GROUNDBENCH: MULTI-RESOLUTION POLYGON GROUNDING EXPOSES THE GEOMETRY GAP IN VISION–LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

A bounding box can locate the right object without describing its shape. This leaves an important question unanswered by box-based grounding scores: can a general-purpose vision–language model turn the same visual and linguistic evidence into a usable object region? We introduce GROUNDBENCH, a benchmark that asks models to express referred objects as polygons through their text-generation interface. It pairs 1,500 image–expression–referent triples across five vertex budgets, holding the questions fixed while varying output complexity. Evaluation reports filled-region overlap, coordinate-sequence completion, and polygon legality separately. Across four hosted models, more vertices do not reliably improve the output: performance degrades at the largest budget despite more faithful reference contours. Target preference can also remain high while conditional region quality is poor, showing that favouring the right object does not ensure a useful geometric description. Controlled interventions further reveal sensitivity to reasoning settings and misleading spatial language. These findings expose a geometric-output gap: locating an object does not ensure reliable explicit geometry. They identify improvement targets in region construction and output completion that a box score alone does not expose. GROUNDBENCH evaluates their combined outcome rather than latent boundary perception in isolation.

1 INTRODUCTION

Referring-expression grounding asks a model to locate the object described by a phrase. RefCOCO-family benchmarks (Yu et al., 2016; Mao et al., 2016), built on COCO (Lin et al., 2014), commonly evaluate this ability with bounding boxes. A successful box prediction establishes where an object is, but says little about how well the model can express its shape: different contours can share the same enclosing rectangle. This distinction matters when the desired output is an object region rather than a coarse location. We ask whether general-purpose vision–language models (VLMs) can turn a referring expression and image into a usable object boundary through their ordinary text-generation interface.

Boxes leave boundary placement largely untested. Dense masks express shape, but often rely on a dedicated decoder or an external segmentation model (Kirillov et al., 2023). Polygons provide a complementary test: their vertices specify a region while remaining directly expressible as text. They let us examine the geometric output of a general-purpose model without adding a task-specific segmentation head. We study this particular interface, not a replacement for mask-based grounding or a pure test of visual perception. Exact vertex budgets make output complexity explicit, so systems can be compared under the same geometric requirements.

The key design choice is to request an exact number of vertices, N. Holding the image, expression, and target object fixed while changing N tests whether additional output capacity produces a better region. A model must identify the referent, place and order boundary points, and complete the requested sequence. More vertices can approximate a contour more closely, but also introduce more opportunities for repetition, crossings, and incomplete output. A box score cannot reveal this trade-off because its output length and rectangular shape are fixed.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

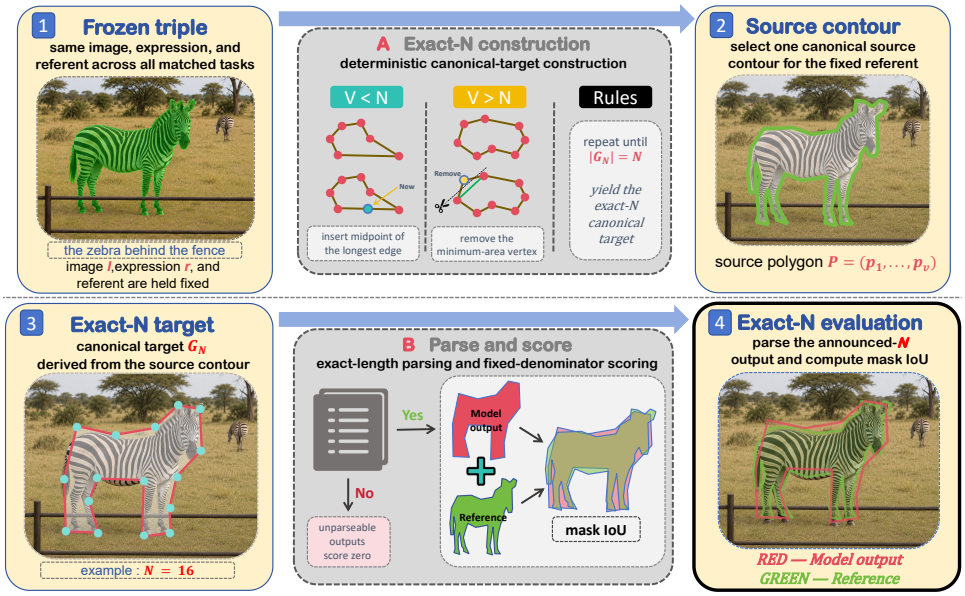

Figure 1: GROUNDBENCH freezes each image–expression–referent triple, derives five exact-budget targets, and scores exact-length outputs with a deterministic rasteriser.

We introduce GROUNDBENCH, a benchmark of 1,500 matched image–expression–referent triples, with 500 drawn from each of RefCOCO, RefCOCO+, and RefCOCOg. Each image contributes one question. Five targets, $N \in \{8, 16, 24, 32, 64\}$, are derived deterministically from the same selected annotation contour; Figure 1 shows the construction and evaluation pipeline. These budgets were fixed before model evaluation: the smaller budgets describe compact contours, while 64 tests a longer output with little additional reference fidelity over 32. Shared question identifiers (UIDs) support paired comparisons across all budgets. Region overlap is measured by intersection over union (IoU). Fixed scores count output failures as zero; parseable-only scores average completed coordinate lists. Legal-polygon coverage checks whether those lists also satisfy the geometric constraints.

The experiments reveal two limitations hidden by a box score. A prediction can favour the intended object yet describe its region poorly. Increasing the vertex budget can also reduce performance despite more faithful reference contours. Controlled interventions help interpret these failures: enabling Qwen's thinking mode improves completion and region quality, while false spatial cues are more disruptive than false colour cues. These findings motivate separate evaluation of target preference, region quality, and output validity.

Contributions. Our contribution is a controlled evaluation of reliable explicit geometric output beyond object localisation. Matched targets vary output complexity without changing the questions. Complementary scores distinguish region overlap from legal polygon construction, while interventions probe input sensitivity and output reliability. The findings concern the sampled COCO referents and tested configurations.

2 RELATED WORK

Referring and polygon grounding. RefCOCO-family benchmarks pair natural-language expressions with COCO instances (Kazemzadeh et al., 2014; Yu et al., 2016; Mao et al., 2016; Lin et al., 2014), and later variants extend the answer to segmentation masks, open phrases, or multiple targets (Kamath et al., 2021; Wu et al., 2020; Liu et al., 2023a). SeqTR and PVD (Zhu et al., 2022; Cheng et al., 2024) formulate localisation as coordinate generation. PolyFormer (Liu et al., 2023c) uses a dedicated model to generate variable-length polygons for referring segmentation. Earlier interactive and automatic contour methods also predict boundaries as vertex sequences (Castrejon

Table 1: GROUNDBENCH filled-region results (0–100). Every metric cell shows fixed / parseable-only scoring under IoU^2. IoU_p uses the source part and IoU_n is fidelity-normalised. Pooled and Mean average budgets and configurations; bold/underline mark best/second.

N	Configuration	$\mathbf{IoU^2}$	IoU_p	IoU_n	Acc@.1	Acc@.3	Acc@.5	Acc@.7	Acc@.9
$N = 8$	Doubao Seed-2.0-Pro	41.2/41.2	41.9/41.9	47.3/47.3	75.0/75.0	59.3/59.3	43.3/43.3	24.2/24.2	1.7/1.7
	Kimi K2.7 Code	38.2/38.2	38.6/38.6	43.6/43.6	83.8/83.8	63.1/63.1	33.8/33.8	10.1/10.1	0.1/0.1
	Qwen 3.7 Plus	48.2/48.4	49.1/49.2	55.4/55.6	81.3/81.7	71.1/71.4	56.0/56.2	30.3/30.4	1.9/1.9
	Gemini 3.5 Flash	**62.6/62.6**	**63.6/63.6**	**71.7/71.7**	**93.0/93.0**	**87.8/87.8**	**76.0/76.0**	**48.8/48.8**	**8.0/8.0**
$N = 16$	Doubao Seed-2.0-Pro	46.8/46.8	46.8/46.8	48.7/48.7	78.7/78.7	65.1/65.1	51.5/51.5	33.4/33.4	3.5/3.5
	Kimi K2.7 Code	37.3/37.3	37.3/37.3	38.8/38.8	82.3/82.3	60.0/60.0	32.9/32.9	10.1/10.1	0.1/0.1
	Qwen 3.7 Plus	49.7/49.9	49.7/49.8	51.6/51.8	81.4/81.7	71.5/71.8	58.1/58.3	33.3/33.4	2.2/2.2
	Gemini 3.5 Flash	**64.7/64.9**	**64.7/64.9**	**67.2/67.3**	**90.6/90.8**	**84.6/84.8**	**76.0/76.2**	**59.8/59.9**	**11.0/11.0**
$N = 24$	Doubao Seed-2.0-Pro	43.4/43.4	43.3/43.3	44.1/44.1	76.9/76.9	60.0/60.0	45.6/45.6	29.1/29.1	2.8/2.8
	Kimi K2.7 Code	35.9/35.9	35.8/35.8	36.5/36.5	81.8/81.8	57.9/57.9	29.7/29.7	7.2/7.2	0.3/0.3
	Qwen 3.7 Plus	49.1/49.3	49.1/49.3	49.9/50.1	82.1/82.5	71.3/71.6	56.6/56.8	32.4/32.5	1.7/1.7
	Gemini 3.5 Flash	**65.9/66.3**	**65.8/66.2**	**67.0/67.4**	**92.0/92.6**	**87.2/87.7**	**79.6/80.1**	**58.2/58.6**	**11.0/11.1**
$N = 32$	Doubao Seed-2.0-Pro	44.4/44.4	44.3/44.3	44.8/44.8	77.1/77.2	61.2/61.2	46.8/46.8	30.6/30.6	3.0/3.0
	Kimi K2.7 Code	33.7/34.0	33.7/33.9	34.1/34.3	78.5/79.1	55.3/55.7	26.8/27.0	6.5/6.6	0.2/0.2
	Qwen 3.7 Plus	48.6/48.8	48.6/48.8	49.1/49.2	**81.4**/81.7	71.3/71.5	56.7/56.9	30.9/31.0	1.9/1.9
	Gemini 3.5 Flash	**57.8/63.3**	**57.8/63.2**	**58.4/63.9**	80.8/**88.4**	**76.4/83.6**	**69.4/75.9**	**49.8/54.5**	**10.0/10.9**
$N = 64$	Doubao Seed-2.0-Pro	37.6/41.6	37.5/41.6	37.8/41.8	**69.3**/76.7	54.6/60.5	39.5/43.8	21.3/23.6	1.7/1.9
	Kimi K2.7 Code	24.4/30.2	24.4/30.2	24.5/30.4	58.3/72.1	38.5/47.6	18.9/23.4	5.1/6.3	0.0/0.0
	Qwen 3.7 Plus	**38.5**/44.9	**38.5**/44.9	**38.7**/45.2	68.8/80.3	**56.3**/65.7	42.5/49.6	22.7/26.5	1.1/1.2
	Gemini 3.5 Flash	37.7/**63.4**	37.7/**63.4**	37.9/**63.7**	52.8/**88.9**	49.4/**83.2**	**44.8/75.4**	**33.6/56.6**	**7.6/12.8**
Pooled	Doubao Seed-2.0-Pro	42.7/43.5	42.8/43.6	44.5/45.4	75.4/76.9	60.1/61.3	45.3/46.3	27.7/28.3	2.5/2.6
	Kimi K2.7 Code	33.9/35.3	34.0/35.4	35.5/37.0	76.9/80.1	55.0/57.2	28.4/29.6	7.8/8.1	0.1/0.2
	Qwen 3.7 Plus	46.8/48.4	47.0/48.5	48.9/50.5	79.0/81.6	68.3/70.5	54.0/55.7	29.9/30.9	1.8/1.8
	Gemini 3.5 Flash	**57.7/64.2**	**57.9/64.3**	**60.4/67.1**	**81.8/90.9**	**77.1/85.6**	**69.2/76.8**	**50.0/55.6**	**9.5/10.6**
	Mean	**45.3/47.9**	**45.4/48.0**	**47.3/50.0**	**78.3/82.4**	**65.1/68.7**	**49.2/52.1**	**28.9/30.7**	**3.5/3.8**

et al., 2017; Acuna et al., 2018; Ling et al., 2019; Peng et al., 2020). These methods use task-specific training. GROUNDBENCH is complementary: it evaluates general-purpose hosted VLMs without task-specific training, announces an exact output budget, and repeats the same referents across a multi-resolution curve.

Grounded multimodal models. Generalist systems increasingly emit boxes, points, or region references in text (Peng et al., 2024; Chen et al., 2023; You et al., 2024; Wang et al., 2023; 2024a; Bai et al., 2025), while grounded assistants connect conversation to dense region machinery (Lai et al., 2024; Li et al., 2024; Zhang et al., 2024; Rasheed et al., 2024). Recent work emits segmentation directly and extends grounding beyond RGB (Yao et al., 2026; Hegde et al., 2026; Zhao et al., 2025); GROUNDBENCH instead audits unmodified hosted APIs. Neither a short box nor a learned mask head establishes that a chat decoder can maintain a long cyclic boundary. Fixed-N polygons make allocation, exact-length compliance, and topology directly auditable between box grounding and specialist segmentation.

Coordinates as an output language. Once localisation is emitted as text (Chen et al., 2022), coordinate vocabulary, sequence length, and ordering become part of the task. Fixing N and reusing each referent across budgets makes the requested output complexity explicit and allows its benefits and failure costs to be compared. Spatial suites report persistent perceptual and relational weaknesses (Fu et al., 2024; Tong et al., 2024; Liu et al., 2023b; Chen et al., 2024); GROUNDBENCH instead requires a dense, metrically scored geometric artefact rather than a categorical judgement.

Polygon evaluation. Rasterising an ordered boundary into a scored region involves explicit choices about disconnected parts, self-intersections, clipping, and discrete integration. We therefore freeze one simple official policy and rescore the same predictions under alternate reference masks and continuous-area intersection, testing how sensitive the conclusions are to these evaluation choices.

Controlled benchmark extensions. Benchmark extensions such as ImageNet-V2 and ReaL (Recht et al., 2019; Beyer et al., 2020), SuperGLUE (Wang et al., 2019b;a), and MMLU-Pro (Hendrycks et al., 2021; Wang et al., 2024b) revisit what existing scores establish. Here the motivation is a limitation of the output representation: even an accurate rectangle does not specify an object's contour.

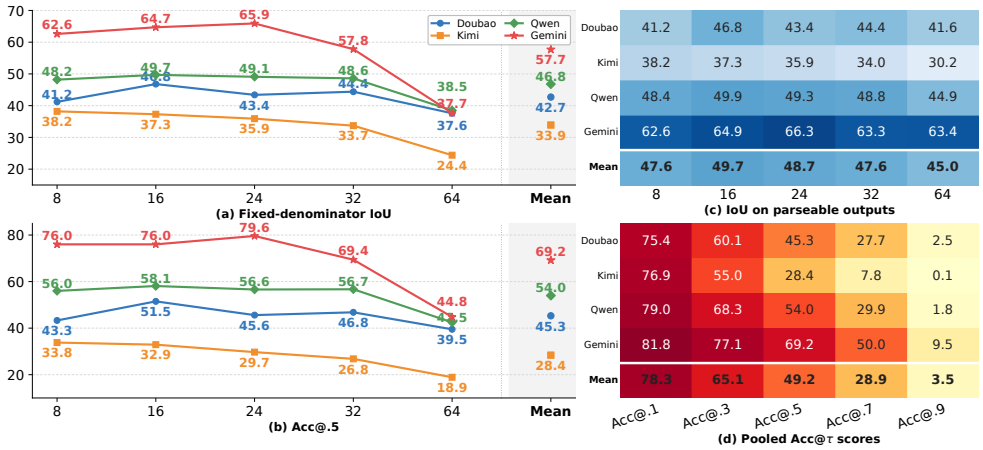

Figure 2: Summary of Table 1: fixed IoU^2, Acc@.5, parseable-only IoU, and pooled threshold accuracy.

Reusing the same referents across vertex budgets allows us to study this limitation without changing the image or expression distribution. We do not assume that the full RefCOCO distribution is saturated.

3 The GroundBench Benchmark

3.1 Task and Canonical Targets

For image I of width W and height H, expression r, and announced budget N, the model outputs $\widehat{P} = (p_1, \ldots, p_N)$ with $p_i = (x_i, y_i)$. The prompt requests one line of $2N$ integers in $\{0, \ldots, 999\}$. Filled scoring maps finite coordinates to $(xW/999,\ yH/999)$, clips them to the canvas, and rasterises the closed region. Strict scoring also requires distinct, in-range integer vertices and a simple positive-area polygon before raster clipping. Parse failures receive zero in both scores; illegal parsed polygons receive zero only in strict scoring. Start point and traversal direction are unrestricted because filled-mask IoU is invariant to cyclic shifts and reversal. A sensitivity audit reconstructs targets after shifting the sampling origin through eight offsets spaced at 12.5% of the contour.

For a multipart annotation, we select the part with the most listed vertices. This affects $148/1{,}500$ UIDs. Within that subset, the selected part covers 71.76% of the all-parts union on average and is the largest-area part in 85.14% of cases. Algorithm 1 and Figure 1 show how we construct G_N. Sparse contours gain vertices by repeatedly inserting the midpoint of the longest edge. Dense contours lose vertices through Visvalingam–Whyatt simplification (Visvalingam & Whyatt, 1993): remove the vertex with the smallest neighbour-triangle area until N remain. Ties use the lowest index.

This frozen representation is rescored against the unreduced part, all-parts union, alternate starting phases, and continuous-area integration. The parser accepts the last explicit or bracketed list only when it contains exactly $2N$ finite values. A compact compliant answer needs about 520 characters at the largest budget, below the smallest output allowance of $1{,}280$ tokens. This does not rule out truncation when a model emits extra text or fails to follow the format.

3.2 Cohort Construction and Rationale

The fixed sampling procedure selects 1,500 unique image–referent pairs: 500 each from RefCOCO, RefCOCO+, and RefCOCOg, one per image. Six frozen groups cover the 80 COCO categories. Distractor bins $D0/D1/D2/D3+$ denote zero, one, two, or at least three other non-crowd target-category instances; the cohort contains $705/276/519$ questions in $D1/D2/D3+$ and none in $D0$. Contour complexity is $Q = \ell^2/(4\pi A)$, where ℓ and A are source-contour perimeter and area; relative scale RS is target-box area divided by image area. $Q1/Q2/Q3$ and $RS1/RS2/RS3$ are

Table 2: **Prompt cues and image evidence.** Fixed IoU (×100); B_{cue} and image controls are same-UID matched.

Cue	$N=8$	$N=16$	$N=24$	Mean
S (true spatial)	45.5	47.2	48.5	**47.1**
C (true colour)	45.3	46.5	45.4	**45.7**
SC (both true)	44.7	46.9	45.8	**45.8**
WC (wrong colour)	38.5	39.4	39.6	**39.2**
WS (wrong spatial)	22.6	24.6	25.8	**24.3**
WSC (both wrong)	20.6	22.6	23.2	**22.1**
No image	14.1	13.2	13.3	**13.5**
Half resolution	42.4	41.6	42.5	**42.2**
Greyscale	46.8	48.2	47.5	**47.5**
B_{cue} (none)	**47.8**	**49.3**	**48.2**	**48.4**

Table 3: **Thinking, decoding, and repeatability.** Fixed IoU (×100); italics report parseable-only scores.

Condition	$N=8$	$N=16$	$N=24$	Mean
Thinking depth				
Max thinking	48.2	49.7	49.1	**49.0**
parse.	48.4	49.9	49.3	49.2
4,096 tokens	48.0	49.0	47.4	**48.1**
parse.	48.7	49.2	49.7	49.2
Thinking off	20.5	14.7	17.0	**17.4**
parse.	27.0	18.7	19.4	21.7
Decoding & repeatability				
Temperature 0	48.2	49.7	49.1	**49.0**
Temperature 0.5	47.2	48.6	50.9	**48.9**
Temperature 1.0	45.9	50.5	49.2	**48.5**
Independent repeat	46.7	49.1	49.9	**48.6**
B_{full} (Qwen max-think)	**48.2**	**49.7**	**49.1**	**49.0**

low/middle/high bins under globally frozen thresholds. Cohort sampling preceded all model outputs.

For phrase word count L and source vertex count V, the composite difficulty score is

$$s = w_D + w_Q + w_{RS} + \mathbf{1}[L \geq 6] + \mathbf{1}[L \geq 10] + \mathbf{1}[V \geq 25] + \mathbf{1}[V \geq 50], \tag{1}$$

Here $w_D = (0, 0, 1, 2)$ for $(D0, D1, D2, D3+)$, $w_Q = (0, 1, 2)$ for $(Q1, Q2, Q3)$, and $w_{RS} = (2, 1, 0)$ for $(RS1, RS2, RS3)$. This yields 436 easy ($s \leq 2$), 554 medium ($3 \leq s \leq 4$), and 510 hard ($s \geq 5$) questions. The components are correlated, so we report both the individual axes and the summary label.

