that humans solve spatial tasks through iterative refinement and feedback-driven self-correction (Shepard & Metzler, 1971; Bohning & Althouse, 1997), we design a benchmark where Tangram pieces must be precisely arranged to cover a target silhouette. Tangrams offer a cognitively natural yet computationally challenging testbed (NP-hard (Yamada et al., 2024)) that bridges cognitive science and AI. We establish a **human-inspired reasoning framework** that incorporates human teaching (few-shot demonstration) and correction (feedback loop), modeling decomposition (position, angle, size) and iterative refinement, yielding interpretable reasoning—position, angle, and size are individually traceable. Through systematic evaluation across VLMs, we find that even frontier systems achieve only $\approx$0.41 IoU, relatively low compared to human performance (Bohning & Althouse, 1997). **Crucially**, we demonstrate that incorporating human-inspired iterative refinement mechanisms, a reward-guided feedback loops that model human self-correction processes which can achieve substantial improvements (IoU gains of +0.30) at test time without model training, bridging the gap between human cognition and AI reasoning.

*Corresponding author

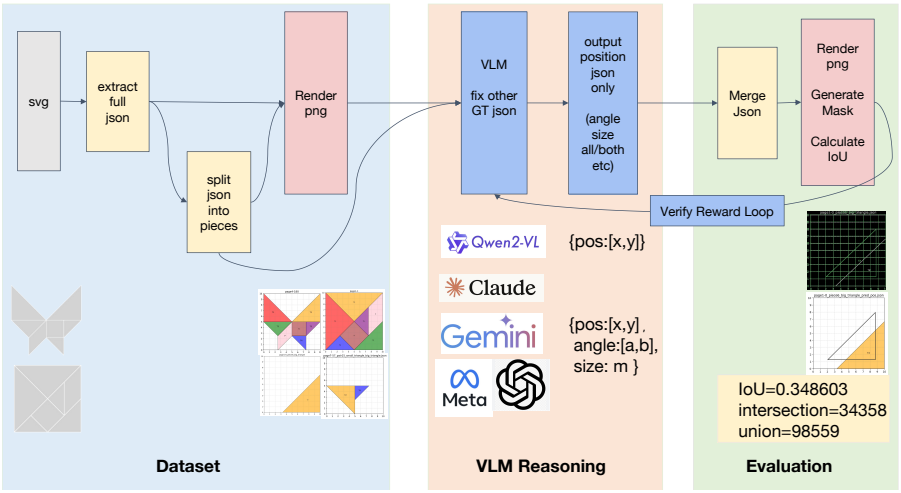


Figure 1: The diagram shows how raw SVG tangram silhouettes are parsed into JSON annotations (type, position, angle, size) rendered into training/evaluation images, and split into single-piece, two-piece, or full-tangram subsets for spacial reasoning and evaluation

Contributions Our key innovations are: (1) We introduce a *human-cognition-inspired* evaluation benchmark for AI spatial reasoning in continuous space, explicitly modeling how humans decompose spatial problems (position, angle, size) and measuring performance using geometric metrics that align with human cognitive evaluation; (2) We conduct a systematic evaluation across leading VLMs and find consistent geometric failures compared to human performance: average IoU on single-piece tasks is only 0.41, relatively low compared to human performance (Bohning & Althouse, 1997), and two-piece composition drops to 0.23, revealing a fundamental gap between human cognition and AI reasoning; (3) We propose a lightweight, **human-inspired test-time refinement framework** that explicitly incorporates models of human cognitive correction mechanisms, a reward-guided refinement loops that mirror how humans iteratively improve spatial estimates through feedback to enhance geometric alignment *without parameter updates*, lifting IoU from 0.63 to 0.93 and bridging human cognition and machine reasoning.