3.3 Scoring

Reported IoU and accuracy scores use a 0–100 scale; per-question z_i and thresholds τ lie in $[0, 1]$. Let $R(\cdot)$ rasterise a polygon onto the original canvas. The indicator $\mathbf{1}[\rho_i]$ is one for an exact-length parse. On failure, $z_i = 0$ without evaluating an absent polygon. For threshold τ, the per-question score and official aggregates over planned cohort S with parsed subset S_{p} are

$$z_i = \mathbf{1}[\rho_i]\ \text{IoU}\big(R(\widehat{P}_i), R(G_{N,i})\big), \tag{2}$$

$$\text{IoU}_{\text{fix}} = \tfrac{100}{|S|}\sum_{i \in S} z_i, \qquad \text{IoU}_{\text{parse}} = \tfrac{100}{|S_{\text{p}}|}\sum_{i \in S_{\text{p}}} z_i, \tag{3}$$

$$\text{Acc@}\tau = \tfrac{100}{|S|}\sum_{i \in S} \mathbf{1}[z_i \geq \tau]. \tag{4}$$

The fixed denominator counts every planned question. We also report parseable-only IoU (Equation (3)), since a model can produce accurate regions on a small fraction of questions while failing the rest. Filled scoring asks whether an emitted sequence still covers a useful region; the separate legality-enforced audit retains the fixed denominator but assigns zero unless the parsed coordinates form a legal polygon. The alternate-mask columns of Table 1 replace the canonical target in Equation (2): with A_i the mask of the unreduced selected source part, set $z_i^{\text{p}} = 0$ on a parse failure and otherwise

$$z_i^{\text{p}} = \text{IoU}(R(\widehat{P}_i), A_i), \quad \text{IoU}_{\text{n}} = 100\,\text{IoU}_{\text{p}}/C_N. \tag{5}$$

where IoU_{p} aggregates z_i^{p} as in Equation (3) and the construction fidelity $C_N = 100\,\text{mean}_i\,\text{IoU}(R(G_{N,i}), A_i)$ is $88.6/96.2/98.3/99.0/99.4$ for $N = 8/16/24/32/64$. Higher C_N means less approximation loss. IoU_{n} is the descriptive representation-fidelity ratio $100\,\text{IoU}_{\text{p}}/C_N$; it is neither a probability nor a mathematical upper bound and may exceed 100. The superscripts in IoU^0, IoU^1, and IoU^2 identify recovery stages, not powers. Stage 0 is the archived primary-run endpoint; stages 1 and 2 add external recovery for unparseable outputs. Main-table scores use stage 2. Setup defines the separate best-effort view.

Algorithm 1 Canonical target construction G_N

Require: source contour $C = (c_1, \ldots, c_V)$, budget N
1: **while** $|C| < N$ **do**
2: $k \leftarrow \arg\max_j \|c_{j+1} - c_j\|$ {longest edge, cyclic}
3: insert $(c_k + c_{k+1})/2$ after c_k
4: **end while**
5: **while** $|C| > N$ **do**
6: $k \leftarrow \arg\min_j \frac{1}{2}|(c_{j-1} - c_j) \times (c_{j+1} - c_j)|$
7: remove c_k {smallest neighbour-triangle area}
8: **end while**
9: **return** $G_N \leftarrow C$

3.4 Target Preference and Region Quality

Diagnostics compare each parsed prediction with the all-parts target mask U_i and each same-category distractor mask $D_{i,d}$. Here d indexes distractors. Unparsed outputs receive $\mathrm{Sel}_i = 0$. Target preference and conditional region IoU for parsed outputs are

$$\mathrm{Sel}_i = \mathbf{1}\Big[\mathrm{IoU}(R(\widehat{P}_i), U_i) > \max_d \mathrm{IoU}(R(\widehat{P}_i), D_{i,d})\Big], \tag{6}$$

$$\mathrm{Con} = 100\,\mathrm{mean}\{\mathrm{IoU}(R(\widehat{P}_i), U_i) : \mathrm{Sel}_i = 1\}. \tag{7}$$

The reported preference rate is $\mathrm{Sel} = 100|S|^{-1}\sum_{i\in S}\mathrm{Sel}_i$, including unparseable questions as zero. Con measures region overlap conditional on target preference. It is not a direct boundary-distance score. This is not verified instance detection: it has no absolute overlap threshold, confidence ranking, margin, or cross-category matching. We additionally audit duplicate points, zero area, self-intersection, alternate reference regions, and continuous-area IoU on already-legal polygons. Paired comparisons match question identifiers; each image intervention uses its own matched control.

4 Experiments

4.1 Setup

We evaluate four API-served models: Doubao Seed-2.0-Pro, Kimi K2.7 Code, Qwen 3.7 Plus, and Gemini 3.5 Flash. Each answers the same 1,500 questions at all five vertex budgets. Doubao uses the provider's default reasoning behaviour, and Kimi uses a fixed-thinking endpoint. For Qwen, *Qwen-max* names the thinking-enabled configuration used in our runs. *Qwen-off* disables thinking at the same model endpoint. Qwen-off is a configuration ablation, not a fifth model. *Gemini-high*, shortened to Gemini in the tables, uses high thinking and high media resolution. Table S5 lists exact API identifiers, temperatures, and output limits.

Main benchmark generations were collected on 10–17 July 2026. Images preserve aspect ratio, with the longest side capped at 1,024 pixels. Results describe these dated configurations, not all models from a provider. IoU^0 denotes the archived primary-run endpoint, which can include internal format-recovery calls. The headline IoU^2 adds at most two external recovery stages for still-unparseable answers and retains the first usable output. The best-effort diagnostic permits one further Gemini stage. These labels identify reporting stages, not equal numbers of physical API calls. We compare complete configured pipelines and keep their reporting views separate.

4.2 Main Results

Table 1 and Figure 2 address two questions: how much region quality is obtained from a polygon output, and whether more vertices improve it. Fixed IoU includes failed outputs; parseable-only IoU describes the subset that supplies an exact-length coordinate list. Reading them together prevents successful examples from hiding a model's completion failures.

Region quality beyond a box. More expressive outputs do not automatically produce better regions. Gemini leads the pooled polygon ranking, followed by Qwen-max, Doubao, and Kimi. When Gemini's direct polygons and predicted rectangles are scored against the same contour target, their

Table 4: Target preference and conditional region IoU by budget in the best-effort view. Sel is target preference, Con is target IoU conditional on Sel, and IoU uses the fixed denominator.

	N=8			N=16			N=24			N=32			N=64		
Config	Sel	Con	IoU	Sel	Con	IoU	Sel	Con	IoU	Sel	Con	IoU	Sel	Con	IoU
Doubao	76.9	53.9	**41.2**	80.5	57.7	**46.8**	78.7	54.5	**43.4**	80.0	54.9	**44.4**	71.2	52.0	**37.6**
Kimi	85.5	44.6	**38.2**	84.7	43.3	**37.3**	83.5	42.1	**35.9**	80.9	40.8	**33.7**	60.7	38.5	**24.4**
Qwen-off	60.9	33.8	**20.5**	62.6	23.4	**14.7**	68.7	24.2	**17.0**	71.3	25.6	**18.8**	3.8	25.8	**1.0**
Qwen-max	85.2	57.5	**48.2**	86.5	57.6	**49.7**	85.7	57.4	**49.1**	85.9	56.7	**48.6**	70.5	54.1	**38.5**
Gemini	94.2	68.1	**62.6**	92.4	70.8	**64.7**	93.4	71.3	**65.9**	86.0	70.1	**59.5**	62.2	67.7	**41.9**
Mean	**80.5**	**51.6**	**42.1**	**81.3**	**50.6**	**42.7**	**82.0**	**49.9**	**42.2**	**80.8**	**49.6**	**41.0**	**53.7**	**47.6**	**28.7**

pooled IoU is 57.7 and 57.3. This near-tie makes the budget curve important: a pooled score can hide where added resolution helps and where output failures offset that benefit. Box-reference and contour-reference scores use different targets. Their difference is not a direct measure of lost visual understanding.

More vertices do not guarantee better output. All four configurations perform worse at $N = 64$ than at $N = 16$, despite higher region fidelity of the target. Gemini illustrates why both scoring views matter: at $N = 64$, fixed IoU is 37.7, while parseable-only IoU is 63.4. Fixed IoU equals parseable-only IoU times parse coverage. For Gemini, $63.4 \times 0.594 \approx 37.7$. The 25.7-point separation therefore reflects the zero contribution of missing outputs under this metric; it does not reveal how accurate those missing outputs would have been. Parsing is not the only obstacle: legal-polygon coverage is 7.5% and legality-enforced IoU is 4.2. Rounded references can also violate the contract: 1,416/1,500 pass at $N = 64$ (Section G.2). Thus strict failures cannot all be assigned solely to the model.

Longer output can hurt without adding contour detail. Reference fidelity rises only from 99.0 at $N = 32$ to 99.4 at 64. The largest budget therefore primarily tests a longer output contract. On stable-boundary subsets, additional vertices subdivide existing edges rather than add source-contour detail. Every configuration still loses fixed IoU from 32 to 64 (Table S32). This links the decline to output requirements, although rounding and rasterisation can slightly alter the scored regions. It does not isolate an internal mechanism. Eight target starting phases preserve the broad high-budget reversal, but direct-versus-box gains depend on phase.

On 12 matched Qwen-max examples at $N = 32$ and 64, boundary assistance improves region overlap more than a target box, while coordinate copying succeeds on 11 examples at each budget (Section F.2). Residual failures suggest that converting visible geometry into a valid sequence remains difficult. The conditions change the available information, so this small diagnostic does not separate latent abilities.

Recovery changes delivered utility. Table 6 separates the primary-run endpoint from the final stage-2 result. Recovery adds 8.15 filled-IoU points for Gemini, versus 1.02–2.65 for the other three configurations. It recovers missing outputs without replacing usable answers. Stage 0 can already include internal format repair; neither column measures strictly first-call reliability. The headline therefore describes utility under the configured recovery policy.

4.3 Ablation Study

We next ask which changes affect polygon production while keeping the original target fixed. A 13-condition study varies Qwen's reasoning settings, visual input, prompt cues, and decoding. Most conditions use all 1,500 questions; no-image, half-resolution, and greyscale tests use matched 500-question subsets. Gemini repeats these three visual interventions and the wrong-spatial and wrong-colour tests; other contrasts are Qwen-specific. Tables 2 and 3 and Figure 3 report the controlled comparisons. The target-preference analysis in Table 4 uses the separately labelled best-effort outputs.

Thinking affects completion and geometry. Qwen-max outperforms Qwen-off at every vertex budget. The gain includes both more parseable answers and better overlap on questions parsed by both configurations. The latter comparison holds the evaluated questions fixed, so it cannot be explained by recovering additional easy-to-parse cases. The setting therefore affects more than format

Table 5: Pooled fixed-IoU slices. Ea/Me/Ha are easy/medium/hard; An./Ind./Out./Per./Veh. are animal/indoor/outdoor/person/vehicle. Difficulty, D, Q, and RS use IoU^2; semantic groups use best-effort. Bold/underline mark best/second.

	Difficulty			**Distr.** D			**Compl.** Q			**Scale** RS			**Semantic group**					
Cfg.	Ea	Me	Ha	$D1$	$D2$	$D3+$	$Q1$	$Q2$	$Q3$	$RS1$	$RS2$	$RS3$	An.	Food	Ind.	Out.	Per.	Veh.
Doubao	50.7	43.0	35.4	45.9	40.5	39.4	46.5	44.5	35.9	37.0	42.9	49.8	42.8	41.9	39.9	37.6	46.7	44.9
Kimi	42.5	33.7	26.8	36.3	33.1	31.0	39.6	33.8	26.5	29.3	34.2	39.6	32.3	34.1	35.0	30.6	33.9	36.5
Qwen-off	17.5	13.7	12.5	15.9	13.8	12.7	15.3	14.0	13.6	12.6	14.6	16.5	17.3	13.9	11.9	12.8	15.6	14.1
Qwen-max	55.6	46.4	39.8	50.3	45.7	42.6	52.2	46.9	39.6	42.9	47.3	51.5	46.7	48.1	45.5	42.7	46.4	50.3
Gemini	**65.1**	**58.1**	**51.0**	**61.8**	**57.6**	**52.3**	**63.4**	**57.8**	**50.2**	**54.2**	**58.8**	**61.4**	**63.4**	**59.3**	**53.6**	**59.4**	**57.4**	**61.3**
Mean	**46.3**	**39.0**	**33.1**	**42.0**	**38.2**	**35.6**	**43.4**	**39.4**	**33.2**	**35.2**	**39.6**	**43.8**	**40.5**	**39.5**	**37.2**	**36.6**	**40.0**	**41.4**

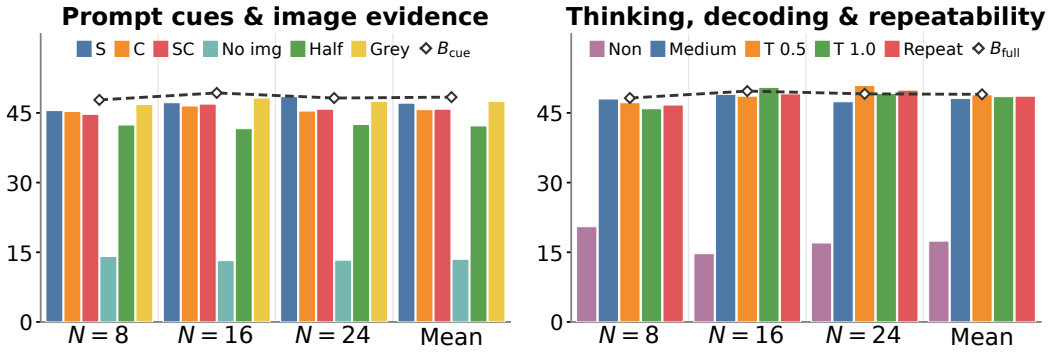

Figure 3: Controlled fixed-IoU ablations. Diamonds mark controls; circles mark matched-image controls.

compliance. It is the largest tested configuration contrast that preserves the visual and linguistic input, but it does not isolate an internal reasoning mechanism. For model development, this result motivates checking completion and region quality separately when changing inference settings.

Conflicting language can override useful visual evidence. Adding a false spatial cue harms the original-target score more than adding a false colour cue. Corresponding true cues yield no reliable gain, so the harm is not explained simply by adding text. The combined false cues have a sub-additive effect. This experiment measures retention of the original referent under contradictory information; it does not define a new correct target. The implication is to evaluate sensitivity to misleading language alongside performance on clean expressions, rather than infer grounding reliability from clean-input accuracy alone.

More image detail is not the main remedy in these tests. Removing the image causes a much larger loss than halving its resolution or removing colour. Residual no-image overlap may reflect language and dataset regularities. The small greyscale effect and larger false-colour cue effect are compatible: one removes visual colour, while the other adds conflicting language. These results concern the tested transformations; they do not establish that resolution or colour is generally unimportant.

Choosing the referent and describing its region are different problems. At $N = 8$, Kimi's prediction overlaps the target more than any same-category distractor on 85.5% of questions, yet its conditional region IoU is only 44.6. Qwen-max has a similar preference rate, 85.2%, but higher conditional IoU, 57.5. This comparison exposes different improvement needs. A model may favour the intended object while representing its region poorly. Improving target choice alone may therefore leave much of the region error unresolved. These diagnostics use best-effort outputs; preference is a relative-overlap event, not verified detection, and conditional IoU is a region measure rather than direct boundary accuracy.

Table 5 slices the frozen answers by composite difficulty, its three geometric components, and six semantic groups, pooled over the five budgets.

Table 6: Primary-run (stage 0) and final (stage 2) scores, pooled over five budgets. Stages are not physical call counts. Strict IoU assigns zero to illegal outputs. Gemini values and Kimi's stage-0 strict value retain the archived reports; see Table S39.

	Filled IoU		Strict IoU	
Configuration	Stage 0	Stage 2	Stage 0	Stage 2
Doubao	40.00	42.65	29.78	30.69
Kimi	32.48	33.90	27.98	28.48
Qwen-max	45.81	46.83	30.85	31.38
Gemini	49.60	57.75	36.76	43.87

Which examples remain difficult? Every configuration declines from the easy to hard subsets. Similar losses among parseable answers indicate that these differences are not explained by output completion alone. More distractors and complex contours are associated with lower IoU, whereas larger targets are associated with higher IoU; no semantic group is consistently easiest. These correlated slices describe where errors occur, not their causes. Alternate-mask mean shifts are at most 1.1 points. Continuous-area scoring closely matches raster scoring on already-legal polygons; repairing invalid polygons is a different policy. The principal ordering is unchanged.

Adjusting inference settings has smaller effects. Reducing Qwen's thinking budget to 4,096 stays close to the maximum setting. Temperature effects change direction across budgets, and one independent repeat remains close to the reference run. These differences are smaller than switching thinking off or removing the image. One repeat cannot establish low variance or an optimal decoding setting.

Region prediction and output reliability. A mask-to-polygon pipeline predicts one mask per UID, then deterministically converts that same mask to $N = 16, 32, 64$ vertices (Section F.3). From $N = 16$ to 64, its fixed IoU falls from 65.85 to 63.63, versus 64.74 to 37.66 for Gemini direct polygons. At $N = 64$, legal coverage is 92.9% versus 7.53%. This contrast supports an output-construction limitation alongside region-prediction difficulty. This comparison concerns complete pipelines. Deterministic conversion supplies boundary tracing and ordered vertices; the difference does not isolate internal visual or geometric abilities. Unlike direct coordinate generation, increasing N here does not lengthen the model response. Structured JSON and generation in four shorter segments also retain substantial failures (Section F.1); changing the output syntax alone need not resolve the bottleneck.

Scope. The benchmark measures explicit geometric output, not latent boundary perception alone. Results concern sampled COCO referents and dated hosted configurations; item-level resampling does not capture provider drift. Target phase affects absolute scores, and contamination is not ruled out.

5 Conclusion

GroundBench exposes a gap between object localisation and reliable explicit geometric output. More vertices do not reliably improve regions: completion and legality deteriorate at long outputs, including when extra vertices add no source-contour detail. Target preference can also remain high despite poor region quality. The mask-to-polygon comparison motivates evaluating alternative output pipelines, while the recovery contrast distinguishes initial-run results from final delivered utility. These findings concern geometric-output reliability, not a general inability to perceive shape.

AI Use Statement

Generative AI tools were used to aid and polish the writing of this manuscript. The authors have reviewed the final manuscript and take full responsibility for its content.

REFERENCES

David Acuna, Huan Ling, Amlan Kar, and Sanja Fidler. Efficient interactive annotation of segmentation datasets with polygon-RNN++. In *CVPR*, 2018.

Shuai Bai et al. Qwen2.5-vl technical report. *arXiv:2502.13923*, 2025.

Lucas Beyer, Olivier J. Hénaff, Alexander Kolesnikov, Xiaohua Zhai, and Aäron van den Oord. Are we done with ImageNet? *arXiv:2006.07159*, 2020.

Lluis Castrejon, Kaustav Kundu, Raquel Urtasun, and Sanja Fidler. Annotating object instances with a polygon-RNN. In *CVPR*, 2017.

Boyuan Chen, Zhuo Xu, Sean Kirmani, Brian Ichter, Danny Driess, Pete Florence, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. SpatialVLM: Endowing vision-language models with spatial reasoning capabilities. In *CVPR*, 2024.

Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. Shikra: Unleashing multimodal llm's referential dialogue magic. *arXiv:2306.15195*, 2023.

Ting Chen, Saurabh Saxena, Lala Li, David J. Fleet, and Geoffrey Hinton. Pix2seq: A language modeling framework for object detection. In *ICLR*, 2022.

Zesen Cheng, Kehan Li, Peng Jin, Xiangyang Ji, Li Yuan, Chang Liu, and Jie Chen. Parallel vertex diffusion for unified visual grounding. In *AAAI*, volume 38, pp. 1326–1334, 2024.

Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A. Smith, Wei-Chiu Ma, and Ranjay Krishna. BLINK: Multimodal large language models can see but not perceive. In *ECCV*, 2024.

Sandesh Hegde, Jaison Saji Chacko, Debarshi Banerjee, and Uma Mahesh. GenSeg-R1: RL-driven vision-language grounding for fine-grained referring segmentation. *arXiv:2602.09701*, 2026.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *ICLR*, 2021.

Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. MDETR: Modulated detection for end-to-end multi-modal understanding. In *ICCV*, 2021.

Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. ReferItGame: Referring to objects in photographs of natural scenes. In *EMNLP*, 2014.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, et al. Segment anything. In *ICCV*, 2023.

Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. LISA: Reasoning segmentation via large language model. In *CVPR*, 2024.

Zhaowei Li, Qi Xu, Dong Zhang, Hang Song, Yiqing Cai, Qi Qi, Ran Zhou, Junting Pan, Zefeng Li, Vu Tu, et al. Groundinggpt: Language enhanced multi-modal grounding model. In *ACL*, 2024.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In *ECCV*, 2014.

Huan Ling, Jun Gao, Amlan Kar, Wenzheng Chen, and Sanja Fidler. Fast interactive object annotation with curve-GCN. In *CVPR*, 2019.

Chang Liu, Henghui Ding, and Xudong Jiang. GRES: Generalized referring expression segmentation. In *CVPR*, 2023a.

Fangyu Liu, Guy Emerson, and Nigel Collier. Visual spatial reasoning. *Transactions of the Association for Computational Linguistics*, 11:635–651, 2023b.

Jiang Liu, Hui Ding, Zhaowei Cai, Yuting Zhang, Ravi Kumar Satzoda, Vijay Mahadevan, and R. Manmatha. Polyformer: Referring image segmentation as sequential polygon generation. In *CVPR*, pp. 18653–18663, 2023c.

Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L. Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *CVPR*, 2016.

Sida Peng, Wen Jiang, Huaijin Pi, Xiuli Li, Hujun Bao, and Xiaowei Zhou. Deep snake for real-time instance segmentation. In *CVPR*, 2020.

Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shuming Ma, and Furu Wei. KOSMOS-2: Grounding multimodal large language models to the world. In *ICLR*, 2024.

Hanoona Rasheed, Muhammad Maaz, Sahal Shaji, Abdelrahman Shaker, Salman Khan, Hisham Cholakkal, Rao M. Anwer, Eric Xing, Ming-Hsuan Yang, and Fahad S. Khan. Glamm: Pixel grounding large multimodal model. In *CVPR*, 2024.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do ImageNet classifiers generalize to ImageNet? In *ICML*, 2019.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? Exploring the visual shortcomings of multimodal LLMs. In *CVPR*, 2024.

Maheswari Visvalingam and J. Duncan Whyatt. Line generalisation by repeated elimination of points. *The Cartographic Journal*, 30(1):46–51, 1993.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. SuperGLUE: A stickier benchmark for general-purpose language understanding systems. In *NeurIPS*, 2019a.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *ICLR*, 2019b.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv:2409.12191*, 2024a.

Wenhai Wang, Zhe Chen, Xiaokang Chen, Jiannan Wu, Xizhou Zhu, Gang Zeng, Ping Luo, Tong Lu, Jie Zhou, Yu Qiao, and Jifeng Dai. Visionllm: Large language model is also an open-ended decoder for vision-centric tasks. In *NeurIPS*, 2023.

Yubo Wang, Xueguang Ma, Ge Zhang, et al. MMLU-Pro: A more robust and challenging multi-task language understanding benchmark. In *NeurIPS*, 2024b.

Chenyun Wu, Zhe Lin, Scott Cohen, Trung Bui, and Subhransu Maji. PhraseCut: Language-based image segmentation in the wild. In *CVPR*, 2020.

Yuan Yao, Qiushi Yang, Humen Zhong, Jiangning Wei, Yifang Men, Shuai Bai, Miaomiao Cui, and Zhibo Yang. Qwen3-VL-Seg: Unlocking open-world referring segmentation with vision-language grounding. *arXiv:2605.07141*, 2026.

Haoxuan You, Haotian Zhang, Zhe Gan, Xianzhi Du, Bowen Zhang, Zirui Wang, Liangliang Cao, Shih-Fu Chang, and Yinfei Yang. Ferret: Refer and ground anything anywhere at any granularity. In *ICLR*, 2024.

Licheng Yu, Patrick Poirson, Shan Yang, Alexander C. Berg, and Tamara L. Berg. Modeling context in referring expressions. In *ECCV*, pp. 69–85, 2016. doi: 10.1007/978-3-319-46475-6_5.

Ao Zhang, Yuan Yao, Wei Ji, Zhiyuan Liu, and Tat-Seng Chua. NExT-Chat: An LMM for chat, detection and segmentation. In *ICML*, 2024.

Tianyi Zhao, Jiawen Xi, Linhui Xiao, Junnan Li, Xue Yang, Maoxun Yuan, and Xingxing Wei. RGBT-GroundBench: Visual grounding beyond RGB in complex real-world scenarios. *arXiv:2512.24561*, 2025.

Chaoyang Zhu, Yiyi Zhou, Yunhang Shen, Gen Luo, Xingjia Pan, Mingbao Lin, Chao Chen, Liujuan Cao, Xiaoshuai Sun, and Rongrong Ji. Seqtr: A simple yet universal network for visual grounding. In *ECCV*, 2022.

APPENDIX

This appendix explains what drives polygon-grounding errors and how the results depend on the evaluation protocol. We first document the benchmark and scoring rules. We then examine model comparisons, input interventions, target preference, and geometric validity. Qualitative examples appear at the end. Throughout, fixed scores include every planned question and assign zero to an unparseable answer. Parseable-only scores use only exact-length finite coordinate lists; such lists need not form legal polygons.

A BENCHMARK COMPOSITION, SETUP, AND PROTOCOL

Scale and terminology. Reported IoU, accuracy, and coverage use the 0–100 scale. Per-question IoU and thresholds τ use $[0, 1]$. Temperatures, rank correlations, and other quantities retain their stated units. The historical run label *Formal* corresponds to the primary-run endpoint (IoU^0), and the recovery stages A01/A02 to IoU^1/IoU^2.

Two scoring views. Headline comparisons use the same external recovery-stage allowance. IoU^2 starts from the archived primary-run endpoint and permits A01 and A02 only for still-unparseable answers. Those stages send the image and expression again in fresh requests. Primary-run records can themselves contain internal format-recovery attempts, so this is not a three-API-call guarantee. Gemini reaches 90.00 parse coverage and 57.75 pooled fixed IoU under the reported IoU^2 view.

The separate best-effort archive gives Gemini one additional retry, A03. Its parse coverage, fixed IoU, and parseable-only IoU are 92.00, 58.93, and 64.05. The other configurations are unchanged. These extra-retry results are diagnostics and never replace the headline ranking. Each diagnostic table identifies its scoring view.

A.1 QUESTION DISTRIBUTION

The five full benchmark files share the same 1,500 question identities. Therefore the counts in Tables S1, S2 and S4 count distinct questions rather than question–resolution pairs. Construction samples the unique-image pool with seed 20260709, chooses one referring expression per image with seed 20260710, and performs the balanced 500-per-source selection with seed 20260711.

Table S1: Dataset and referring-expression split distribution.

Source	val	test	testA	testB	Total
RefCOCO+	236	0	61	203	500
RefCOCO	252	194	11	43	500
RefCOCOg	345	155	0	0	500
Total	833	349	72	246	1,500

Table S2: Semantic category-group quotas. Availability caps prevent exact equality for outdoor/accessory and vehicle.

Source	Animal	Food/kitchen	Indoor/object	Outdoor/accessory	Person	Vehicle	Total
RefCOCO+	100	100	99	39	99	63	500
RefCOCO	99	99	98	42	98	64	500
RefCOCOg	89	89	89	57	88	88	500
Total	288	288	286	138	285	215	1,500

The frozen label-to-group mapping used by these quotas is listed in Table S3.

A.2 EXACT DIFFICULTY HEURISTIC

The three component axes are frozen metadata. D counts *other* non-crowd instances of the target's COCO category, excluding the target itself: $D0, D1, D2, D3+$ mean zero, one, two, or at least three

Table S3: Frozen mapping from the 80 COCO labels to six semantic groups.

Group	COCO category labels
Animal	bear, bird, cat, cow, dog, elephant, giraffe, horse, sheep, zebra
Food/kitchen	apple, banana, bottle, bowl, broccoli, cake, carrot, cup, donut, fork, hot dog, knife, orange, pizza, sandwich, spoon, wine glass
Indoor/object	bed, book, cell phone, chair, clock, couch, dining table, hair drier, keyboard, laptop, microwave, mouse, oven, potted plant, refrigerator, remote, scissors, sink, teddy bear, toaster, toilet, toothbrush, tv, vase
Outdoor/accessory	backpack, baseball bat, baseball glove, bench, fire hydrant, frisbee, handbag, kite, parking meter, skateboard, skis, snowboard, sports ball, stop sign, suitcase, surfboard, tennis racket, tie, traffic light, umbrella
Person	person
Vehicle	airplane, bicycle, boat, bus, car, motorcycle, train, truck

Table S4: Same-class distractor and composite heuristic difficulty distributions. The empty $D0$ bin is a property of this selected release, not of all RefCOCO data.

Source	$D0$	$D1$	$D2$	$D3+$	Easy	Medium	Hard
RefCOCO+	0	225	62	213	156	186	158
RefCOCO	0	197	85	218	159	179	162
RefCOCOg	0	283	129	88	121	189	190
Total	0	705	276	519	436	554	510

same-class distractors. Since this release contains no $D0$ question, no $D0$ model score exists. Shape complexity is the selected source contour's

$$Q = \frac{P^2}{4\pi A}, \tag{S1}$$

where P and A are contour perimeter and area. Relative size is $RS = A_{\mathrm{bbox}}/(WH)$. The exact globally frozen cut points are

$$\begin{aligned} Q1 &: Q < 1.84479414323066, \\ Q2 &: 1.84479414323066 \le Q < 2.96810193360072, \\ Q3 &: Q \ge 2.96810193360072, \end{aligned} \tag{S2}$$

$$\begin{aligned} RS1 &: RS < 0.10942010120738635, \\ RS2 &: 0.10942010120738635 \le RS < 0.21650858507462686, \\ RS3 &: RS \ge 0.21650858507462686. \end{aligned} \tag{S3}$$

They were computed on the complete RefCOCO/RefCOCO+/RefCOCOg validation/test candidate pool, not fitted to the selected 1,500 questions or to model scores. The half-open intervals make boundary assignment deterministic. Larger Q means a relatively elongated or intricate contour, while $RS1$ contains the smallest targets. For word count L and source-contour vertex count V, the released metadata score is

$$\begin{aligned} s = w_D + w_Q + w_{RS} + \mathbf{1}[L \ge 6] + \mathbf{1}[L \ge 10] \\ + \mathbf{1}[V \ge 25] + \mathbf{1}[V \ge 50]. \end{aligned} \tag{S4}$$

Weights are $w_D = (0, 0, 1, 2)$ for $(D0, D1, D2, D3+)$, $w_Q = (0, 1, 2)$ for $(Q1, Q2, Q3)$, and $w_{RS} = (2, 1, 0)$ for $(RS1, RS2, RS3)$. Easy, medium, and hard mean $s \le 2$, $3 \le s \le 4$, and $s \ge 5$. The score is deterministic metadata, not a human visual-difficulty rating. In particular, RefCOCOg has fewer $D3+$ questions than the other sources but longer expressions on average, so D and composite difficulty must not be conflated.

A.3 What the Balancing Does and Does Not Guarantee

Equal source counts do not make every type of question equally common. The cohort balances the three datasets and seeks coverage of six semantic groups, subject to available examples. Outdoor/accessory and vehicle groups are therefore smaller than the other four groups. Every selected

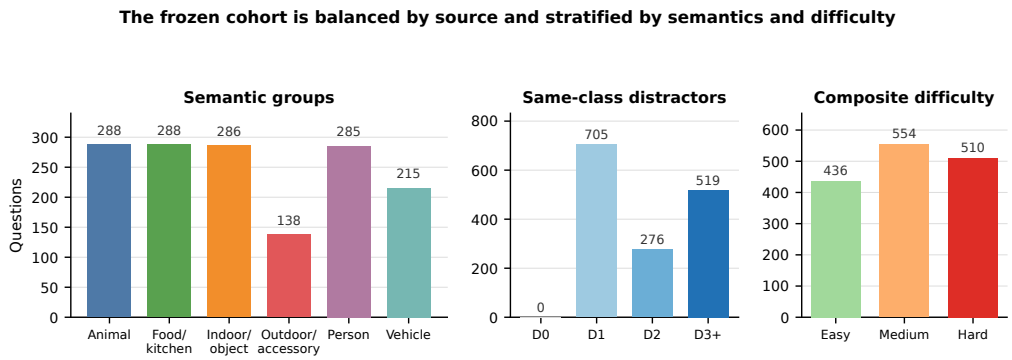

Figure S1: Frozen cohort composition. The three panels visualise the semantic-group totals, same-class distractor counts, and composite difficulty totals tabulated above. The three source datasets each contribute exactly 500 questions.

question has at least one same-class distractor, so $D0 = 0$. The category, split, and difficulty tables describe this evaluation population; they do not imply uniform coverage of all RefCOCO conditions.

A.4 Multipart Targets in Plain Language

COCO can encode one target instance with several disconnected polygon parts, for example when occlusion splits a visible region. The canonical benchmark chooses the well-formed source part with the longest coordinate list and deterministically normalises it to N vertices. "Original selected part" later means this same part before vertex deletion or insertion. "All-parts union" instead fills every well-formed disconnected part from the same annotation. It does *not* combine the $N = 8, 16, 24, 32, 64$ polygons or join different object instances. Reporting all three regions reveals whether a conclusion depends on fixed-N approximation or on omitted disconnected pieces. The cohort contains $148/1{,}500$ multipart UIDs (9.87%). Under the unchanged Pillow masks, the longest-list part covers 71.76% of its all-parts union on average (median 76.70%) and is also the largest raster-area part in 85.14% of multipart cases. These construction statistics are computed before and independently of any model output. The target selection rule is deliberately mechanical. Longest coordinate list is reproducible but is not claimed to be the perceptually dominant component in every image. This is why the canonical score remains the benchmark definition while selected-part and all-parts scores are sensitivity analyses. No alternative target is chosen after seeing a model prediction. The large single- versus multipart gap is therefore not attributed to one mechanism. Multipart annotations simultaneously correlate with occlusion, fragmented visible support, contour difficulty, and mismatch between a single selected component and the whole perceived object. The all-parts audit measures sensitivity to the reference choice. It does not separate these factors causally. Multipart status is therefore a descriptive slice, not evidence that component selection alone causes the loss.

A.5 Output, Parsing, and Score

For a multipart COCO annotation, construction selects the flattened polygon part with the most listed coordinates. If it has fewer than N vertices, the midpoint of a longest edge is inserted repeatedly; if it has more than N, iterative Visvalingam–Whyatt removal deletes the smallest local-area vertex. The source ordering fixes the phase, and deterministic tie handling makes all five G_N files reproducible. The all-parts reference used later is the filled union of every valid polygon part from the same annotation, not a union over values of N. The largest budget tests output reliability with little added reference detail. The $8/16/24$ budgets cover compact contour descriptions. At 32, mean reference fidelity is already about 99%; 64 doubles sequence length for a small further gain. All budgets were fixed before model evaluation. They do not define a uniformly increasing difficulty scale: more points can improve a representation while making its generation less reliable. Parsing tests whether an answer supplies the requested coordinate list, not whether that list forms a valid polygon. The model receives the image, expression, range $[0, 999]$, and vertex count N. It is asked for $[x_1, y_1, \ldots, x_N, y_N]$ as integers. The parser accepts the last explicit answer list or a bracketed

list with exactly $2N$ finite numeric values. Decimal notation alone is accepted at this stage. Answers without such a list receive zero in fixed-denominator scores.

The scorer maps normalised coordinates to $(xW/999, yH/999)$, then clips them to $[0, W-1] \times [0, H-1]$ for rasterisation. Thus a coordinate of 999 reaches the outer image boundary and is clipped to the last pixel. This differs from scaling directly by $W-1$ and $H-1$. We use the same mapping for predictions, canonical targets, boxes, quantised references, and specialist outputs. Prediction and target are filled with the same Pillow binary-mask convention before computing IoU.

A.6 Metric Family and Denominators

For an evaluation set S, fixed-denominator IoU is

$$\mathrm{mIoU}_{\mathrm{fixed}} = \frac{100}{|S|} \sum_{i \in S} \mathrm{IoU}_i, \tag{S5}$$

An unparseable response has $\mathrm{IoU}_i = 0$; we do not evaluate an absent polygon. Parseable-only IoU uses the same numerator but divides by the number M of parseable outputs. For $M > 0$, writing parse coverage as $p = M/|S|$ gives $\mathrm{IoU}_{\mathrm{fix}} = p\,\mathrm{IoU}_{\mathrm{parse}}$. A high conditional score can therefore coexist with a low fixed score when few answers parse. This identity accounts for output coverage; it does not separate independent visual and output abilities. We also report $\mathrm{Acc}@\tau = 100|S|^{-1} \sum_i \mathbf{1}[\mathrm{IoU}_i \geq \tau]$ at $\tau \in \{.1, .3, .5, .7, .9\}$. Acc@.1 distinguishes a coarse overlap from a complete miss; Acc@.5 is a medium-quality operating point; Acc@.9 demands near-exact masks. A configuration can consequently rank differently at .1 and .7 even when its mean IoU is lower. This threshold family prevents a single mean from hiding whether performance consists of many approximate localisations or a smaller number of precise polygons. The resampling unit is the question. Per-budget comparisons resample question identities and keep each pair of condition scores together. Five-budget analyses instead resample 1,500 UID clusters, retaining all five budget rows for each selected UID. This preserves dependence across budgets. The resulting uncertainty concerns variation across evaluated questions, not repeated generations for a fixed prompt.

Unless otherwise stated, main analyses use 2,000 replicates, Qwen thinking slice interactions use 1,000, and transport-clean cue comparisons use 10,000 paired replicates; all use base seed 20260718. Within the primary controlled grid, conditions use the full 1,500 UIDs per budget except the three image ablations: no image, half resolution, and greyscale each use a pre-specified matched 500-UID cohort and the archived Qwen max-thinking outputs on those exact UIDs as their paired reference.

The frozen protocol initially designated filled-sequence fixed IoU as its system-utility endpoint. The legality-enforced parsed-polygon score applies a stricter zero-on-invalid rule to the same predictions. We report both fixed-denominator axes: filled-region utility and legal-polygon completion. The latter is not retroactively presented as a second preregistered endpoint; claims about polygon-contract completion use legality-enforced scoring, whereas filled scoring asks whether an emitted sequence still fills a useful region.

Multiplicity-controlled inference is limited to two separately adjusted five-contrast families. Family A is filled-sequence direct polygon minus box-derived polygon across the five N, and Family B is the corresponding legality-enforced comparison. The frozen 13-condition grid is a run inventory, not an external preregistration. The controlled studies are structured diagnostics rather than one 39-test confirmatory family: Family C groups no-image, half-resolution, and greyscale across three budgets, while Family D groups wrong spatial, wrong colour, and combined wrong spatial–colour cues across three budgets. Their paired analyses assess direction within each family without a familywise confirmatory claim. Truthful cues, parameter changes, the single A21 repeat, slices, and alternate metrics are exploratory or sensitivity analyses.

Images preserve aspect ratio, have longest side at most 1,024 pixels, and are encoded as JPEG quality 85. The frozen prompt is:

> **System:** You are a 2D referring polygon localizer. Return exactly one line and nothing else: `ANSWER: [x1,y1,...,xN,yN]`. Exactly $2N$ numbers, each in 0..999 (normalized image coordinates). No reasoning.
> **User:** Object: {`description`}. Image size: {`width`}x{`height`}. For each point $k = 1..N$: x_k maps across image width, y_k maps across height; use 0..999 for full span. Output format (required): `ANSWER: [x1,y1,...,xN,yN]`. Constraints: $2N$ numbers only, each in 0..999; no explanation, markdown, or extra words.

"No reasoning" constrains the visible answer format; it does not disable provider-side thinking when a configuration explicitly enables it.

B Complete Main-Experiment Results

B.1 Evaluated Configurations

Table S5: Main evaluated configurations. Gemini additionally uses high media resolution. Doubao is not labelled thinking-on because no controlled thinking field was sent.

Paper label	Exact model identifier	Reasoning request	Temperature	Maximum output
Doubao default	`doubao-seed-2-0-pro-260215`	field omitted; provider default	0	384/512/640/768/1280 by N
Kimi fixed-think	`kimi-k2.7-code`	fixed-thinking endpoint	1	8192
Qwen thinking-off	`qwen3.7-plus`	disabled	0	8192
Qwen max-think	`qwen3.7-plus`	enabled, maximum/default budget	0	8192
Gemini high-think	`gemini-3.5-flash`	`thinking_level=high`	provider default	32768

Results describe dated hosted configurations, rather than whole model families. Main generations were collected on July 10–17, 2026, using the identifiers in Table S5. Providers may update the weights served under an identifier. Archived outputs therefore support offline rescoring, but do not guarantee identical future generations. The study compares four providers and one Qwen setting contrast; it is not a probability sample of vision–language models.

B.2 Same-Cohort Box and Specialist Calibration

The box calibration reuses all 1,500 images, expressions, and UIDs. The user prompt changes only the output contract, requesting one tight normalised rectangle as `ANSWER: [x1,y1,x2,y2]`; the system, image preprocessing, and per-configuration provider settings remain unchanged. It uses the same uniform finalisation rule as the polygon headline: the primary-run endpoint followed by A01/A02 only for still-unparseable answers, retaining the first usable answer and never replacing a parseable answer. A01/A02 are parse-failure-triggered stateless retries, not edits of the earlier response.

Parse below is coverage after A02; failures remain zero in box mIoU and Acc@.5. Both metrics use the frozen COCO instance bounding box. For the Box→poly column, the four predicted rectangle edges are sampled at equal perimeter intervals to obtain exactly N cyclic points and passed through the canonical polygon rasteriser. The reported value pools all five budgets against G_N, measuring what a short successful localisation retains when its rectangular region is judged as the benchmark's shape output. For box calibration, the recovery messages replace the polygon-vertex clause with the exact four-coordinate contract `ANSWER: [x1,y1,x2,y2]`; the image, system message, all other request fields, stateless retry rule, and first-usable selection rule are unchanged.