2 Related Work

Extensive research in cognitive science demonstrates that humans excel at spatial reasoning through processes involving mental rotation (Shepard & Metzler, 1971), iterative refinement, and visual feedback. Cognitive science research shows that human spatial cognition operates on fundamental primitives: position, orientation, and scale (Biederman, 1987; Spelke & Kinzler, 2007). Tangram puzzles have been extensively studied as a probe of human spatial intelligence (Bohning & Althouse, 1997), with developmental work on decomposition and iterative correction (Lee et al., 2009; Antrilli, 2019; Liben, 2007). Recent work on incorporating human cognitive models into AI systems includes approaches that model human mental states (Goodman & Frank, 2016), incorporate human teaching and correction into learning processes (Guo, 2021), and use human cognitive patterns to make AI systems more interpretable (Lake et al., 2017). In the context of test-time refinement, approaches like ReAct (Yao et al., 2022) and Reflexion (Shinn et al., 2023) show that language models can refine reasoning traces based on feedback, while self-consistency decoding (Wang et al., 2022) and verifier-guided generation (Cobbe et al., 2021) demonstrate iterative refinement. However, these focus on discrete or symbolic reasoning. Prior benchmarks examine spatial reasoning through tasks like mental rotation (Shepard & Metzler, 1971), spatial relation matching (Johnson et al., 2017), and compositionality (Wu et al., 2023; Hesham et al., 2025), but largely reduce spatial reasoning to discrete judgments without measuring *continuous* geometric errors. Our work extends this to *continuous geometric reasoning* by explicitly modeling human cognitive correction mechanisms (iterative refinement, feedback-driven correction) in a continuous metric space.

Algorithm 1 VLM + ICL + Reward Loop (simplified)

Input: image I, model M, mode $\in \{\text{pos}, \text{angle}, \text{size}, \text{all}\}$, ICL size k, loop iters T, threshold τ
Output: best JSON prediction $J^\star$, best IoU

$\mathcal{S} \leftarrow$ sample k few-shot (image, JSON) pairs for ICL Initialize best $\leftarrow$ (iou $= 0,\ J = \varnothing$) $t = 1$ to T Query M with I + $\mathcal{S}$ + refinement hint Parse output $\rightarrow J_t$ (JSON fields) Compute $\text{iou}_t = \text{IoU}(J_t,\ G)$ $\text{iou}_t >$ best.iou best $\leftarrow (J_t,\ \text{iou}_t)$ best.iou $\geq \tau$ **break** Optionally run small local search around best.pos **return** $J^\star =$ best.J, best.iou

3 METHODOLOGY

We build a Tangram benchmark with two splits: **Single-piece**—each image contains a *single* canonical piece with GT $(\mathbf{p}, \alpha, s)$, supporting pos/angle/size decoupling aligned with cognitive primitives (position, orientation, scale) (Biederman, 1987; Spelke & Kinzler, 2007); **Two-piece**—each image contains two pieces whose relative arrangement matters, probing compositional reasoning. Annotations are JSON: `type`, `pos`$= [x, y]$, `angle` (in degrees), `size` ($s > 0$). Shapes are rendered from canonical templates for exact polygon IoU (no boundary noise). Tasks: **pos-only**, **angle-only**, **size-only**, **two-piece**, **joint**; details in Appendix A.2 and A.1.

We report rasterized intersection-over-union (IoU) on a 512×512 canvas with 1–2 px dilation for morphology-tolerant evaluation. We evaluate Vision–Language Models (VLMs) and compare their performance to human cognitive capabilities. Each model receives a Tangram silhouette image and outputs a minimal JSON with the requested field(s): position (`pos`), orientation (`angle`), or scale (`size`). Predictions are evaluated in a normalized $[0, 1]^2$ coordinate frame. Geometric consistency is assessed using Euclidean position error, angular deviation, scale difference, and IoU, enabling comparison to human performance (Bohning & Althouse, 1997). We report results under both zero-shot and few-shot ICL settings (typically $k = 15$).

Humans solve spatial reasoning tasks through a functional cognitive process: trial and error $\rightarrow$ observation $\rightarrow$ correction. Although we do not claim to replicate neural mechanisms, our framework functionally mimics this behavioral pattern, incorporating the 'mental rotation' and 'iterative refinement' process as described by Shepard & Metzler (Shepard & Metzler, 1971) into an AI reasoning pipeline. This cognitive process involves explicit decomposition (position, angle, size), mental simulation, and feedback-driven correction, a mechanisms we explicitly model in our AI framework. For a single tangram piece, the pose is $\Theta = (\mathbf{p}, \alpha, s)$ with position $\mathbf{p} \in [0, 10]^2$, angle α (deg), and size $s > 0$. Let $\mathcal{U}(\Theta)$ be the rendered polygon from the canonical template under $(\mathbf{p}, \alpha, s)$, and S the ground-truth polygon. We use a scalar reward that trades off geometric coverage (IoU) against position error:

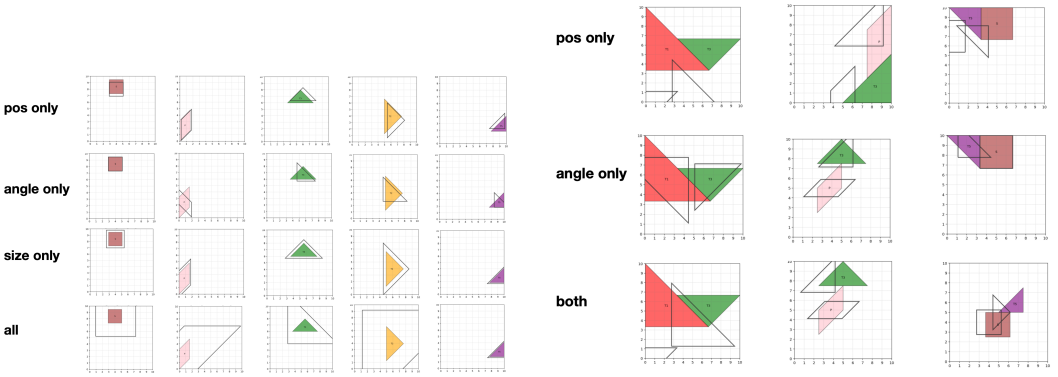
$$\mathcal{R}(\Theta) = \text{IoU}(\mathcal{U}(\Theta), S) - \lambda \cdot \|\hat{\mathbf{p}} - \mathbf{p}\|_2 / 10 \tag{1}$$

where $\hat{\mathbf{p}}$ is the GT position and $\lambda > 0$ is a small weight. Inspired by how humans iteratively refine spatial estimates through visual feedback, we run T iterations: (i) build k few-shot pairs excluding the current sample (modeling how humans use prior examples); (ii) query the VLM with a JSON instruction, appending numeric feedback hints from the second iteration onward (modeling human feedback-driven correction); (iii) keep the candidate with the highest $\mathcal{R}$ in Eq. equation 1, with early stop once $\text{IoU} \geq \tau$. If $\text{IoU} < \tau$ and the task involves position, we perform a tiny grid search around the current best (x, y) using a 3×3 neighborhood at step sizes $0.6 \rightarrow 0.3 \rightarrow 0.15$ (canvas units), modeling how humans perform fine-grained local adjustments. This *training-free* approach treats the VLM as a proposal generator with a geometry-based verifier, explicitly incorporating human cognitive mechanisms (iterative refinement, feedback-driven correction, local search) into AI reasoning processes.

4 RESULTS AND ANALYSIS

4.1 BASELINE VLM PERFORMANCE: SINGLE-PIECE AND TWO-PIECE TASKS

Setup. We evaluate VLMs on both single-piece and two-piece Tangram assembly tasks, comparing their performance to human cognitive capabilities. For *single-piece tasks*, we evaluate *pos-only*, *angle-only*, *size-only*, and *joint (all)* predictions. The factorized task design (mirroring how humans



(a) Single-piece setup and outputs: position, angle, size, and all.

(b) Two-Piece setup and outputs: position, angle and both.

Figure 2: Spatial reasoning tasks: single-piece and two-piece Tangram assembly.

decompose spatial problems) enables us to isolate specific geometric reasoning failures. For *two-piece tasks*, we test compositional reasoning with three modes: (A) fix both (α, s), predict $(\mathbf{p}_1, \mathbf{p}_2)$; (B) fix $\mathbf{p}, s$ and predict *angles*; (C) predict positions + angles jointly (scaled fixed). We report mean$\pm$95%CI over the test set. Metrics: task-specific errors (L2, angular degrees, relative scale) and IoU, with human baseline (Bohning & Althouse, 1997).

Table 1: Unified results for VLM one-piece spatial reasoning (↑ higher is better). Human baseline: relatively high performance (Bohning & Althouse, 1997).

Method	**Pos IoU ↑**	**Angle IoU ↑**	**Size IoU ↑**	**All IoU ↑**
Gemini-2.5-pro	**0.443**	**0.434**	**0.432**	**0.417**
GPT-4o mini	0.427	0.429	0.393	0.413
Qwen-72B	0.415	0.432	0.425	0.408
Claude-Sonnet-4	0.419	0.394	0.372	0.395

Findings. Single-piece tasks. Table 1 shows that larger models reduce L2 and scale errors, but angle remains fragile across all models, achieving only $\approx$0.41 IoU, relatively low compared to human performance (Bohning & Althouse, 1997). Joint prediction aggregates noise across multiple axes, amplifying errors versus factorized tasks (mirroring how humans decompose spatial problems into subcomponents). IoU is highly sensitive to angular mismatch even when position errors are small, revealing a fundamental gap between human cognitive capabilities and current AI reasoning.