Table S6: Matched box calibration on the same distractor-rich 1,500-question cohort. Box→poly scores the predicted rectangle as an exact-N polygon; the final column is direct polygon minus that baseline.

Configuration	Parse	Box IoU	Box Acc@.5	Box→poly	Direct−rect. Δ
Doubao default	99.87	83.57	94.13	53.81	−11.16
Kimi fixed-think	99.60	81.86	92.73	52.68	−18.78
Qwen thinking-off	98.93	76.84	87.39	49.43	−35.02
Qwen max-think	99.73	86.69	96.07	55.92	−9.09
Gemini high-think	99.80	88.17	97.07	57.30	+0.45

Table S7: Direct polygon versus predicted-box→polygon by configuration and vertex budget. Each N group reports Box→poly (Box) and direct minus Box→poly (Δ); direct scores appear in Table S9. Averaging the five point estimates reproduces the pooled values.

	$N = 8$		$N = 16$		$N = 24$		$N = 32$		$N = 64$	
Configuration	Box	Δ	Box	Δ	Box	Δ	Box	Δ	Box	Δ
Doubao	54.88	−13.72	54.21	−7.37	53.72	−10.36	53.39	−9.04	52.85	−15.29
Kimi	53.96	−15.75	53.14	−15.84	52.61	−16.75	52.15	−18.41	51.54	−27.15
Qwen-off	50.82	−30.31	49.95	−35.21	49.41	−32.44	48.92	−30.11	48.05	−47.01
Qwen-max	57.16	−8.96	56.35	−6.65	55.80	−6.68	55.40	−6.76	54.89	−16.39
Gemini	58.32	+4.32	57.56	+7.18	57.08	+8.78	56.91	+0.92	56.63	−18.97

A strong box score does not guarantee that direct polygon output improves the represented region. On this $D \geq 1$ cohort, box parse coverage is 98.93–99.87 and box Acc@.5 is 87.39–97.07 under the matched finalisation policy. Four configurations remain below their own rectangle-derived baselines at every budget, with the largest deficits at $N = 64$. Gemini gains from direct polygons at $N = 8$–24, has an unresolved contrast at $N = 32$, and reverses at $N = 64$ (Table S7). Its pooled $+0.45$ difference therefore hides a strong dependence on budget.

To test the Gemini budget interaction without selecting only the visually largest pair, we use a UID-clustered Wald procedure. For UID i and budget N, let d_{iN} be direct minus box-derived IoU, and let C contain four independent contrasts of the five-element mean vector. We estimate the covariance of $\bar{\mathbf{d}}$ from 2,000 whole-UID bootstrap resamples that retain all budgets jointly, then compute $W = (C\bar{\mathbf{d}})^{\mathsf{T}}(C\widehat{\Sigma}C^{\mathsf{T}})^{-1}(C\bar{\mathbf{d}})$. The result is $W = 684.3$ on four degrees of freedom ($p < .001$), rejecting a constant mean contrast across budgets.

As a separate sensitivity analysis, $T = \sum_N (\bar{d}_N - \bar{d}_\cdot)^2$ is evaluated under the stronger within-UID budget-label exchangeability null using 10,000 independent within-UID permutations. Its Monte Carlo p-value is also $< .001$; we do not interpret this test as testing equal means alone.

Within Gemini's five filled-sequence comparisons, Holm adjustment retains the positive decisions at $N = 8, 16, 24$ and the negative decision at $N = 64$, while $N = 32$ remains unresolved. The corresponding Gemini legality-enforced family retains positive decisions at $N = 8, 16, 24$ and negative decisions at $N = 32, 64$. These decisions use the frozen phase-zero target: the eight-phase audit in Table S31 lacks matched box-derived rows and therefore cannot validate them across phase. Figure S2 visualises both contrast families.

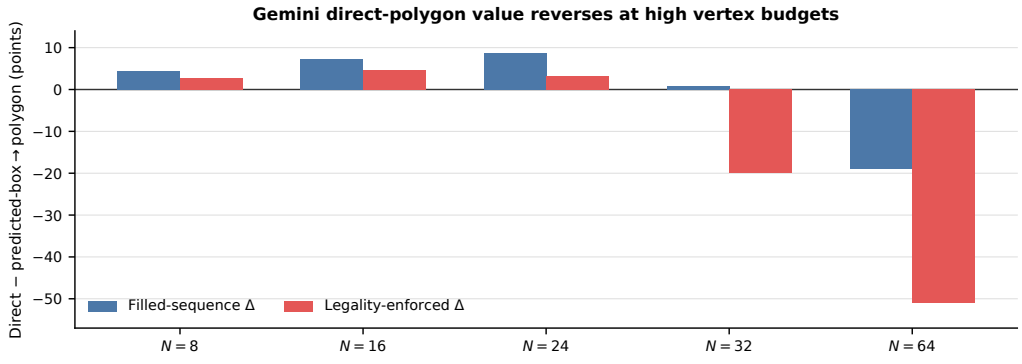

Figure S2: Matched direct-polygon value relative to predicted-box→polygon for Gemini. Bars visualise the filled-sequence and legality-enforced paired point differences in Tables S7 and S38. For Gemini, both tracks are positive at moderate budgets and reverse as output length and legality demands increase.

Gemini's near-zero pooled difference reflects cancellation across budgets, not equivalence at each budget. Predicted rectangles score only 49.43–57.30 IoU against contour targets across configurations. The common-reference comparison therefore asks whether the model converts a more expres-

sive output into a better region. It does not identify an internal geometric mechanism or establish equivalent boundaries. Figure S5 shows the pooled comparisons under the canonical starting phase.

We separately calibrate the representation and a task-specific model. "GT box$\rightarrow N$" converts the annotation's tight rectangle into the same exact-N sequence, measuring the tight annotation rectangle as a shape baseline. A perfect target rectangle reaches only 59.16 against the fixed-N shape. The quantised oracle maps the canonical G_N into the announced 0–999 alphabet and back before rasterisation. The specialist uses the public PolyFormer-L model-zoo checkpoints corresponding to RefCOCO, RefCOCO+, and RefCOCOg from the authors' Apache-2.0 release. PolyFormer-L has a Swin-L visual encoder, BERT-base text encoder, and 12-layer transformer encoder and decoder. We use the checkpoint matching each question's source benchmark and its published evaluation split; exact downloaded-file hashes are retained in the artifact manifest.

The release pretrains referring-expression comprehension jointly on Visual Genome, Flickr30k Entities, RefCOCO, RefCOCO+, and RefCOCOg, then fine-tunes REC and referring segmentation on the three RefCOCO-family datasets. A test image from one source may therefore have appeared through another source during mixed pretraining. This result is an *illustrative specialist calibration*. Training overlap and deployment differences prevent a controlled comparison with the hosted models. Across the five equally weighted vertex budgets, its filled-sequence mean is 61.34.

For conversion, we decode PolyFormer's native multipart polygons, rasterise their filled union on the original canvas, retain the largest-area foreground component, ignore interior holes, and trace its external boundary. We fix phase at the topmost then leftmost boundary pixel, use clockwise traversal, resample by cumulative arc length to exactly N vertices, map them once to integer 0–999 coordinates, and pass that sequence to the unchanged canonical scorer. No target mask, instance identity, test-time refinement, component search against ground truth, or post-hoc choice among candidate outputs is used.

Table S8: Representation and specialist calibration. All entries use the unchanged rasteriser and fixed denominator.

N	GT box$\rightarrow N$	Quantised G_N oracle	PolyFormer mask$\rightarrow N$
8	60.87	99.08	61.54
16	59.43	99.39	61.12
24	58.79	99.53	62.61
32	58.51	99.64	62.08
64	58.19	99.72	59.37
Mean	59.16	99.47	61.34

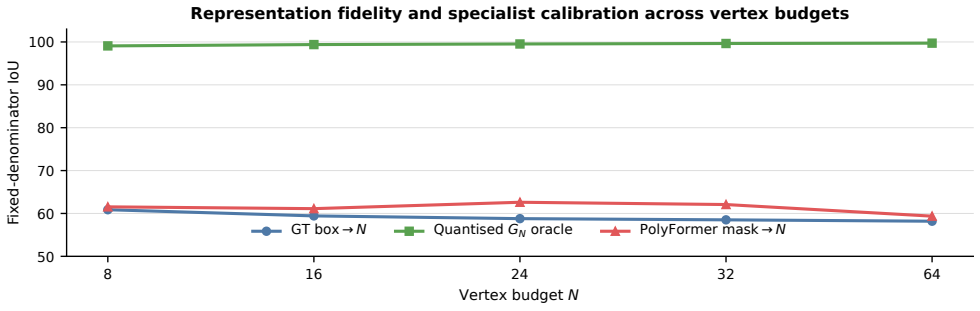

Figure S3: Representation and specialist calibration across vertex budgets. Quantisation preserves the canonical target almost perfectly, while the best possible rectangle and the illustrative specialist occupy a much lower and similar filled-region range.

Coordinate quantisation preserves nearly all reference region IoU (Table S8 and Figure S3). This is an area-fidelity result, not a guarantee of polygon legality; Section G.2 checks the rounded vertex lists separately. PolyFormer's mean of 61.34 provides a separate task-specific calibration, but its training overlap prevents a leakage-free comparison. We omit its legality-enforced score because no row-level export matched to the validator was frozen.

B.3 Per-Model, Per-Resolution Scores

Output completion explains only part of the performance gap. At low budgets, fixed and parseable-only scores are close because most answers can be parsed. They separate when long coordinate lists fail. Qwen-off rises from 14.41 fixed IoU to 21.14 on parseable outputs, but remains well below the other configurations. Kimi shows a different limitation: it usually completes the format, yet its conditional region quality trails Doubao and Qwen-max. Gemini has the strongest conditional overlap but loses substantial coverage at $N = 64$.

Every configuration faces the same questions at every vertex budget. This matching makes differences in Table S9 interpretable without changes in cohort composition. Cross-model and cross-budget analyses retain exact question intersections. The remaining views answer different questions: Table S12 reports threshold accuracy, Table S39 separates primary-run and additional-recovery outcomes, and Table S36 tests reference-mask sensitivity.

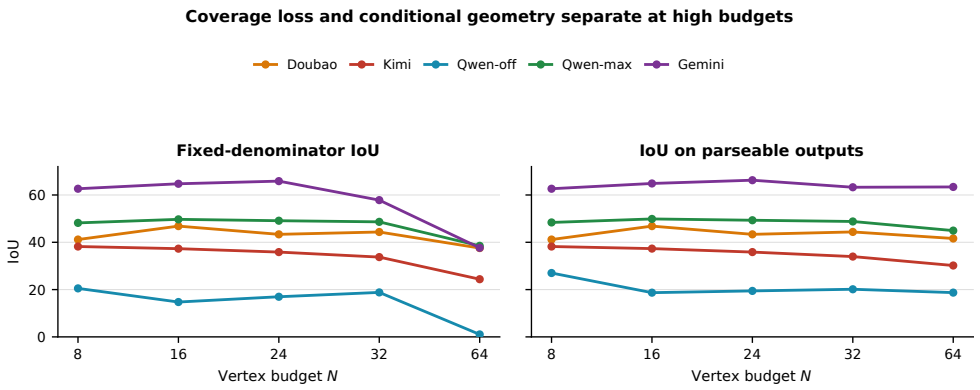

Figure S4: Main score curves under fixed and parseable-only denominators. The left panel combines geometry with output coverage; the right conditions on usable exact-length lists. Their separation at high N shows that delivery loss compounds, rather than fully explains, the remaining geometric error.

B.4 Micro Summary and Paired Analysis

Pooling all five resolutions gives the operational summary in Table S10. Headline five-budget inference resamples 1,500 UIDs and carries the five associated N rows together; per-budget matched inference resamples question identities. The question, rather than 7,500 correlated question–budget instances, is therefore the sampling unit. The compact main-table *Mean* row averages the four displayed one-decimal configuration cells in each column, whereas the micro table below retains the frozen full-precision configuration values. This display convention can change only the final reported tenth (for example, 47.9 versus 47.8 for the parseable-only mean); it does not change any configuration score or comparison.

B.5 Fair Same-Question Contrasts

Different models can lose score at different stages. In the best-effort comparison, Gemini has lower parse coverage than Qwen-max but higher overlap on questions parsed by both. Its fixed-IoU advantage is $58.93 - 46.83 = 12.10$, consistent with the paired mean. In contrast, Qwen-max improves on Qwen-off in both coverage and jointly parseable geometry. Its advantage therefore cannot be explained solely by assigning zero to failed outputs.

Paired comparisons use exact same-question rows at each N. Table S11 reports pre-specified same-question model contrasts rather than subtracting unmatched cohort means; these diagnostics read the best-effort archive (Section A), and only the Gemini row differs between views.

Table S9: Complete per-resolution results. Unparsable outputs count as zero in Fixed IoU and fixed-denominator accuracy; Parseable IoU conditions on exact-length numeric lists.

Configuration	N	Parse	Fixed IoU	Parseable IoU	Acc@.5
Doubao default	8	100.00	41.16	41.16	43.33
	16	100.00	46.84	46.84	51.47
	24	100.00	43.36	43.36	45.60
	32	99.93	44.35	44.38	46.80
	64	90.27	37.56	41.61	39.53
Kimi fixed-think	8	100.00	38.21	38.21	33.80
	16	99.93	37.30	37.33	32.87
	24	100.00	35.86	35.86	29.67
	32	99.33	33.74	33.96	26.80
	64	80.80	24.39	30.18	18.93
Qwen thinking-off	8	75.93	20.51	27.01	19.07
	16	78.80	14.74	18.70	11.00
	24	87.27	16.97	19.44	9.73
	32	93.40	18.81	20.14	9.80
	64	5.53	1.04	18.72	0.53
Qwen max-think	8	99.60	48.20	48.39	56.00
	16	99.67	49.70	49.87	58.07
	24	99.60	49.12	49.32	56.60
	32	99.67	48.64	48.80	56.73
	64	85.67	38.50	44.94	42.53
Gemini high-think	8	100.00	62.64	62.64	76.00
	16	99.80	64.74	64.87	76.00
	24	99.40	65.86	66.25	79.60
	32	91.40	57.83	63.27	69.40
	64	59.40	37.66	63.40	44.80

Table S10: Five-resolution micro results on the shared 1,500-question set. Matched contrasts are reported separately rather than repeated for every marginal mean.

Configuration	Parse	Fixed IoU	Parseable IoU	Acc@.5
Doubao	98.04	42.65	43.51	45.35
Kimi	96.01	33.90	35.31	28.41
Qwen-off	68.19	14.41	21.14	10.03
Qwen-max	96.84	46.83	48.36	53.99
Gemini	90.00	57.75	64.16	69.16

B.6 Full Accuracy Curves

Table S12 lists the complete fixed-denominator accuracy ladder for every model–N cell, complementing the fixed/parseable-only results in Tables S9 and S10; Figure S6 visualises the same threshold–budget surface.

B.7 Rank Stability Is Metric-Dependent

The best model depends on the required overlap threshold. In the best-effort archive, fixed IoU and Acc@.5 rank Gemini $>$ Qwen-max $>$ Doubao $>$ Kimi $>$ Qwen-off at every budget. Other thresholds change this ordering. Kimi often exceeds Doubao at Acc@.1 but falls below it in mean IoU and Acc@.5: coarse overlap is more frequent, while precise regions are less frequent. At $N = 64$, Doubao can lead at low thresholds even though Gemini leads at .5, .7, and .9. All comparisons use the same questions and the scoring view in Section A; only Gemini differs from the uniform view.

The exclusion in Table S13 keeps its best-effort comparison on one common four-configuration view: it excludes Gemini and includes Qwen thinking-off. As a separate uniform-IoU^2 check,

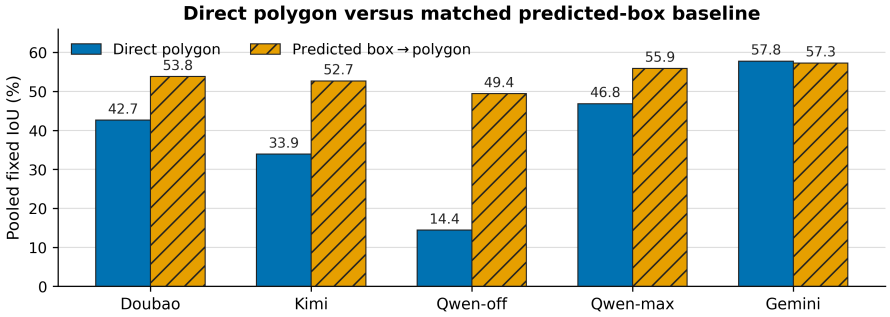

Figure S5: Pooled direct polygon versus matched predicted-box→polygon. Bars reproduce the fixed-denominator point estimates in Tables S6 and S10. Only Gemini is approximately level with its rectangle-derived baseline after pooling all five budgets.

Table S11: Question-paired fixed-IoU contrasts in the best-effort diagnostic archive. The displayed mean contrast equals the difference of the corresponding archive means.

Contrast $A - B$	Fixed difference	A/tie/B
Qwen-max − Doubao	+4.18	50.9/3.9/45.2
Qwen-max − Kimi	+12.93	67.4/3.2/29.3
Qwen-max − Qwen-off	+32.42	80.4/6.7/12.9
Gemini − Qwen-max	+12.10	69.8/3.6/26.7

adding Gemini to the Acc@.1 rankings yields minimum cross-budget Kendall $\tau = -0.20$, mean $\tau = 0.44$, and four distinct orders; the extra model therefore increases, rather than hides, low-threshold rank instability. In 2,000 bootstrap resamples, Qwen-max is first in every four-model micro replicate. The $N = 64$ cells are less certain: Qwen-max ranks first in 82.7 of four-model resamples while Doubao occasionally overtakes it. This is compatible with the broad ordering while warning against treating one small high-N point estimate as deterministic.

B.8 Parse-Failure-Triggered Stateless Retries

Retries mainly recover missing outputs rather than improve already usable polygons. Fixed IoU rises as formerly unparseable questions acquire an answer, while parseable-only IoU changes little (Table S14). No parseable answer is replaced by a better candidate. Gemini's extra archive retry has the same pattern. Qwen-off gains coverage but still trails the other configurations. We retain both the archived primary-run score, IoU^0, and the uniform capped-retry score, IoU^2, because they describe different operating conditions.

The archive preserves the primary-run endpoint and two external parse-failure-triggered recovery stages, A01 and A02; these produce the IoU^0, IoU^1, and IoU^2 views used in the main text. Gemini has one further authorised retry, A03, which defines its best-effort archive view (Section A); the table below stops at the uniform-protocol endpoint IoU^2. The primary-run endpoint, historically labelled *Formal*, is immutable.

Each external recovery stage queries only questions for which no earlier call produced a parseable exact-length polygon and retains the first subsequently authorised usable answer. It can therefore add coverage but cannot replace or rerank an existing parseable answer. Because the image is resent and the prompt requests renewed inspection of the referent, each retry is a conditional generation opportunity rather than format-only repair. The compact trajectory table reports the four configurations displayed in the main result table; Qwen thinking-off primary-run and final endpoints are reported separately in Table S39.

Exact recovery request. Each A01/A02/A03 stage begins with a fresh, stateless request with the original image re-sent, the unchanged system prompt printed in the benchmark protocol, and the same model identifier, image preprocessing, temperature, thinking field, output limit, and provider

Table S12: Fixed-denominator accuracy at five IoU thresholds. Low-threshold ordering can differ from IoU/Acc@.5 ordering, especially for Kimi.

Configuration	N	@.1	@.3	@.5	@.7	@.9	Configuration	N	@.1	@.3	@.5	@.7	@.9
Doubao	8	75.00	59.33	43.33	24.20	1.73	Qwen-max	8	81.33	71.13	56.00	30.27	1.93
Doubao	16	78.73	65.13	51.47	33.40	3.47	Qwen-max	16	81.40	71.53	58.07	33.27	2.20
Doubao	24	76.87	60.00	45.60	29.13	2.80	Qwen-max	24	82.13	71.27	56.60	32.40	1.73
Doubao	32	77.13	61.20	46.80	30.60	3.00	Qwen-max	32	81.40	71.27	56.73	30.87	1.93
Doubao	64	69.27	54.60	39.53	21.27	1.73	Qwen-max	64	68.80	56.27	42.53	22.73	1.07
Kimi	8	83.80	63.07	33.80	10.07	0.13	Gemini	8	93.00	87.80	76.00	48.80	8.00
Kimi	16	82.27	60.00	32.87	10.07	0.13	Gemini	16	90.60	84.60	76.00	59.80	11.00
Kimi	24	81.80	57.93	29.67	7.20	0.27	Gemini	24	92.00	87.20	79.60	58.20	11.00
Kimi	32	78.53	55.33	26.80	6.53	0.20	Gemini	32	80.80	76.40	69.40	49.80	10.00
Kimi	64	58.27	38.47	18.93	5.07	0.00	Gemini	64	52.80	49.40	44.80	33.60	7.60
Qwen-off	8	44.13	31.80	19.07	6.27	0.20							
Qwen-off	16	36.00	22.07	11.00	2.80	0.07							
Qwen-off	24	46.07	25.13	9.73	1.47	0.07							
Qwen-off	32	52.00	27.60	9.80	1.93	0.00							
Qwen-off	64	2.87	1.67	0.53	0.00	0.00							

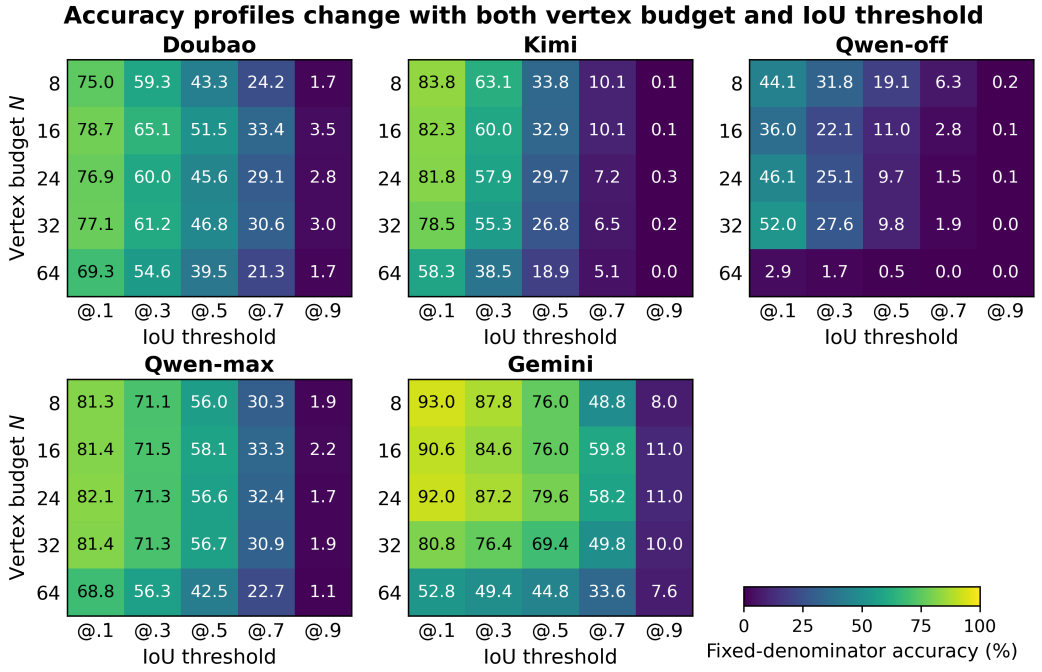

Figure S6: Full fixed-denominator accuracy surfaces. Each panel visualises one configuration's values from Table S12 over all five vertex budgets and five IoU thresholds. The surfaces make visible why a single threshold and a single budget can induce a different ranking from the full curve.

parameters as the primary-run configuration. The malformed response is not included in the retry context. The original user prompt is replaced by the following complete A01 user message:

> Object: {`description`}. Image size: {`width`}x{`height`}. The previous attempt could not be parsed because it did not contain exactly {`2N`} finite coordinates. Inspect the image again and return exactly {`N`} polygon vertices in normalized 0..999 image coordinates. Output exactly one line: `ANSWER: [x1,y1,...,xN,yN]`. Use exactly {`2N`} numbers; no explanation, markdown, labels, or extra text.

Table S13: Across-N ranking stability among the four configurations other than Gemini. Kendall τ summarises pairwise agreement between two resolution-specific rankings.

Cohort/metric	Minimum Kendall τ	Mean τ	Distinct orders
Four-model fixed IoU	1.000	1.000	1
Four-model Acc@.1	0.000	0.533	3
Four-model Acc@.3	0.667	0.867	2
Four-model Acc@.5	1.000	1.000	1
Four-model Acc@.7	0.667	0.867	2
Four-model Acc@.9	0.333	0.733	2

For A02, the complete user message is:

> FORMAT RECOVERY. Object: {`description`}. Image size: {`width`}x{`height`}. Re-evaluate the object and output exactly {`N`} cyclic polygon vertices in normalized 0..999 coordinates. Count before sending: the bracket must contain exactly {`2N`} finite numbers. Return one line only: `ANSWER: [x1,y1,...,xN,yN]`. Do not output reasoning, markdown, or any other text.

A03, authorised only for the frozen Gemini archive view, repeats the A02 system, user message, image, and request fields exactly; it is an additional opportunity rather than a different retry instruction. No previous-stage response or conversation history is supplied at stage entry; selection never replaces an already parseable answer.

Table S14: Stage-by-stage recovery trajectory for the four main-table configurations, pooled over all five N values. IoU^0 is the immutable primary-run endpoint, possibly including internal format repair, IoU^1 includes A01, and IoU^2 includes A01/A02. Fixed IoU retains the complete denominator; parseable-only IoU conditions on exact-length numeric lists. Finalisation never searches among multiple parseable answers for a higher score.

Configuration	IoU^0	IoU^1	IoU^2	Parseable IoU: $0 \rightarrow 2$
Doubao	40.00	42.13	42.65	43.66→43.51
Kimi	32.48	33.79	33.90	35.19→35.31
Qwen-max	45.81	46.49	46.83	48.27→48.36
Gemini	49.60	55.46	57.75	64.79→64.16

If p_0, p_1 are parse coverage after the primary-run endpoint and A01, the number of scheduled primary/recovery stages per question is $1 + (1 - p_0) + (1 - p_1)$. The recorded values are 1.116/1.108/1.853/1.089/1.370 for Doubao/Kimi/Qwen thinking-off/Qwen max-thinking/Gemini. They count outer stages, not physical API calls: internal format recovery and transport retries are not captured by this expression. Table S39 keeps the primary-run and additional-recovery endpoints separate.

C Dataset, Type, and Difficulty Slices

All tables in this section are five-N micro averages on the shared 1,500-question set and support within-model slice trends. The dataset, semantic-group, and cross-slice tables read the best-effort archive (Section A); the difficulty tables (composite and D/Q/RS) follow the uniform protocol and match the main-text difficulty slices exactly.

C.1 Source Dataset

Table S15 reports fixed IoU separately for the three source corpora; the balanced 500-per-source design makes the rows directly comparable.

C.2 Semantic Category Group

Table S16 slices fixed IoU by the six frozen semantic groups; the configuration-specific maxima and minima are discussed together with the other heterogeneity results below.

Table S15: Fixed IoU by source dataset. All five configurations are highest on RefCOCO, but the source gap is smaller than most geometric difficulty gaps.

Source	Doubao	Kimi	Qoff	Qmax	Gemini
RefCOCO+	42.40	33.57	12.95	46.31	58.00
RefCOCO	43.41	35.82	15.77	47.39	60.51
RefCOCOg	42.15	32.30	14.52	46.80	58.27

Table S16: Fixed IoU by semantic group. The ordering is model-dependent, so category differences are descriptive rather than a universal difficulty ranking.

Category group	Doubao	Kimi	Qwen-off	Qwen-max	Gemini	Mean
Animal	42.80	32.28	17.28	46.71	63.44	40.50
Food/kitchen	41.93	34.15	13.86	48.10	59.25	39.46
Indoor/object	39.92	34.96	11.86	45.49	53.59	37.16
Outdoor/accessory	37.56	30.61	12.80	42.68	59.41	36.61
Person	46.74	33.87	15.64	46.44	57.38	40.01
Vehicle	44.92	36.46	14.12	50.26	61.29	41.41

C.3 Composite Difficulty and Component Axes

Table S17 stratifies the same scores by the composite difficulty label and its three geometric components; every configuration degrades monotonically along each pooled axis, and the full parse/fixed/parseable accounting follows in the next subsections.

C.4 Difficulty Components

Harder geometric conditions reduce region quality even among parseable outputs. Every configuration declines across the pooled easy-to-hard, distractor-count, and contour-complexity axes; larger targets score better (Table S17). Similar trends in parseable-only scores show that formatting failures are not the sole explanation. Some source-specific bins reverse, and the axes are correlated. These are descriptive patterns, not universal causal laws. Qwen-off's smaller absolute slopes reflect its low score floor, rather than evidence of greater robustness.

C.5 Semantic Heterogeneity Is Configuration-Specific

No semantic category is hardest for every configuration. Doubao peaks on person, Kimi and Qwen-max on vehicle, and Qwen-off and Gemini on animal. Outdoor/accessory is weakest for Doubao, Kimi, and Qwen-max; indoor/object is weakest for Qwen-off and Gemini. Categories also differ in target size, multipart frequency, and language. These slices help locate failures, but cannot attribute them to category identity alone. The source effect is comparatively modest. All five rows peak on RefCOCO, yet their within-configuration source ranges are only about 1.1–3.5, smaller than most composite, Q, or RS gaps. The agreement could reflect split and expression differences rather than an intrinsic dataset effect. Reporting the source table beside the balanced counts makes this composition visible instead of silently pooling three corpora.

C.6 Multipart Structure and Additional Descriptive Slices

Multipart targets are harder for every configuration, but several factors can contribute. The penalties are 16.90, 15.11, 4.24, 19.21, and 24.57 points in the displayed configuration order (Figure S7). Such objects may have disconnected visible regions and stronger occlusion, while the canonical target retains one part. All-parts rescoring raises means modestly without removing the gap or changing the ordering. Qwen-off's smaller penalty again occurs near a low score floor. Additional correlated exploratory slices are omitted because they do not support confirmatory claims or alter the reported conclusions. The stable-boundary paired analysis in Table S32 is the stronger test of output-length behaviour.

Table S17: Fixed IoU by composite and component difficulty (uniform protocol). Every configuration follows easy $>$ medium $>$ hard, $D1 > D2 > D3+$, $Q1 > Q2 > Q3$, and $RS3 > RS2 > RS1$.

Axis	Bin	Doubao	Kimi	Qwen-off	Qwen-max	Gemini
Composite	Easy	50.66	42.52	17.48	55.57	65.14
	Medium	43.00	33.66	13.74	46.41	58.08
	Hard	35.44	26.78	12.53	39.82	50.99
Distractors	$D1$	45.87	36.30	15.90	50.34	61.79
	$D2$	40.50	33.13	13.84	45.75	57.64
	$D3+$	39.43	31.05	12.69	42.64	52.32
Complexity	$Q1$	46.52	39.60	15.32	52.24	63.44
	$Q2$	44.45	33.80	14.02	46.94	57.82
	$Q3$	35.91	26.53	13.59	39.64	50.22
Relative size	$RS1$ (small)	36.96	29.28	12.64	42.89	54.17
	$RS2$	42.92	34.22	14.60	47.34	58.80
	$RS3$ (large)	49.80	39.60	16.55	51.48	61.42

Table S18: Single- versus multipart source annotations. The $1{,}352/148$ UID split yields $6{,}760/740$ question–budget rows after pooling five N values.

Configuration	Single fixed	Multipart fixed
Doubao	44.32	27.42
Kimi	35.39	20.28
Qwen-off	14.83	10.59
Qwen-max	48.73	29.52
Gemini	61.36	36.79

D Controlled Model and Input Analyses

The primary controlled grid uses Qwen 3.7 Plus with max-thinking as its reference. One condition lowers the thinking budget to 4096 tokens. Selected visual and cue interventions are also replicated on Gemini. The on/off table uses all 1,500 questions at each N. The final Qwen grid contains 13 conditions at $N = 8, 16, 24$: six cue arms, no image, half resolution, greyscale, a 4096-token thinking budget, two temperatures, and one independent stability repeat.

Cue conditions use the full 1,500-question cohort and the same-UID unmodified prompt as their primary control. The no-image, half-resolution, and greyscale conditions each use a predefined 500-question cohort per N and are paired with archived Qwen max-thinking outputs on those exact UIDs. The thinking-budget, temperature, and stability-repeat conditions use all 1,500 questions per N against the full-1,500 reference.

D.1 Coverage and Interpretation Boundary

All 39 cells have a frozen fixed-denominator point estimate. The image and cue studies use paired analyses to assess the direction of the large effects, and the compact tables emphasise effect sizes. Parameter and repeat contrasts remain descriptive. Fourteen cells retain at least one terminal infrastructure row; those rows remain zero under the declared fixed denominator, while a qualified-only sensitivity removes only failures logged before model completion. The image-evidence studies use predefined 500-question sets independently selected for each N and intervention, so each is paired only with its own exact UID control. Cue conditions use the full 1,500 questions at $N = 8, 16, 24$, and the cue baseline is recomputed on those same questions. These are the control conventions behind the deltas in main Table 2. A10/A19/A21 continue to use the final full-1,500 Qwen-max score: $48.20/49.70/49.12$ at $N = 8/16/24$.

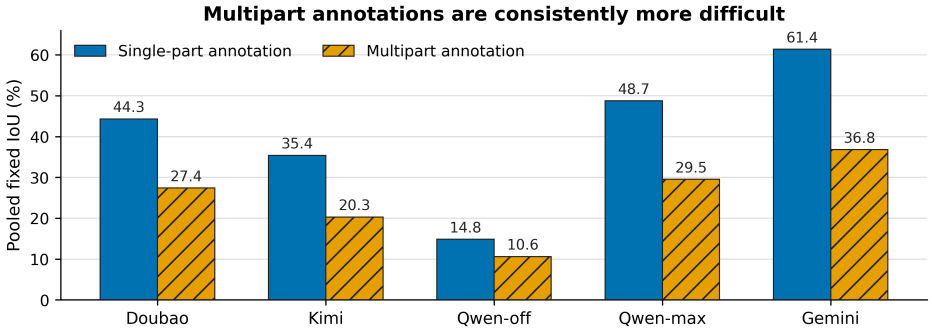

Figure S7: Single- versus multipart fixed IoU. Bars reproduce Table S18. Every configuration scores lower on the multipart subset; the slice remains descriptive because multipart structure covaries with occlusion, target extent, and category.

D.2 Qwen Thinking-Setting Contrast

Enabling thinking improves geometry as well as output completion. Qwen-max exceeds Qwen-off by 28.17 IoU points on questions parsed by both configurations (Table S19). The advantage also appears at every examined budget, source, and $D/Q/RS$ bin. This is a comparison of complete hosted settings; it does not isolate the causal effect of an internal reasoning trace.

Table S19: Same-model finalised-configuration comparison, paired on 1,500 questions per N. The pooled fixed-IoU difference is $+32.42$ points.

N	Max-think	Think-off	Difference	Parse gap
8	48.20	20.51	+27.69	+23.67
16	49.70	14.74	+34.96	+20.87
24	49.12	16.97	+32.15	+12.33
32	48.64	18.81	+29.83	+6.27
64	38.50	1.04	+37.46	+80.13
Micro	46.83	14.41	+32.42	+28.65

The positive configuration difference appears across every examined source and $D/Q/RS$ slice, so the aggregate is not confined to one subgroup. Slice magnitudes are descriptive: no paired test of between-bin interactions was pre-specified, and high-budget behaviour remains entangled with the configuration-specific output protocol.

D.3 Visual Evidence, Resolution, and Colour

Access to the image matters more than preserving all its detail or colour. Removing the image lowers paired IoU by $34.58/35.91/35.47$ points at $N = 8/16/24$ (Table S20). Halving both image dimensions causes smaller losses of $6.05/7.66/6.65$. Paired analyses support both losses at every budget. Greyscale changes are smaller still, $-1.25/-1.56/-1.54$, and the aggregate effect remains unresolved. This ordering holds after matching each intervention to its own control cohort.

The no-image loss is mainly geometric, rather than a collapse of the response format. Residual overlap may come from language, category, or spatial priors; we do not label it "above chance." Half resolution preserves the scene, aspect ratio, prompt, and coordinate frame, but removes detail. Its effect on small targets varies with budget, so the exploratory size splits do not establish a general rule. The unresolved greyscale effect also does not establish that colour is irrelevant. Removing image colour and adding a false colour statement change different information channels.

Table S20: Image ablations on predefined 500-question condition cohorts. Each reference uses the exact same UIDs as its intervention.

Condition	N	Fixed IoU	Matched reference	Paired difference
No image	8	14.05	48.63	−34.58
	16	13.17	49.08	−35.91
	24	13.32	48.79	−35.47
Half width/height	8	42.42	48.47	−6.05
	16	41.65	49.31	−7.66
	24	42.47	49.12	−6.65
Greyscale	8	46.77	48.02	−1.25
	16	48.23	49.79	−1.56
	24	47.49	49.03	−1.54

D.4 Correct and Incorrect Cues

Truthful hints are safer than false hints, but are not necessarily helpful beyond the original expression. Correct spatial cues do not reliably outperform the unmodified prompt (Table S21). False spatial cues reduce fixed IoU by 22.5–25.3 points across the three budgets, compared with 8.6–9.9 for false colour cues. Their combination loses 25.0–27.2 points. These results show greater susceptibility to spatial conflict under the tested templates, not an intrinsic ordering of spatial and colour information.

Table S21: Complete cue-condition grid on the shared full 1,500-question cohort. Matched controls are $47.84/49.26/48.22$ at $N = 8/16/24$. A dagger marks a cell containing a logged infrastructure-terminal row; the qualified-only sensitivity is summarised below. We standardise the final name as `SC`; `CS` in historical run paths denotes the same condition. The compact main table averages its displayed one-decimal budget cells for the Mean column; this table retains the frozen two-decimal cell values.

Cue condition	Meaning	N	Fixed IoU	Difference from matched control
C	correct colour	8	45.26	−2.58
		16	46.50†	−2.75
		24	45.39†	−2.83
S	correct spatial	8	45.52	−2.32
		16	47.21	−2.05
		24	48.49	+0.27
SC	correct spatial+colour	8	44.70	−3.14
		16	46.85†	−2.40
		24	45.81†	−2.40
WC	wrong colour	8	38.49	−9.34
		16	39.36	−9.90
		24	39.57†	−8.65
WS	wrong spatial	8	22.56	−25.27
		16	24.55†	−24.71
		24	25.76†	−22.45
WSC	wrong spatial+colour	8	20.61	−27.22
		16	22.57†	−26.69
		24	23.21†	−25.00

The combined false cues are most harmful, but their losses are sub-additive. One misleading cue may already move the prediction away from the intended target, leaving less room for a second cue to worsen it. This is a possible interpretation, not a demonstrated internal mechanism. Wrong colour often preserves parse coverage despite lowering overlap, which also shows why valid formatting alone is insufficient. Factorial interactions outside the cleanest cell remain descriptive.

D.5 Thinking Budget and Temperature

Small changes to inference settings have weaker and less consistent effects than removing visual evidence or adding false spatial cues. A 4096-token thinking budget still enables reasoning and should not be confused with thinking-off (Table S22). Temperature effects change direction across budgets, so the single-sample grid does not identify a monotone temperature benefit. The threshold scores show the same broad separation between the large input effects and smaller parameter shifts.

Table S22: Complete $N = 8/16/24$ parameter grid ($n = 1,500$ per cell), compared with the full-1,500 Qwen-max uniform-IoU^2 final mean at the same N. All parameter contrasts are descriptive; daggers mark cells with a logged infrastructure-terminal row.

Study	Setting	N	Fixed IoU	Full-1,500 control	Difference
Thinking budget	4096	8	48.00	48.20	-0.20
		16	49.02	49.70	-0.68
		24	$47.43^{\dagger}$	49.12	-1.69
Temperature	0.5	8	47.20	48.20	-1.00
		16	48.60	49.70	-1.10
		24	$50.87^{\dagger}$	49.12	$+1.75$
Temperature	1.0	8	45.91	48.20	-2.29
		16	50.50	49.70	$+0.80$
		24	$49.22^{\dagger}$	49.12	$+0.10$

D.6 Cross-Budget Synthesis of the Final Controlled Grid

Table S23 compresses the final 13-condition, 39-cell matrix into its primary differences. Cue interventions use the shared same-UID full-1,500 cue control; A15/A16 use their exact same-UID 500-question controls; A10, A19, and A21 retain the full-1,500 Qwen-max mean at the same N and remain descriptive. Each difference uses the appropriate control cohort. Parameter and repeat comparisons remain descriptive.

Table S23: Primary fixed-IoU differences across the final controlled grid. Cue and image rows use same-UID controls; parameter and repeat rows use the full-1,500 Qwen-max uniform-IoU^2 final mean. Daggers mark cells with residual infrastructure-terminal rows.