Two-piece tasks. Table 2 shows that arrangement is *significantly* harder than single-piece: IoU drops $\sim$0.3 even when single-piece IoU exceeds 0.7, revealing that compositional reasoning remains challenging for current VLMs compared to human performance. Typical failure modes include mutual collision, near-miss adjacency, and mirrored angles, highlighting the gap between human cognitive capabilities and current AI reasoning.

4.2 Part II: Human-Inspired Test-Time Refinement via Reward-Guided Feedback

Setup. We focus on the *medium triangle* subset (single-piece), starting from the VLM's JSON output and running T human-inspired refinement loop iterations with reward $\mathcal{R}$ that models how humans iteratively correct spatial estimates through feedback. We allow a tiny local search over $\mathbf{p}$ at the end if IoU remains low, modeling how humans perform fine-grained local adjustments.

Findings. Table 3 shows that the human-inspired refinement loop elevates IoU from 0.65 to **0.932** (**+0.282**), far exceeding ICL-only gains (+0.145). The loop mechanism contributes the majority of improvement, effectively correcting small positional and angular errors. Combining ICL with the loop achieves similar performance (0.932) while improving initialization stability. The recom-

Table 2: Unified results for VLM two-piece spatial reasoning (↑ higher is better). Human baseline follows the same reference as Table 1, that human perform well in these tasks

Model	Pos IoU ↑	Angle IoU ↑	All IoU ↑
Gemini-2.5-pro	**0.340**	0.397	0.340
Qwen-72B	0.253	**0.495**	0.248
GPT-4o mini	0.276	0.394	0.278
Claude-Sonnet-4	0.318	0.394	0.235

Table 3: Medium triangle IoU across different settings (baseline start = 0.65).

Setting Number	Description	ICL (k)	Loop	Threshold	Temp.	IoU (final)
1	`VLM + ICL + Loop`	15	6	0.9	0	**0.9320**
2	`VLM + Loop`	n/a	6	0.9	0	0.9320
3	`VLM + ICL + Loop`	20	6	0.9	0	0.9300
4	`VLM + ICL`	15	n/a	n/a	0	0.7950
5	`VLM + ICL + temp`	15	n/a	n/a	0.5	0.7690
6	`VLM only`	n/a	n/a	n/a	0	0.6500

mended setting is **ICL** k=15 **+ Loop** =6 **+** τ=0.9 **+ temperature** =0 (see Appendix A.3 for detailed ablation studies).

5 Discussion

Our results demonstrate that explicitly modeling human cognitive correction mechanisms—reward-guided refinement loops that mirror how humans iteratively refine spatial estimates—enables substantial test-time improvement in VLM geometric reasoning, achieving IoU gains of over 40% without parameter updates. Refinement loops converge within 6 iterations on average, indicating efficient human-inspired feedback-driven self-correction. For single-piece tasks, most models achieve moderate IoU (0.2–0.45) without refinement, far below human performance (relatively low compared to human performance (Bohning & Althouse, 1997)). Human-inspired test-time refinement consistently improves alignment, with small geometric offsets corrected over 1–2 iterations. For two-piece tasks, performance is lower overall since errors compound, but relative gains from human-inspired refinement loops are consistent with the single-piece setting. VLMs consistently perform poorly in continuous-space evaluation (only ≈0.40 IoU) due to training distribution mismatch, output mismatch (autoregressive decoders force continuous quantities into discrete tokens), and absence of geometry-aware feedback during training. However, our human-inspired test-time refinement framework demonstrates that geometric feedback can effectively guide refinement even without training-time supervision. Our structured task decomposition (position, angle, size) based on cognitive primitives not only improves performance but also enhances interpretability by making the reasoning process transparent and aligned with human cognitive evaluation. Extensive developmental studies show that school-age children achieve near-perfect Tangram performance through explicit problem decomposition, mental simulation, and iterative correction. Our results demonstrate that modeling these human cognitive mechanisms can substantially bridge the gap between human cognition and AI reasoning, achieving IoU improvements from 0.63 to 0.93.