Family	$N = 8$	$N = 16$	$N = 24$
Correct colour (C)	-2.58	$-2.75^{\dagger}$	$-2.83^{\dagger}$
Correct spatial (S)	-2.32	-2.05	$+0.27$
Correct spatial+colour (SC)	-3.14	$-2.40^{\dagger}$	$-2.40^{\dagger}$
Wrong colour (WC)	-9.34	-9.90	$-8.65^{\dagger}$
Wrong spatial (WS)	-25.27	$-24.71^{\dagger}$	$-22.45^{\dagger}$
Wrong spatial+colour (WSC)	-27.22	$-26.69^{\dagger}$	$-25.00^{\dagger}$
No image (A15)	-34.58	-35.91	-35.47
Half resolution (A16a)	-6.05	-7.66	-6.65
Greyscale (A16b)	-1.25	-1.56	$-1.54^{\dagger}$
Thinking budget 4096 (A10)	-0.20	-0.68	$-1.69^{\dagger}$
Temperature .5 (A19)	-1.00	-1.10	$+1.75^{\dagger}$
Temperature 1.0 (A19)	-2.29	$+0.80$	$+0.10^{\dagger}$
A21 repeated run	-1.48	-0.56	$+0.73^{\dagger}$

Infrastructure failures do not explain the direction of the large cue effects. Removing only terminal failures logged before model completion raises affected means by 0.08–2.29 points and shifts contrasts by $+0.10$ to $+2.28$. Large cue contrasts keep their signs; small parameter effects remain unresolved. This sensitivity uses each affected cell's remaining UIDs. It does not replace the fixed-denominator results, which retain all planned questions. The largest effects are both directionally and numerically stable across budgets. Removing the image loses a paired 34.58–35.91 IoU, showing that textual priors alone recover only coarse target regions. Wrong spatial information loses 22.5–25.3, and combining wrong spatial and colour cues loses 25.0–27.2. Wrong colour alone is

less destructive but still consistently costs about 9. These patterns support a substantive conclusion: the model uses the appended language, and false location information is more disruptive than redundant truthful wording in this benchmark. Half resolution forms a second tier, with paired losses of 6.05–7.66. Greyscale is much smaller at 1.25–1.56 and remains unresolved. This suggests that spatial structure and luminance retain most of the usable evidence for the aggregate cohort, but does not prove that colour is irrelevant: the false-colour arm remains strongly harmful, and image greyscale and linguistic colour misinformation intervene on different information channels. Parameter perturbations are smaller and less monotone. Budget 4096 is modestly negative at all three budgets, while both nonzero-temperature series change sign. The A21 repeat similarly moves from a small negative primary gap at $N = 8/16$ to a small positive gap at $N = 24$: $-1.48, -0.56, +0.73$ (mean -0.44), closely matching the qualitative small-effect picture, but one repeat cannot estimate stochastic variance. Visual removal and misleading cues have larger observed effects than these parameter changes and the repeat. The smaller shifts do not establish invariance.

D.7 Targeted Cross-Configuration Replication

Gemini reproduces the broad ordering of the selected input effects. Losses are largest without the image, followed by false spatial cues, false colour cues, and half resolution (Figure S8). Exact-length coverage stays high, so widespread output failure does not explain these changes. The transformations, cue templates, cohorts, and uniform IoU^2 rule match the Qwen study. However, visual interventions use 500-question cohorts and cue tests use all 1,500 questions. The cross-intervention ordering is therefore descriptive, not a common-cohort inferential result. We do not pool providers or extend the unreplicated Qwen findings to Gemini.

Table S24: Targeted Gemini replication of four core controlled interventions. Fixed IoU^2 is measured on exact same-UID controls. Image transformations use the pre-specified 500-UID cohort per budget; cue conflicts use all 1,500 UIDs. Paired analyses support a negative effect in every displayed cell.

Intervention	N	Same-UID control	Intervention IoU	Paired Δ
No image	8	62.64	23.26	−39.38
	16	64.74	25.69	−39.05
	24	65.86	26.52	−39.34
Half resolution	8	62.64	55.52	−7.12
	16	64.74	56.63	−8.11
	24	65.86	58.08	−7.78
Wrong spatial cue	8	62.64	33.63	−29.01
	16	64.74	34.88	−29.86
	24	65.86	36.18	−29.68
Wrong colour cue	8	62.64	50.09	−12.55
	16	64.74	51.22	−13.52
	24	65.86	53.71	−12.15

E Target Selection and Contour Quality

A model can prefer the intended object yet describe its region poorly. Kimi exceeds Doubao in fixed target preference by 1.59 points but trails it in conditional target IoU by 12.61 (Table S25 and Figure S9). Qwen-max improves on Qwen-off in both preference (+29.28) and target overlap (+32.99). These examples expose different failure profiles that one aggregate IoU would hide. Conditional IoU measures region overlap, not boundary accuracy directly.

Target preference asks which instance a predicted region overlaps most, rather than whether its boundary is accurate. For each parseable answer, we compare overlap with the intended instance and every same-category distractor. Each instance uses the union of its annotated polygon parts. These masks differ from canonical G_N, so the diagnostic IoU need not equal the headline polygon score. A target is preferred only when its overlap strictly exceeds every distractor's. There is no absolute overlap threshold, confidence ranking, NMS, or cross-category matching. All results here use the best-effort archive in Section A.

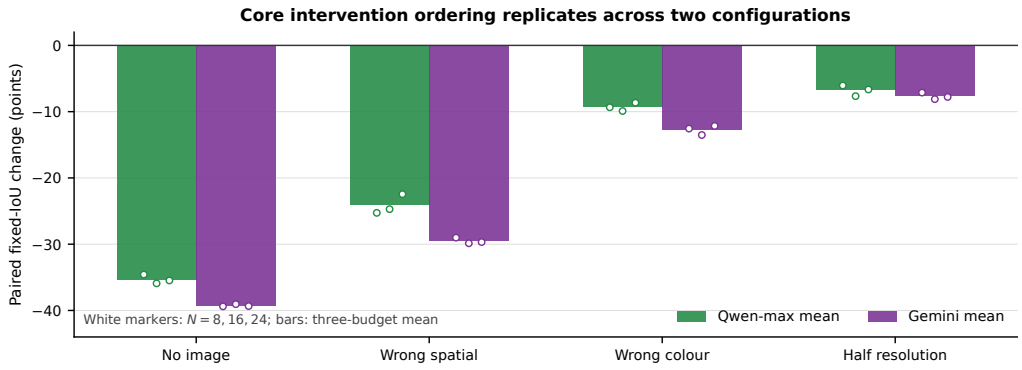

Figure S8: Cross-configuration replication of core interventions. Bars average the three reported $N = 8, 16, 24$ paired effects; white markers retain the individual-budget values from Tables S23 and S24. Across their separately declared cohorts, both configurations show the same descriptive loss ordering: no image > wrong spatial cue > wrong colour cue > half resolution.

Table S25: Target-selection diagnostic. Parse and Selected/Tie/Other are percentages; the latter three are shares of parseable outputs. "Selected" means that target overlap is strictly greater than every same-category distractor overlap under the operational definition in the main text. Contour|selected is target IoU conditional on that event. Rows use the best-effort diagnostic archive.

Configuration	Parse	Selected	Tie	Other	Fixed selection	Contour|selected
Doubao	98.0	79.0	3.5	17.5	77.45	54.69
Kimi	96.0	82.3	3.0	14.7	79.04	42.08
Qwen-off	68.2	78.4	2.6	18.9	53.47	26.59
Qwen-max	96.8	85.4	1.9	12.7	82.75	56.74
Gemini	92.0	93.1	0.7	6.2	85.64	69.71

E.1 Resolution-by-Resolution Selection and Contour Table

Successful outputs alone can hide a large loss of target coverage. Among Gemini's parseable $N = 64$ answers, 93.7% favour the target over distractors. Fixed target preference is only 62.2% because 66.4% of all answers parse (Table S26). Qwen-off is more extreme: 68.7 conditional preference becomes 3.8 with the fixed denominator. Kimi instead retains relatively strong preference but weak conditional region quality across budgets. Both denominators are needed to distinguish these behaviours.

F Vertex Budget and Polygon Geometry

F.1 High-Budget Interface Diagnostic

Structured JSON and four-part generation do not remove the observed high-budget bottleneck. We compare structured JSON and four-part generation with the free-list interface at $N = 64$ for Qwen-max and Gemini. Both alternative interfaces request the JSON payload {"coordinates":[c1,...,c128]}, with integer values in $[0, 999]$. The protocols below differ in provider support and call count, so this comparison describes complete interface configurations.

Gemini uses the Interactions structured-output field with `type:text`, `mime_type:application/json`, and `schema:S`. The schema requires one `coordinates` array, integer items bounded by 0 and 999, `minItems=maxItems=128`, and no additional properties. Qwen uses its supported JSON mode, `response_format={type:json_object}`, plus the same schema rendered in the prompt. Because that endpoint guarantees JSON syntax rather than full schema adherence, the scorer still validates length and bounds. We therefore call the joint condition structured JSON rather than a provider-invariant grammar constraint.

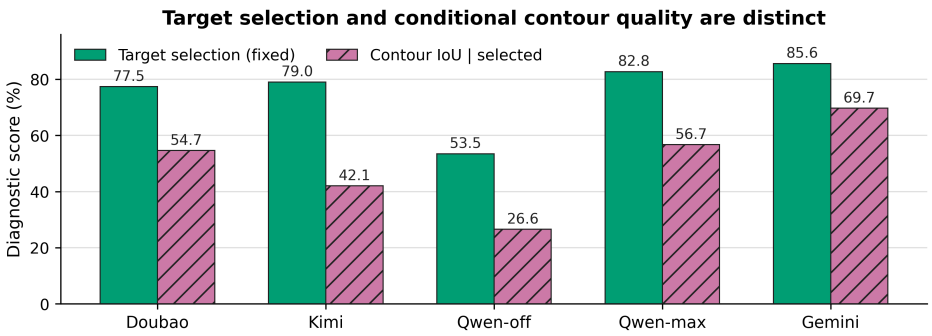

Figure S9: Target selection and conditional contour quality. Bars reproduce the pooled diagnostics in Table S25. The two quantities use different denominators and should be read as a decomposition rather than a common-score ranking.

Table S26: Target-selection and contour decomposition in all 25 configuration–N cells. "Target IoU" uses the union of all source parts, explaining its small numerical difference from canonical fixed-N IoU. Rows use the best-effort diagnostic archive.

Configuration	N	Selection fixed	Selection parseable	Target IoU fixed	Target IoU parseable	Contour|selected
Doubao	8	76.9	76.9	42.5	42.5	53.9
	16	80.5	80.5	47.5	47.5	57.7
	24	78.7	78.7	44.0	44.0	54.5
	32	80.0	80.1	44.9	45.0	54.9
	64	71.2	78.9	38.1	42.2	52.0
Kimi	8	85.5	85.5	39.1	39.1	44.6
	16	84.7	84.8	37.8	37.8	43.3
	24	83.5	83.5	36.3	36.3	42.1
	32	80.9	81.4	34.2	34.4	40.8
	64	60.7	75.1	24.6	30.5	38.5
Qwen-off	8	60.9	80.2	21.1	27.8	33.8
	16	62.6	79.4	15.0	19.1	23.4
	24	68.7	78.8	17.1	19.6	24.2
	32	71.3	76.3	18.9	20.3	25.6
	64	3.8	68.7	1.0	18.9	25.8
Qwen-max	8	85.2	85.5	49.7	49.9	57.5
	16	86.5	86.8	50.5	50.6	57.6
	24	85.7	86.1	49.7	49.9	57.4
	32	85.9	86.2	49.3	49.5	56.7
	64	70.5	82.3	38.9	45.5	54.1
Gemini	8	94.2	94.2	64.6	64.6	68.1
	16	92.4	92.6	65.8	66.0	70.8
	24	93.4	93.8	66.8	67.1	71.3
	32	86.0	91.3	60.6	64.3	70.1
	64	62.2	93.7	42.3	63.7	67.7

Each structured-JSON condition uses one stateless call per UID and no recovery. The free-list comparator is the uniform IoU[2] endpoint, while chunking uses the four calls defined below and likewise has no recovery. The table is therefore an interface-system diagnostic, not an equal-call capability comparison.

The complete added instruction is: *Return JSON only as* {`"coordinates":[...]`}. *The array must contain exactly 128 integers ordered cyclically around the visible boundary, each from 0 through 999; include no other key or text.*

Chunk protocol. Four sequential calls produce four 16-vertex segments. Each call receives the same image, expression, coordinate range, and clockwise-order instruction. Calls 2–4 also receive the accepted segments and request the next continuation. Responses use {`"segment":k,"coordinates":[32 integers]`}. Calls are stateless and have no recov-

ery. A missing segment, repeated index, wrong length, non-finite value, or out-of-range coordinate invalidates the example. Otherwise, segments are concatenated in index order and the polygon is closed. We do not delete endpoints, deduplicate, reorder, or select among alternative answers. The four calls yield one prediction per question.

Table S27: High-budget interface diagnostic on the same 1,500 UIDs. Every interface retains the fixed denominator and the same geometric target.

Configuration	Output interface	Parse	Fixed IoU	Parseable IoU
Qwen max-think	Free list	85.67	38.50	44.94
	JSON mode + prompt schema	85.73	38.23	44.59
	Four indexed chunks	86.87	38.84	44.71
Gemini high-think	Free list	59.40	37.66	63.40
	Schema-constrained JSON	59.47	37.48	63.02
	Four indexed chunks	58.73	37.70	64.19

Neither structured JSON nor chunking substantially changes the recorded coverage or geometric scores (Table S27). The similarity concerns filled IoU and parseable-only overlap. These interface rows are not used to compare strict scores, because their row-level outputs have not been aligned with the strict validator used below. These interfaces therefore do not resolve the bottleneck in the tested configurations. Because provider constraints and call budgets differ, the result does not establish that all constrained-generation methods are ineffective.

F.2 Assisted-Output Diagnostic

Boundary assistance improves region quality, but does not eliminate geometric-output failures. We use a 12-example Qwen diagnostic to interpret this observation alongside the full-cohort high-budget decline. Each example is tested at $N = 32$ and 64 with progressively more target information. The aim is to locate failures that persist under assistance, not to estimate separate recognition, perception, and serialisation abilities.

Sample and conditions. We select six simpler and six more complex single-part referents using annotation geometry, without selecting for observed model failures. Screening excludes 148 multipart cases and 46 cases with an invalid quantised oracle, leaving 1,306 feasible candidates. After sorting by $(Q, \mathrm{row_idx})$, we sample six from each of the bottom and top 435 candidates with seed 20260926. The zero-based benchmark indices are `131, 597, 991, 1056, 1367, 1496` and `25, 90, 491, 986, 1332, 1396`, respectively. These are benchmark row indices, not COCO image IDs. Table S28 defines the four conditions on these same examples.

Table S28: Information supplied in the assisted-output diagnostic. Every condition requests the same exact-N coordinate-list format at $N = 32$ and 64.

Condition	Input and requested output
A: original	Original image and referring expression; generate the target polygon.
B: target	A plus the correct target box, drawn as a thin outline; trace the object, not the box.
C: boundary	B plus a marked oracle contour, with no vertex markers or coordinate values; trace that boundary.
D: coordinate	The ordered oracle coordinates as text, without the image or expression; reproduce the list in the required format.

Matched geometry and scoring. Both budgets show the model the same boundary. We quantise the canonical 32-vertex target to the nearest even integer in $\{0, 2, \ldots, 998\}$, resolving ties downwards. The retained vertices form a simple positive-area polygon H_{32} and are distinct and in range. Inserting one midpoint per edge yields 64 distinct integer vertices with the same continuous region. The mean H_{32} fidelity to the original annotation is 98.53 IoU. Separate Pillow rasterisation of the two lists gives overlaps of 98.26–99.82, despite continuous-region equality. We therefore display the identical H_{32} boundary in both C conditions. A longer requested list adds no displayed shape detail.

For A–C, both fixed and strict IoU use the original selected-part mask and all 12 questions as the denominator. Parse failures receive zero in fixed IoU; strict IoU additionally assigns zero to every illegal polygon. A legal answer has exactly N distinct, in-range integer vertices and simple positive-area topology. D's Exact count measures equality of the ordered numeric coordinate list to the supplied list, not string-level whitespace agreement. Its copying score is not pooled with visual-grounding IoU.

Execution and provenance. A uses archived first calls, whereas the assisted outputs were collected later. The archived Qwen maximum-thinking runs date from 16–17 July 2026 and use nonstreaming requests, temperature zero, and an 8,192-token cap. We retain only the first `initial_image` call. A failed first call stays a failure even if a retry later succeeded; retained responses are checked against their first-call hashes before rescoring. Assisted conditions use the same requested settings, coordinate mapping, and parser, with one response per example and condition. Matching identifiers and settings cannot exclude provider drift. These are descriptive controls, not contemporaneous causal estimates.

In B/C, the target box identifies the referent without replacing the original expression; C additionally instructs the model to follow the marked contour. D asks for exact reproduction of the supplied ordered list. The boundary overlay has no opaque fill or coordinate annotations. The archived A records, selected identities, oracle coordinates, and original contours are retained in `sections/diagnostic12_records.json`; `sections/diagnostic12_rescore.py` reproduces A's scores. The assisted-condition aggregate results are recorded separately from these A-only raw records.

Table S29: Assisted-output diagnostic on 12 matched examples. Counts use the complete 12-example denominator. Fixed and strict IoU are on a 0–100 scale and share the original target reference. Exact applies only to coordinate reproduction; n/a denotes an inapplicable metric.

Condition	N	Parse / 12	Legal / 12	Fixed IoU	Strict IoU	Exact / 12
A: original	32	9	8	24.71	24.56	n/a
	64	6	2	19.04	12.72	n/a
B: target	32	9	8	27.08	25.54	n/a
	64	7	4	24.22	22.67	n/a
C: boundary	32	10	9	50.44	48.09	n/a
	64	9	8	48.73	47.15	n/a
D: coordinate	32	12	11	n/a	n/a	11
	64	11	11	n/a	n/a	11

Interpretation. Explicit boundary information helps more than the target box alone. Box assistance gives modest fixed-IoU gains over A. Boundary assistance raises IoU to 50.44 and 48.73 at $N = 32$ and 64, but still leaves three and four illegal or incomplete outputs. In contrast, D reproduces the supplied coordinate list exactly on 11 of 12 examples at each budget. This pattern suggests that converting a visible boundary into ordered coordinates adds difficulty beyond copying a long list. It does not show which internal operation causes the failures.

The assisted diagnostic supports a narrower conclusion than a high-budget collapse. C falls by only 1.71 points from 32 to 64. Its remaining errors show that boundary information alone does not guarantee a usable polygon, but do not establish a large residual budget effect. The conditions also change salience, instructions, and available information. Their differences cannot be added up as independent ability components. With one model and 12 examples, the study neither isolates latent perception nor establishes a general remedy.

F.3 Mask-to-Polygon Pipeline Comparison

The mask-to-polygon pipeline is more stable across vertex budgets than direct coordinate generation in the reported results (Table S30). For each UID, it predicts a mask once and deterministically converts that same prediction into three polygons, at $N = 16, 32, 64$. The converter does not use ground-truth information. The resulting polygons are evaluated against the same target definitions

with the polygon evaluator used in the main study. This is a comparison of complete output pipelines, not an isolated intervention on a model's internal capabilities.

Table S30: Direct-polygon and mask-to-polygon pipeline results. IoU uses the fixed denominator; legal coverage is a percentage. Direct scores are the reported Gemini results. Δ is mask-to-polygon minus direct fixed IoU. Each predicted mask is reused across all three vertex budgets.

	Fixed IoU			Legal coverage (%)	
N	Direct	Mask→polygon	Δ IoU	Direct	Mask→polygon
16	64.74	65.85	+1.11	94.00	98.8
32	57.83	64.12	+6.29	55.00	96.4
64	37.66	63.63	+25.97	7.53	92.9

The two pipelines have similar region overlap at $N = 16$, and their difference grows with the vertex budget. From $N = 16$ to 64, direct fixed IoU falls by 27.08 points, whereas mask-to-polygon IoU falls by 2.22. The latter pipeline retains 92.9% legal coverage at $N = 64$, although deterministic conversion does not achieve universal legality in these results. Because its mask is fixed across budgets, its remaining decline cannot be attributed to a longer model-generated coordinate sequence. Quantisation and contour-conversion failures are possible explanations, but identifying their contributions requires per-example failure analysis.

The comparison supports evaluating region prediction and geometric-output reliability separately. It does not measure target preference against same-category distractors or establish a pure serialisation effect. In particular, post-processing supplies boundary tracing and ordered vertices rather than requiring the model to emit them directly. Pre-conversion mask IoU is not reported here, so the conversion loss relative to the predicted mask cannot be quantified from this table. The results are descriptive pipeline aggregates, not a claim of equal inference cost or identical model components.

F.4 Reference Fidelity and Sampling Phase

Table S31: Reference and phase-robust configuration-gap diagnostics in the best-effort diagnostic archive. "Half-step phase" shifts the sampling origin by half an average vertex interval. Phase spread is the maximum minus minimum cohort-mean reference IoU over offsets $\{0, 12.5, 25, 37.5, 50, 62.5, 75, 87.5\}$. The final columns give the largest change in any same-N pairwise fixed-IoU configuration gap across those phases, respectively on the four-configuration full cohort and the five-configuration common-Gemini subset. Predicted-box-derived rows are not included.

N	G_N vs selected part	G_N vs all parts	Half-step phase	Phase spread	Max gap spread (4)	Max gap spread (5)
4	–	–	44.52	55.78	–	–
8	88.56	86.12	70.09	30.89	4.76	6.55
16	96.20	93.52	85.49	15.13	1.77	2.71
24	98.26	95.53	91.14	9.14	0.59	1.11
32	99.00	96.24	93.98	6.20	0.28	0.49
64	99.39	96.65	97.57	2.43	0.08	0.08

The model ordering and high-budget reversal survive changes in target starting phase. Every tested same-budget configuration gap keeps its sign across eight phases, and every configuration retains the negative $32 \to 64$ transition. Unreduced-part rescoring gives the same pooled conclusions (Table S36). Individual low-budget targets are less stable: shifting the sampling origin changes the sampled shape, not just the names of its vertices. The robustness claim therefore concerns comparative results, not phase-invariant target geometry. The main text defines $\mathrm{IoU_n}$ as the descriptive representation-fidelity ratio $100\,\mathrm{IoU_p}/C_N$. It is neither a probability nor a mathematical upper bound and may exceed 100.

Because this audit predates the matched box-derived rows, it cannot establish phase robustness of the moderate-budget direct-minus-box gains. Those contrasts and their Holm decisions remain conditional on the declared phase-zero target. Five-model values use common-Gemini subsets; $N = 4$ is source-only.

F.5 Stable-Boundary Adjacent-N Comparisons

Performance can fall even when a larger vertex budget adds no new contour detail. We restrict each adjacent-budget comparison to source polygons whose vertex count is no greater than the lower budget. Additional points then subdivide existing source edges rather than introduce new contour detail. Rounding stored coordinates and rasterising them can still produce small region differences. Table S32 reports higher-budget minus lower-budget fixed IoU on these matched subsets.

Table S32: All stable-boundary adjacent-N comparisons. Δ is higher-N minus lower-N fixed IoU on the paired rows. Rows use the best-effort diagnostic archive.

Configuration	Transition	Δ IoU	Configuration	Transition	Δ IoU
Doubao	$8 \to 16$	$+6.74$	Kimi	$8 \to 16$	$+0.66$
Doubao	$16 \to 24$	-4.76	Kimi	$16 \to 24$	$+0.82$
Doubao	$24 \to 32$	-0.21	Kimi	$24 \to 32$	-2.78
Doubao	$32 \to 64$	-6.29	Kimi	$32 \to 64$	-8.26
Qwen-off	$8 \to 16$	-1.46	Qwen-max	$8 \to 16$	$+0.11$
Qwen-off	$16 \to 24$	$+3.76$	Qwen-max	$16 \to 24$	$+0.84$
Qwen-off	$24 \to 32$	$+1.48$	Qwen-max	$24 \to 32$	-2.82
Qwen-off	$32 \to 64$	-16.89	Qwen-max	$32 \to 64$	-11.24
Gemini	$8 \to 16$	-14.97	Gemini	$16 \to 24$	-0.89
Gemini	$24 \to 32$	-3.32	Gemini	$32 \to 64$	-21.93

The $32 \to 64$ loss persists in every tested configuration on these stable-boundary subsets. Earlier transitions are mixed: Doubao declines at $16 \to 24$, Kimi and Qwen-max at $24 \to 32$, while Qwen-off improves at $16 \to 24$. The comparison therefore ties the high-budget loss to a change in output requirements rather than a need to represent additional source detail. It remains observational with respect to the model's internal mechanism.

F.6 Topology Warnings

High-budget failure has several forms: incomplete lists, invalid topology, and spatial error (Figure S10). These must be read together. Qwen-off at $N = 64$, for example, has only 83 parseable answers. Its 87.95 self-intersection rate describes that small subset, alongside parse coverage of just 5.53. Mask IoU ignores cyclic start and traversal direction. Even after choosing the best cyclic or reversed alignment, vertex error remains nonzero (Table S34).

The topology, cyclic-alignment, and point-set-distance diagnostics below read the frozen best-effort archive. The available export contains 1,500 questions per budget for Doubao, Kimi, and both Qwen settings, but 500 per budget for Gemini. These historical diagnostics must not be treated as row-matched checks of the main manuscript's reported Gemini full-cohort strict scores.

F.7 Order-Free Point-Set Distances

Ignoring vertex order does not remove spatial error. We measure symmetric Chamfer and bidirectional Hausdorff distances on parseable outputs, normalised by image diagonal (Table S35). Chamfer averages nearest-point mismatches in both directions. Hausdorff takes the worst mismatch, making it more sensitive to a misplaced extreme point. Both compare vertex sets rather than polygon edges or filled regions.

Point-set distances complement IoU but cannot replace completion and topology checks. Across the 25 model–budget cells, per-question Pearson correlations with parseable-only IoU range from -0.713 to -0.235 for Chamfer and from -0.760 to -0.244 for Hausdorff (Figure S11). The negative association is consistent with spatial mismatch even after ignoring vertex order. Yet these distances omit unparseable answers and ignore polygon edges. Gemini at $N = 64$ illustrates the limitation: low conditional point-set distance coexists with low fixed IoU because coverage and topology deteriorate.

Table S33: Displayed topology diagnostics among parseable outputs. Rates after Parse use the parseable count as denominator. Self-intersection uses boundary-clipped coordinates; degenerate and repeated points use stored pixel coordinates before raster clipping. Beyond-canvas warnings concern that pixel canvas, not the raw normalised-coordinate contract. Repeated counts any duplicate vertex. Warning categories overlap and must not be added. In the frozen row-level audit, every other clipped-geometry warning is a subset of the self-intersection event, so the aggregate "any clipped warning" rate is numerically identical and is omitted rather than shown as a duplicate column. Zero-area, consecutive-repeat, and union-warning fields remain in the machine table. Rows use the best-effort diagnostic archive.

Configuration	N	Parse	Self-intersection	Degenerate (raw)	Repeated (raw)	Beyond canvas
Doubao	8	100.00	6.40	0.00	2.13	22.47
	16	100.00	22.27	0.00	9.27	18.60
	24	100.00	30.07	0.00	7.67	23.87
	32	99.93	45.76	0.00	20.88	25.02
	64	90.27	67.58	0.74	37.00	30.21
Kimi	8	100.00	3.07	0.13	1.27	20.93
	16	99.93	12.68	0.07	5.34	20.95
	24	100.00	19.60	0.07	7.07	23.53
	32	99.33	24.77	0.34	9.80	25.84
	64	80.80	35.73	0.91	28.55	24.09
Qwen-off	8	75.93	64.18	12.20	47.85	25.46
	16	78.80	69.88	14.13	50.34	20.56
	24	87.27	71.05	7.49	46.07	20.24
	32	93.40	72.09	3.78	40.83	21.13
	64	5.53	87.95	1.20	60.24	36.14
Qwen-max	8	99.60	12.12	0.54	8.23	23.69
	16	99.67	22.14	1.47	10.90	21.40
	24	99.60	30.79	1.14	16.33	22.09
	32	99.67	40.80	1.20	28.29	21.61
	64	85.67	56.11	0.31	42.88	21.63
Gemini	8	100.00	9.00	0.00	8.00	26.60
	16	99.80	26.85	0.00	17.23	29.06
	24	99.60	37.15	0.00	20.68	26.71
	32	94.20	47.77	0.00	28.24	22.29
	64	66.40	52.41	0.00	29.22	21.69

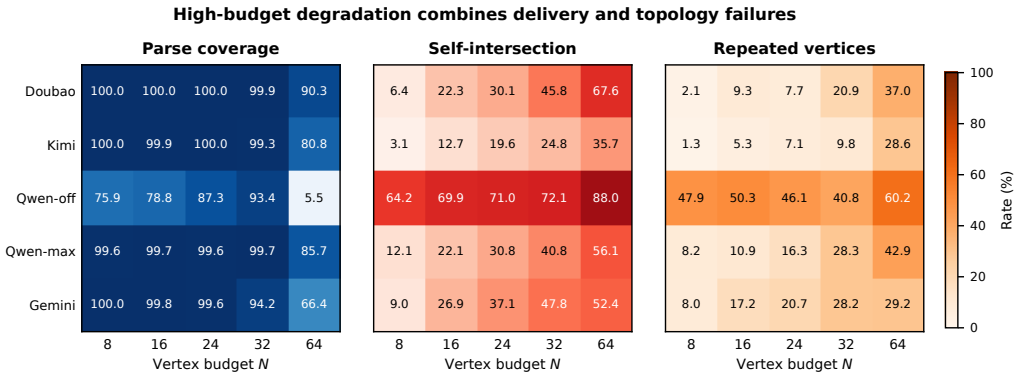

Figure S10: Delivery and topology diagnostics across vertex budgets. Parse coverage uses the complete denominator; self-intersection and repeated-vertex rates condition on parseable outputs and may overlap. The three heatmaps visualise the distinct failure channels tabulated in Table S33.

Table S34: Best cyclic-or-reversed aligned mean vertex distance normalised by image diagonal / fraction choosing reversed order, among parseable outputs. Choosing reversal only means that alignment has lower distance; it does not prove an incorrect traversal direction. Rows use the best-effort diagnostic archive.

Configuration	$N = 8$	$N = 16$	$N = 24$	$N = 32$	$N = 64$
Doubao	14.07/34.87	12.42/36.60	12.87/39.67	12.65/38.43	13.26/39.36
Kimi	12.54/37.00	12.07/38.23	12.22/38.60	12.70/39.40	14.30/37.46
Qwen-off	17.00/35.38	18.54/38.07	20.45/40.95	19.37/40.04	29.22/43.37
Qwen-max	12.01/34.87	11.27/34.98	11.14/36.41	10.99/37.12	12.03/36.34
Gemini	8.54/35.00	8.92/37.47	8.07/36.14	9.03/36.94	8.86/40.36

Table S35: Symmetric Chamfer / bidirectional Hausdorff distance among parseable outputs, normalised by image diagonal; lower is better. Rows use the best-effort diagnostic archive and its per-model, per-resolution parse denominators.

Configuration	$N = 8$	$N = 16$	$N = 24$	$N = 32$	$N = 64$
Doubao	8.77/19.56	6.66/16.50	6.63/17.00	6.26/16.53	6.29/17.35
Kimi	8.09/17.40	6.78/16.28	6.65/16.89	6.83/17.65	7.92/20.69
Qwen-off	9.05/20.64	8.69/21.23	9.42/23.32	8.99/22.90	13.05/31.51
Qwen-max	7.23/16.83	5.79/14.85	5.55/14.82	5.25/14.46	5.55/16.53
Gemini	5.05/12.84	4.33/11.79	3.54/10.13	4.01/11.21	3.10/10.46

G Reference-Mask and Backend Sensitivity

G.1 Three Reference Regions

Including all annotated parts raises scores modestly without changing the displayed ordering. Some predictions cover valid disconnected regions that the single-contour target omits. The selected-part reference is the original unreduced contour used to construct G_N. The all-parts reference fills every annotated component of that instance; it does not combine targets from different budgets.

G.2 Legality-Enforced Parsed-Polygon Track

The filled-sequence scorer is intentionally permissive, allowing exact-length parsing and downstream geometry to be analysed separately. The legality-enforced headline axis validates the coordinate list extracted by that parser. A row passes only if the list contains exactly $2N$ finite numerical values, every value is integral and lies in $[0, 999]$ before clipping, all N announced vertex pairs are distinct, and the cyclic polygon before raster clipping is nonempty, positive-area, simple, and free of self-intersection. Every failure receives zero; passing rows use the unchanged canonical Pillow IoU.

The audit validates the original parsed response before coordinate clamping. Testing only the stored clamped list would incorrectly accept some out-of-range answers. For Doubao, Kimi, and both Qwen settings, the revised strict tables are recomputed from 1,500 selected outputs per budget; their filled scores reproduce the main table. The input hashes, counts, and aggregates are recorded in `sections/strict_track_audit.json`. Gemini values retain the main manuscript's reporting view; the available 500-question diagnostic export is not used to replace them. Its full-cohort strict rows, and Kimi's primary-pass strict aggregate with a missing raw prefix, were not independently recomputed in this audit.

Because the parser may extract a bracketed list from surrounding prose or select the last explicit answer list, this axis does not enforce the prompt's one-line/no-Markdown wrapper. It enforces numeric and polygon legality after parsing, repairs no invalid geometry, and retains the complete fixed denominator.

Reference-grid validity. High raster fidelity does not guarantee a legal integer vertex list. We normalise each canonical target to $[0, 999]$ and round to the nearest integer, using ties-to-even and no repair. Under the same range, distinctness, and simple-positive-area checks, $1{,}496$, $1{,}481$, $1{,}473$,

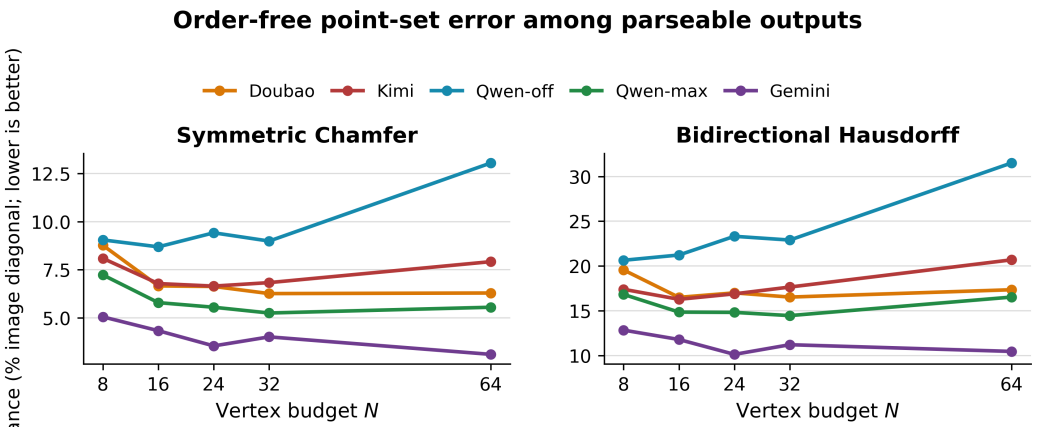

Figure S11: Order-free point-set error among parseable outputs. The two panels visualise the Chamfer and Hausdorff entries in Table S35. Gemini has the lowest conditional distances, whereas Qwen-off deteriorates sharply at $N = 64$; these conditional curves must still be read with parse coverage and polygon legality.

Table S36: Fixed IoU under three reference regions on the shared question set. The point-estimate order and substantive gaps are stable. Rows read the best-effort archive (Section A).

Reference	Doubao	Kimi	Qwen-off	Qwen-max	Gemini
Canonical fixed-N G_N	42.65	33.90	14.41	46.83	58.93
Original selected source part	42.77	33.98	14.49	46.97	59.09
Union of all source parts	43.40	34.40	14.65	47.64	60.03

1,457, and 1,416 of 1,500 targets pass at $N = 8, 16, 24, 32, 64$. Repeated vertices and nonsimple or zero-area polygons account for the failures. These counts describe this rounding rule; an invalid rounded target does not prove that no valid exact-N approximation exists. We therefore do not attribute every strict-score failure solely to the model. The full counts, failing row indices, and input hashes are in `sections/reference_grid_audit.json`.

Contract coverage below checks output length, integrality, and range. Strict coverage also checks distinct vertices and simple positive-area topology. The reported full-cohort denominator is 7,500 question–budget rows. The provenance distinction above applies to Gemini.

Table S37 and Figure S12 give the pooled contract and legality diagnostics; the following table resolves the same track by vertex budget.

Table S40 applies the same zero-on-invalid rule before thresholding.

Direct polygons retain a legal-region advantage only at the moderate budgets in this comparison. Gemini's strict gains over box-derived polygons are $+2.68/ + 4.54/ + 3.11$ at $N = 8/16/24$, followed by $-19.87/ - 51.07$ at $N = 32/64$ (Table S38). Thus extra representational capacity is useful only when the model can deliver a valid contour. Strict scoring lowers all absolute scores, especially Qwen-off's, while retaining the pooled point-estimate ordering. That stable ranking does not itself validate the task construct. Figures S13 and S14 show the budget curves and retry effects.

G.3 Pillow Raster versus Continuous-Area Geometry

The choice between raster and continuous-area IoU has little effect on already-legal polygons. Shapely minus Pillow differences are about -0.0018 to -0.0009 on the raw scale, or -0.18 to -0.09 points on the paper's scale. Within-cell correlations are 0.9999–1.0000. We use the same legal outputs for each paired backend comparison. Because legal coverage varies by model, this subset does not support a separate cross-model ranking. Table S41 reports a different analysis that repairs illegal outputs with `make_valid`.

Table S37: Central legality-enforced parsed-polygon track on the uniform IoU^2 archive. Coverage and scores use the complete fixed denominator; illegal extracted polygons are zero, never repaired. Four configurations are recomputed from raw responses; Gemini retains the manuscript reporting values, as explained above.

Configuration	Official parse	Contract cov.	Strict cov.	Strict IoU / Acc@.5
Doubao default	98.04	98.04	64.96	30.69 / 33.71
Kimi fixed-think	96.01	95.96	78.23	28.48 / 24.44
Qwen thinking-off	68.19	65.29	19.65	6.03 / 4.68
Qwen max-think	96.84	88.16	60.15	31.38 / 37.04
Gemini high-think	90.00	85.92	68.36	43.87 / 52.76

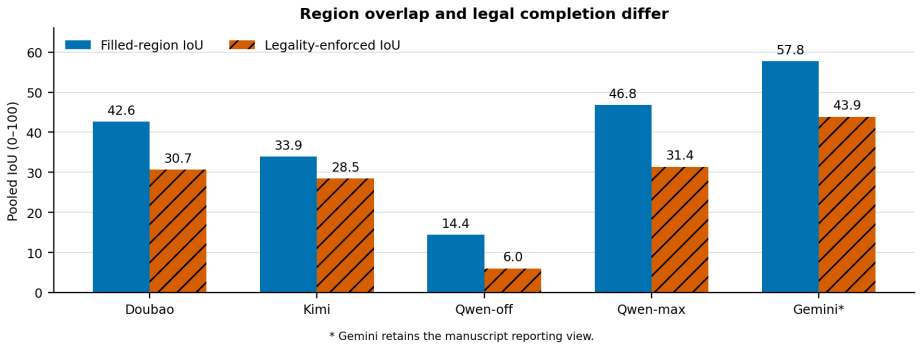

Figure S12: Filled-region utility versus legal-polygon completion. The pooled legality-enforced score is lower for every configuration because an invalid extracted polygon receives zero rather than being repaired.

Repairing an invalid polygon changes the evaluation policy, not just the integration method. This explains the larger shifts for Qwen-off and other invalid outputs. Pillow filling and Shapely repair interpret self-intersections differently, so repaired scores are not corrections to the official score. Even with repair, the micro ordering remains Gemini $>$ Qwen-max $>$ Doubao $>$ Kimi $>$ Qwen-off. We retain Pillow as the canonical scorer and treat continuous geometry as a sensitivity check. The main text's approximately one-point robustness summary refers to alternate-reference mean shifts together with ordinary Pillow-versus-continuous disagreement on already-legal polygons. The larger differences in Table S41 include polygon repair. They should not be attributed to the integration method alone.

H Additional Reproducibility Details

Compute environment. Hosted inference uses provider-managed APIs, whose accelerator and memory details are not exposed. The local preprocessing and offline-scoring pipeline is CPU-only; a basic reproduction environment is an x86-64 Linux machine with 16 GB RAM and Python 3.10 or later. The requirements file in the prepared code archive lists direct dependencies; the recorded geometry checks used Pillow 12.2.0 and Shapely 2.1.2. No local GPU is required.

Configuration labels. Doubao is provider-default because its request did not send a thinking field. Kimi is a fixed-thinking endpoint. Qwen max-thinking explicitly enables the highest thinking mode; Qwen thinking-off explicitly disables it. Gemini sets thinking level to high. The image longest side is at most 1,024 pixels. Doubao and Qwen use temperature zero, Kimi uses one, and Gemini uses provider-default temperature.

Aggregation. For a slice S, let $u_i \in [0,1]$ be each question's region IoU, with unparseable answers assigned zero. Fixed IoU is $100|S|^{-1}\sum_{i\in S} u_i$. Parseable-only IoU uses the same numerator but

Table S38: Legality-enforced performance and the matched strict comparison under uniform IoU^2 finalisation. Panel A uses the complete 1,500-UID denominator at each N; cells are strict coverage / strict fixed IoU. Panel B pairs the same 1,500 UIDs under the frozen phase-zero target; illegal direct outputs receive zero.