6 Conclusion

We presented a human-inspired reasoning framework that explicitly incorporates models of human cognitive capabilities into AI reasoning processes, bridging human cognition and AI reasoning in continuous geometric space. By modeling how humans iteratively refine spatial predictions through feedback-driven correction, our reward-guided refinement loops achieve IoU improvements from 0.63 to 0.93 without retraining, substantially closing the gap between human cognition and AI reasoning. Our work demonstrates that explicitly modeling human cognitive mechanisms for iterative refinement, feedback-driven correction, and problem decomposition can enable AI systems to reason more like humans, contributing to the broader vision of human-inspired AI reasoning systems.

REFERENCES

Nicola K Antrilli. *Tangrams 101: Effects of Tangible and Digital Play on Children's Spatial Reasoning and Parental Spatial Language*. PhD thesis, University of California, Santa Cruz, 2019.

Irving Biederman. Recognition-by-components: a theory of human image understanding. *Psychological review*, 94(2):115, 1987.

Gerry Bohning and Jody Kosack Althouse. Using tangrams to teach geometry to young children. *Early childhood education journal*, 24(4):239–242, 1997.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Noah D Goodman and Michael C Frank. Pragmatic language interpretation as probabilistic inference. *Trends in cognitive sciences*, 20(11):818–829, 2016.

Tong Guo. Learning from how humans correct. *arXiv preprint arXiv:2102.00225*, 2021.

Syed Ariff Syed Hesham, Yun Liu, Guolei Sun, Henghui Ding, Jing Yang, Ender Konukoglu, Xue Geng, and Xudong Jiang. Exploiting temporal state space sharing for video semantic segmentation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 24211–24221, 2025.

Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2901–2910, 2017.

Brenden M Lake, Tomer D Ullman, Joshua B Tenenbaum, and Samuel J Gershman. Building machines that learn and think like people. *Behavioral and brain sciences*, 40:e253, 2017.

Joohi Lee, Joo Ok Lee, and Denise Collins. Enhancing children's spatial sense using tangrams. *Childhood Education*, 86(2):92–94, 2009.

Lynn S Liben. Education for spatial thinking. *Handbook of child psychology*, 4, 2007.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Roger N Shepard and Jacqueline Metzler. Mental rotation of three-dimensional objects. *Science*, 171(3972):701–703, 1971.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. *Advances in Neural Information Processing Systems*, 36:8634–8652, 2023.

Elizabeth S Spelke and Katherine D Kinzler. Core knowledge. *Developmental science*, 10(1):89–96, 2007.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2022.

Chenwei Wu, Li Erran Li, Stefano Ermon, Patrick Haffner, Rong Ge, and Zaiwei Zhang. The role of linguistic priors in measuring compositional generalization of vision-language models. In *Proceedings on*, pp. 118–126. PMLR, 2023.

Fernanda Miyuki Yamada, Harlen Costa Batagelo, João Paulo Gois, and Hiroki Takahashi. Generative approaches for solving tangram puzzles. *Discover Artificial Intelligence*, 2024.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. In *The eleventh international conference on learning representations*, 2022.

A APPENDIX

A.1 DATASET CONSTRUCTION PIPELINE

Algorithm 2 Tangram Dataset Pipeline (SVG → JSON → PNG)

Input: Directory of SVG files (`IN_SVG_DIR`)
Output: JSON annotations and optional rendered PNGs

`svg_path` ∈ `IN_SVG_DIR` Parse SVG into polygon list $(\texttt{polys}, W, H)$ Fit polygons to canonical tangram templates Save piece parameters as JSON (`pos`, `angle`, `flip`, `scale`) rendering enabled Render shapes via geometry engine and save PNG aligned outline available Compute IoU between rendered union and outline **return** dataset (`JSON`, `PNG`, optional IoU logs)

A.2 TASK DESIGN DETAILS

We design four tasks that progressively increase geometric difficulty:

1. **pos-only**: fix GT (α, s), predict $\hat{\mathbf{p}}$;
2. **angle-only**: fix GT $(\mathbf{p}, s)$, predict $\hat{\alpha}$;
3. **size-only**: fix GT $(\mathbf{p}, \alpha)$, predict $\hat{s}$;
4. **two-piece arrangement**: fix each piece's (α, s), predict two positions $(\hat{\mathbf{p}}_1, \hat{\mathbf{p}}_2)$, and assess composition.

We also include a **joint** setting that predicts all three fields simultaneously to expose compounding errors.

Algorithm 3 Evaluation Pipeline (concise)

Input: `IN_DIR` (PNGs), `GT_DIR` (JSONs), `OUT_DIR`, model, mode; optional `geometry.py`, `overlay.py`.
Output: Pred JSONs, basic metrics (L2 / angle / size / IoU), GT/PRED/OVERLAY visualizations, optional IoU summary CSV.

$I \in$ IN_DIR $G \leftarrow$ pair JSON by stem (may be empty), two $\leftarrow$ IsTwoPiece(G) *Predict:* call OpenRouter with strict JSON; retry $\leq R$; parse & validate fields by `mode`, `two` *Merge:* update requested fields into G; if size updated then mirror to `scale` *Save:* write `out_json=stem_pred_{mode}.json` *Metrics (if G exists):* compute L2 / angle diff / size rel. (two-piece = per-piece + mean); print JSON *Viz:* prefer `geometry.py` to render GT/PRED/OVERLAY; else use internal renderer *Overlay:* save `overlay_tmp/stem/pred.json` (alias by GT basename) not deferred run `overlay.py` once for this sample (anchors / size_mode / dilate) deferred run `overlay.py` on `overlay_tmp/` to produce `iou_summary.csv`

A.3 Additional Ablation Studies

Table 4: Ablation on loop count and threshold (keep ICL = 15 fixed). Baseline IoU = 0.65.

Setting Number	Description	Loop	Threshold	IoU (final)
1	ICL + Loop	6	0.9	**0.9320**
7	ICL + Loop	4	0.9	0.9287
8	ICL + Loop	2	0.9	0.9291
9	ICL + Loop	6	0.5	0.8410
10	ICL + Loop	4	0.5	0.8609
11	ICL + Loop	2	0.5	0.7200
12	ICL + Loop	6	0.8	0.9310
13	ICL + Loop	6	0.7	0.9233
14	ICL + Loop	6	0.6	0.9063
15	ICL + Loop	8	0.9	0.9345
16	ICL + Loop	10	0.9	0.9258
17	ICL + Loop	12	0.9	0.9323

Table 5: Ablation on ICL window size (k), keep loop and threshold constant.

ICL (k)	Loop	Threshold	IoU (final)
15	8	0.90	0.9345
20	8	0.90	0.9311
25	8	0.90	0.9310

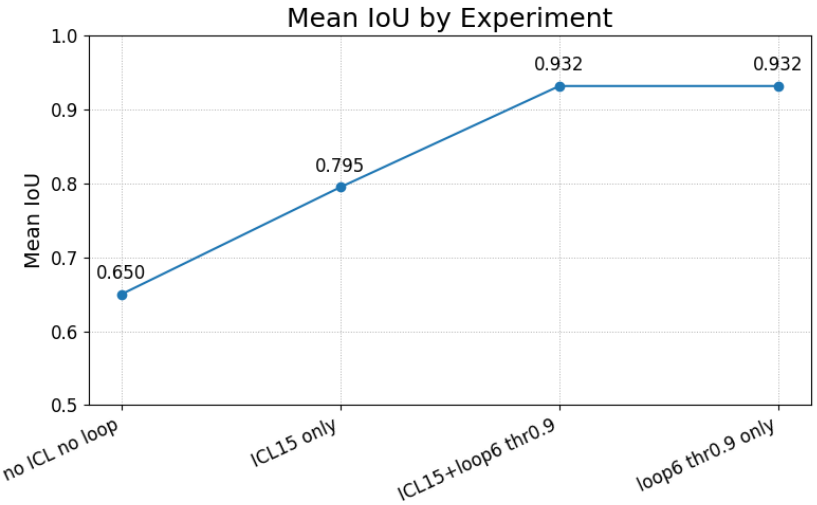


Figure 3: Mean IoU across ablations on the *medium triangle*. The test-time self-refinement loop (ICL + reward) yields the largest gain.

Loop Parameter Sensitivity. Table 4 compares different loop counts and acceptance thresholds under constant ICL (k=15). Performance quickly saturates after only a few refinement iterations, indicating that most geometric errors are corrected early in the process. Varying the acceptance threshold highlights its decisive role: comparing Settings 1 (τ=0.9, 0.9320) with 9 (τ=0.5, 0.8410), a steep IoU decline appears as the gating becomes more lenient. Higher thresholds act as a filter that stabilizes updates, while additional iterations beyond six offer diminishing returns.

B. THE USE OF LLMS

We use LLMs to check and correct grammar and spelling mistakes and to polish the sentences in our paper to make them more fluent.