Panel A: strict coverage / strict fixed IoU

Configuration	$N = 8$	$N = 16$	$N = 24$	$N = 32$	$N = 64$	Pooled
Doubao	93.60/39.23	77.80/39.39	69.93/33.39	54.20/27.30	29.27/14.14	64.96/30.69
Kimi	96.93/37.26	87.27/33.56	80.40/29.76	74.80/25.93	51.73/15.92	78.23/28.48
Qwen-off	26.60/11.27	22.73/6.91	23.73/6.00	24.53/5.92	0.67/0.06	19.65/6.03
Qwen-max	77.87/39.31	71.60/39.01	63.27/33.30	53.40/28.67	34.60/16.59	60.15/31.38
Gemini	95.47/59.72	94.00/60.83	89.80/58.91	55.00/35.74	7.53/4.15	68.36/43.87

Panel B: Gemini strict direct versus predicted-box→polygon

N	Direct	Box→poly	Paired Δ
8	59.72	57.04	+2.68
16	60.83	56.29	+4.54
24	58.91	55.80	+3.11
32	35.74	55.61	−19.87
64	4.15	55.22	−51.07

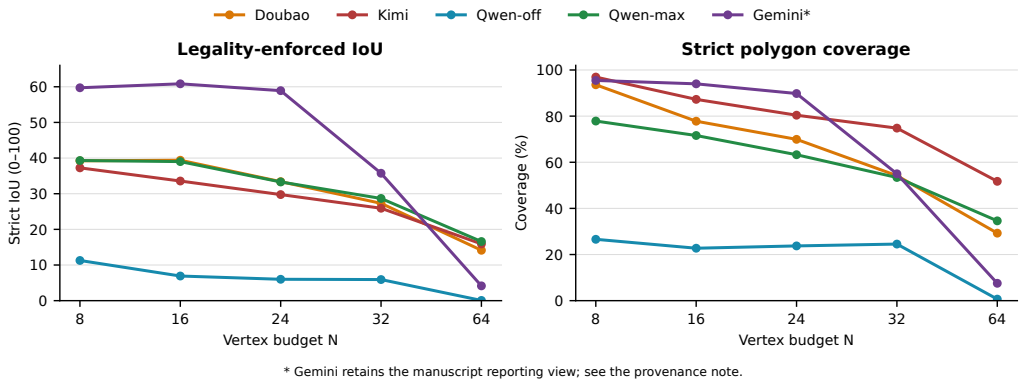

Figure S13: Legality-enforced performance across vertex budgets. Strict fixed IoU and strict polygon coverage reproduce Panel A of Table S38. Moderate-budget quality is followed by a marked high-budget legality collapse, especially for Gemini and Qwen-off.

divides by the number of parseable answers. Micro results pool question–budget rows; they do not reweight bins unless explicitly stated.

Pairing and resampling. All paired differences match exact question IDs within the same N. Prespecified inferential decisions and controlled-study direction checks use question-level paired bootstrap resampling: main analyses use 2,000 replicates, Qwen thinking-slice interactions use 1,000, and both use base seed 20260718. Five-resolution analyses and micro rank probabilities sample 1,500 UID clusters while retaining all five N rows. These resamples quantify item-sampling variation rather than generation-to-generation stochasticity or future service updates. Comparisons must use the actual shared UID set of the selected reporting view. The historical Gemini geometry export contains only 500 questions per budget and cannot be treated as a 1,500-row export. A21 supplies one repeated run at each of $N = 8, 16, 24$ as a same-UID grid check.

H.1 Controlled-Condition Implementations

Cue templates. The natural short spatial sentence is "The target's center is in the `<position>` of the image." The colour sentence is "The target has a prominent `<color>` region." (The inserted surface strings keep their original US spelling exactly as sent to the models.) The nine strings actually inserted into the natural prompt are `top-left`, `top-center`, `top-right`, `middle-left`,

Table S39: Archived primary-run versus stage-2 finalisation. Stage 0 may include internal format-recovery calls; stage 2 adds at most two external recovery stages. Parse-failure-triggered stateless retries increase coverage but never replace an already usable answer; strict columns revalidate each selected answer without repair.

Configuration	Perm. IoU^0	Perm. IoU^2	Strict IoU^0	Strict IoU^2
Doubao	40.00	42.65	29.78	30.69
Kimi	32.48	33.90	27.98	28.48
Qwen-off	10.79	14.41	4.69	6.03
Qwen-max	45.81	46.83	30.85	31.38
Gemini	49.60	57.75	36.76	43.87

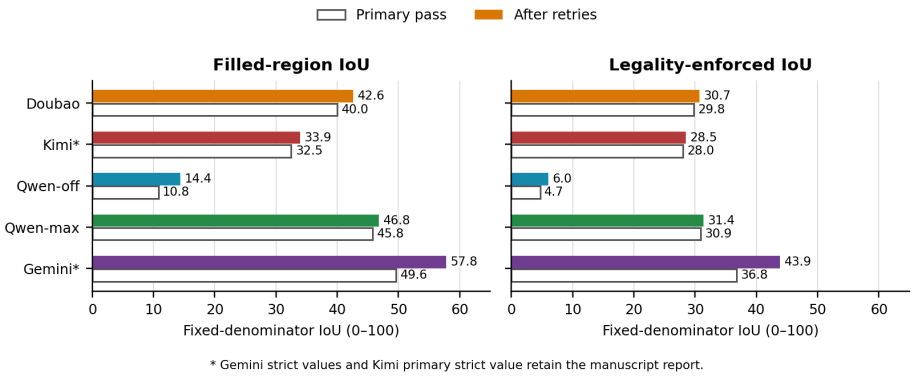

Figure S14: Primary-run and additional-recovery endpoints. Gemini and the Kimi primary strict value retain the manuscript reporting values; other strict values are recomputed. Each pair shows the primary-pass IoU^0 and uniform two-retry IoU^2 from Table S39. Retries recover previously unparsable rows but never replace an already usable answer.

`center`, `middle-right`, `bottom-left`, `bottom-center`, and `bottom-right`. Internal metadata uses an upper/center/lower naming scheme, but those internal labels are not the surface words seen by the model. Correct spatial labels come from the original target centre. Wrong spatial labels use a distant cell that is false of that target. Correct colours are reviewed visible target colours. A wrong colour occupies at most 2% of the original target mask and is preferentially more evident on a same-category distractor. SC/WSC concatenate the corresponding two sentences mechanically. The correct-cue conditions retain a single unambiguous original target. The wrong-cue prompts are deliberately internally conflicting: the original referring expression still points to target A, while the appended statement is false of A. Their score is therefore best read as original-target retention under misinformation, not as accuracy against a newly defined target B. The experiment does not require a unique alternative object satisfying the wrong cue.

Image transformations. Half resolution uses $W' = \max(1, \lfloor W/2 \rfloor)$ and $H' = \max(1, \lfloor H/2 \rfloor)$ with `Pillow.Image.Resampling.LANCZOS` in Pillow 12.2.0. The materialised image is encoded once as JPEG quality 85 with `optimize=False` and `progressive=False`; the execution path sends those frozen bytes and is not allowed to recompress them at runtime. This preserves the whole scene and aspect ratio. The prompt retains the original image-size text and the returned 0–999 coordinates are evaluated in the original frame, isolating supplied image detail from a simultaneous prompt change. Greyscale applies RGB $\rightarrow$ Pillow L $\rightarrow$ three-channel RGB at the original dimensions, preserving luminance, texture, and geometry. No-image removes the visual payload while leaving the referring text and coordinate request intact.

Parameter conditions. Budget 4096 changes only the requested reasoning budget relative to Qwen-max. Temperature .5 and 1.0 change only sampling temperature relative to temperature zero. Every question contributes one finalised selected output; protocol delivery or format-validation at-

Table S40: Strict fixed-denominator Acc@.5 by budget. Illegal outputs are zero before thresholding; the pooled column averages equal-size budget cells. Gemini retains the manuscript reporting values.

Strict Acc@.5	$N = 8$	$N = 16$	$N = 24$	$N = 32$	$N = 64$	Pooled
Doubao	41.60	44.40	36.53	30.27	15.73	33.71
Kimi	32.93	30.33	25.13	21.07	12.73	24.44
Qwen-off	11.40	5.40	3.80	2.73	0.07	4.68
Qwen-max	46.07	46.53	39.20	34.33	19.07	37.04
Gemini	75.13	74.80	73.60	38.87	1.40	52.76

Table S41: Continuous-area backend sensitivity. Geom.-valid / parseable is geometric validity among parsed rows, not central fixed-denominator strict coverage. Strict Δ is Shapely minus Pillow on that subset; repair fixed Δ applies `make_valid` on the original denominator.

Configuration	Geom.-valid / parseable	Strict Δ	Repair fixed Δ
Doubao	77.61	-0.14	-0.09
Kimi	89.40	-0.17	-0.15
Qwen-off	39.38	-0.17	$+6.29$
Qwen-max	79.84	-0.14	$+1.15$
Gemini	81.00	-0.15	$+0.32$

tempts are not independent repeated generations, and no repeated-sample vote is taken. These cells therefore measure complete hosted configurations, not isolated access to a latent reasoning trace.

H.2 Independent Max-Thinking Stability Repeat

The independent repeat stays close to the reference run, but does not establish low variance. Its mean is 48.57, a shift of -0.44 points; the largest absolute budget-specific shift is 1.48, with mixed signs (Table S42). This reduces concern that the reported scale reflects a single unusual run. One repeat, however, combines generation variability with any change in the hosted service.

H.3 Coordinate, Raster, and Identity Checks

The canonical scorer clips coordinates to the original $W \times H$ image, rasterises prediction and target with a shared Pillow binary-mask convention, and computes intersection over union. Clockwise versus counterclockwise traversal and cyclic starting-point changes leave the filled mask invariant. The separate alignment and point-set diagnostics operate on vertices before this invariance is applied; they must not be mixed into the official mask score. Each analysis joins records by model label, vertex budget, and question identity. Duplicate keys, mismatched targets, and discrepancies between stored and recomputed scores must be checked before aggregation. The four configurations with complete selected-response archives reproduce the headline filled-score means. Gemini's full-cohort reporting values remain separately identified where the corresponding row-level export is unavailable. The three-reference audit additionally joins raw annotations by dataset and annotation ID. It verifies that selected-part and all-parts masks belong to the same target instance and that canonical G_N matches the frozen question target. The target-selection diagnostic uses the same annotation inventory to enumerate same-category distractors. These joins prevent an apparently plausible score from being produced by matching a prediction to the wrong annotation.

I Scope and Broader Impact

These results are specific to the sampled COCO referents and evaluated interfaces. Training contamination and external-distribution performance are not ruled out. The phase audit supports the configuration ordering and high-budget reversal, but direct-versus-box gains remain conditional on the canonical phase. Specialist calibration reports filled-region scores only. Chamfer and Hausdorff measure vertex-set distances, not boundary accuracy. The mask-to-polygon comparison evaluates

Table S42: One independent end-to-end repeat of Qwen max-thinking at temperature zero. Differences use the full-1,500 final reference on the same questions.

N	Parse	Repeat fixed	Full-1,500 reference	Difference
8	99.2	46.72	48.20	-1.48
16	99.6	49.14	49.70	-0.56
24	99.4	49.85	49.12	$+0.73$

pipeline reliability; it does not establish a same-model, equal-call decomposition of region prediction and serialisation.

Cost and latency were not recorded consistently, so we make no efficiency claim. Annotation, accessibility, and robotics are potential uses; surveillance risks favour transparent evaluation over unattended deployment.

J QUALITATIVE EXAMPLES

The remainder of the supplement presents observed five-configuration examples spanning food, person, accessory, vehicle, animal, electronics, drinkware, and articulated-object referents. Each page fixes one image and referring expression, overlays the exact-N target and the five tested configuration outputs, and reports IoU from the same frozen evaluator used elsewhere. The first six examples cover all five budgets; later pages show only the budgets jointly available for every configuration. No unavailable configuration–budget cell is imputed, and these selected cases illustrate behaviour rather than estimate its frequency.

Figures S16 to S26 contain the complete example sequence below.

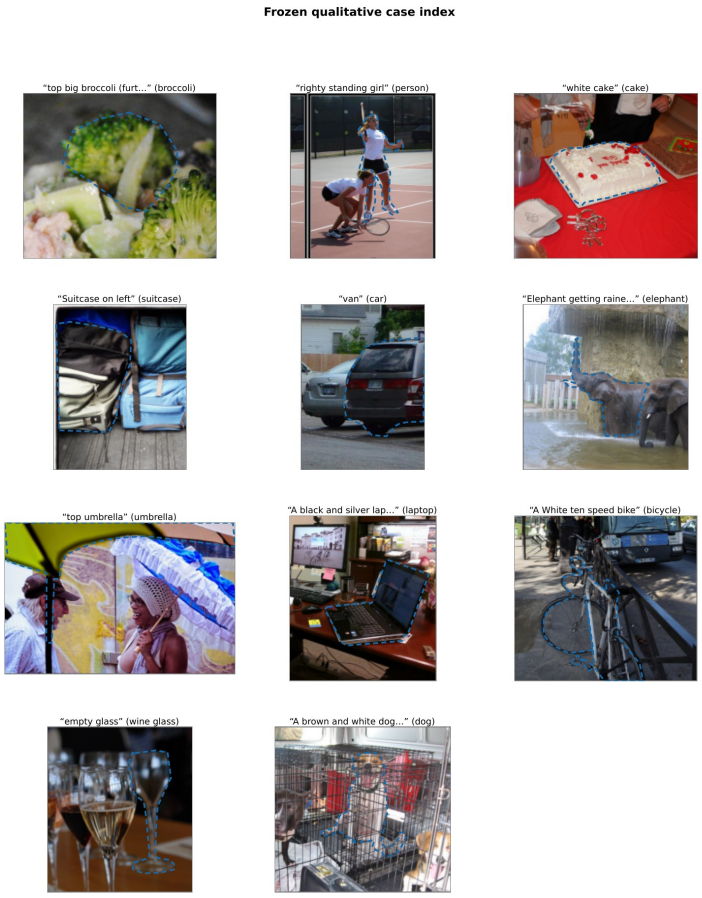

Figure S15: Index of the eleven frozen qualitative referents. Blue dashed contours show the highest jointly available exact-N target for each case. The following full-page plates compare all five configurations and every jointly available budget.

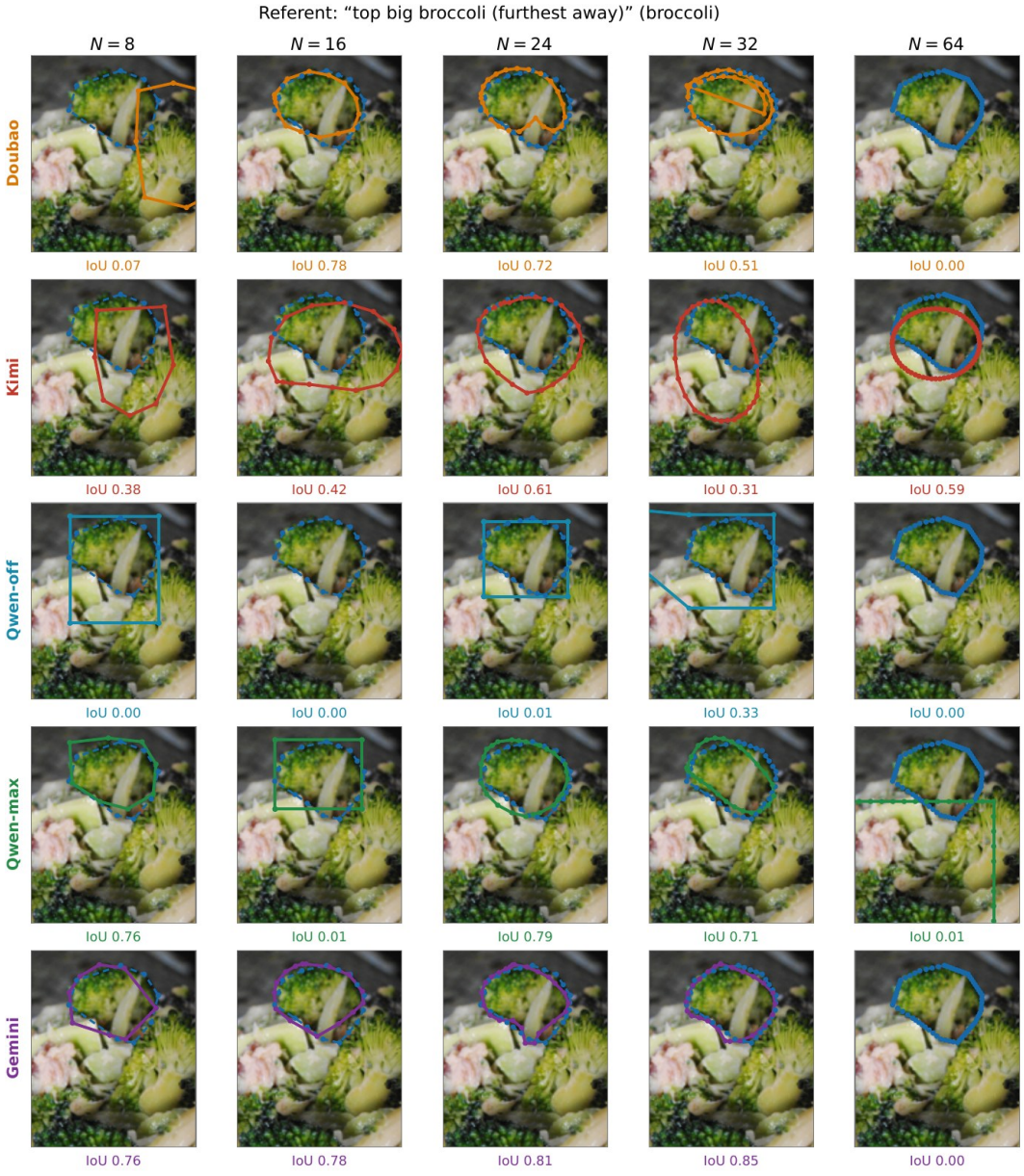

Figure S16: Five-configuration comparison on a broccoli referent. Doubao, Kimi, Qwen thinking-off, Qwen max-thinking, and Gemini predict polygons for the same image and referring expression at $N = 8, 16, 24, 32, 64$. Blue dashed contours are the exact-N targets; coloured contours are predictions from the frozen qualitative view, with IoU below each panel.

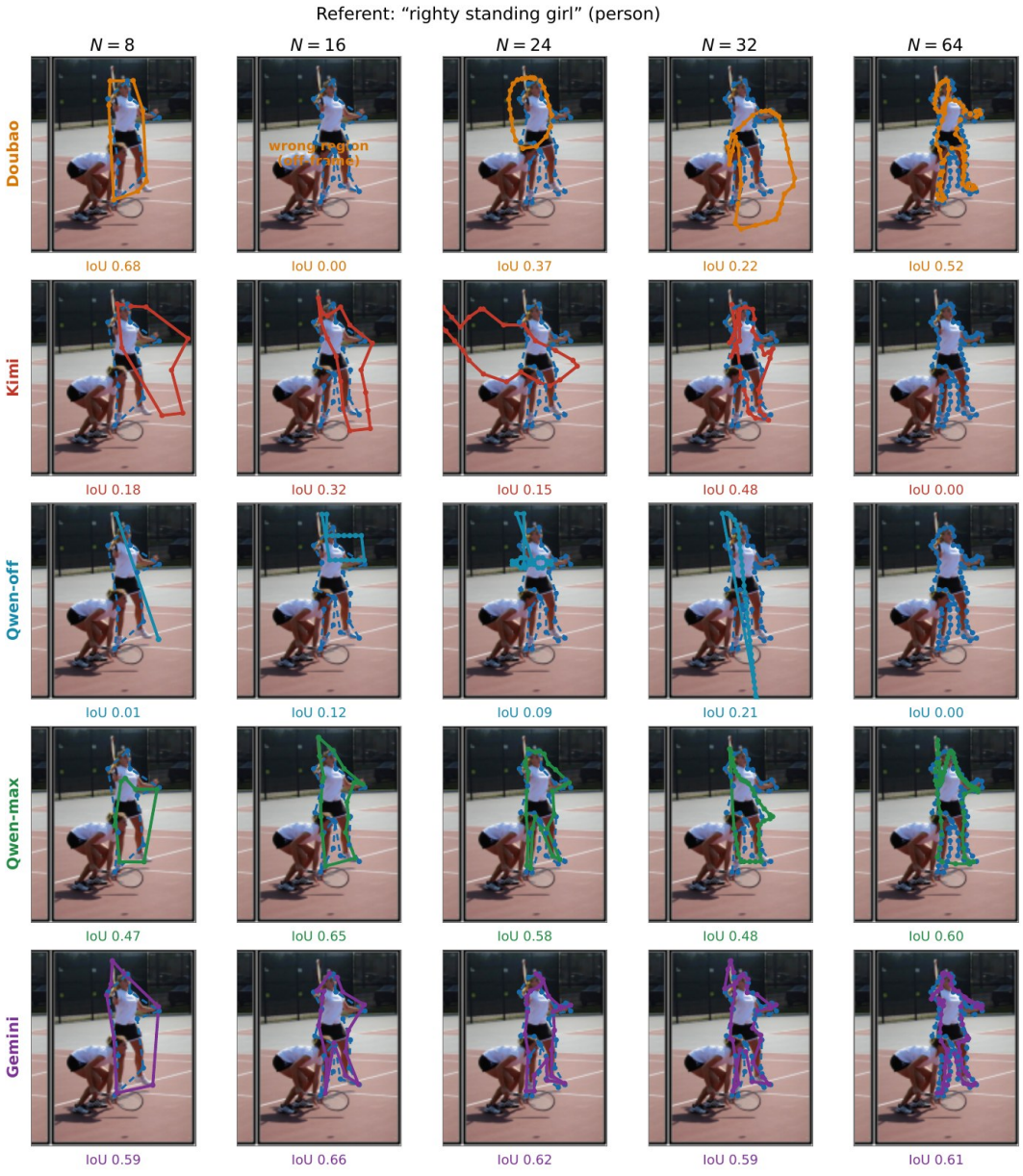

Figure S17: Five-configuration comparison on a person referent. Each row shows one tested configuration on the same image and referring expression across all five budgets. The two Qwen rows expose the same-endpoint thinking contrast.

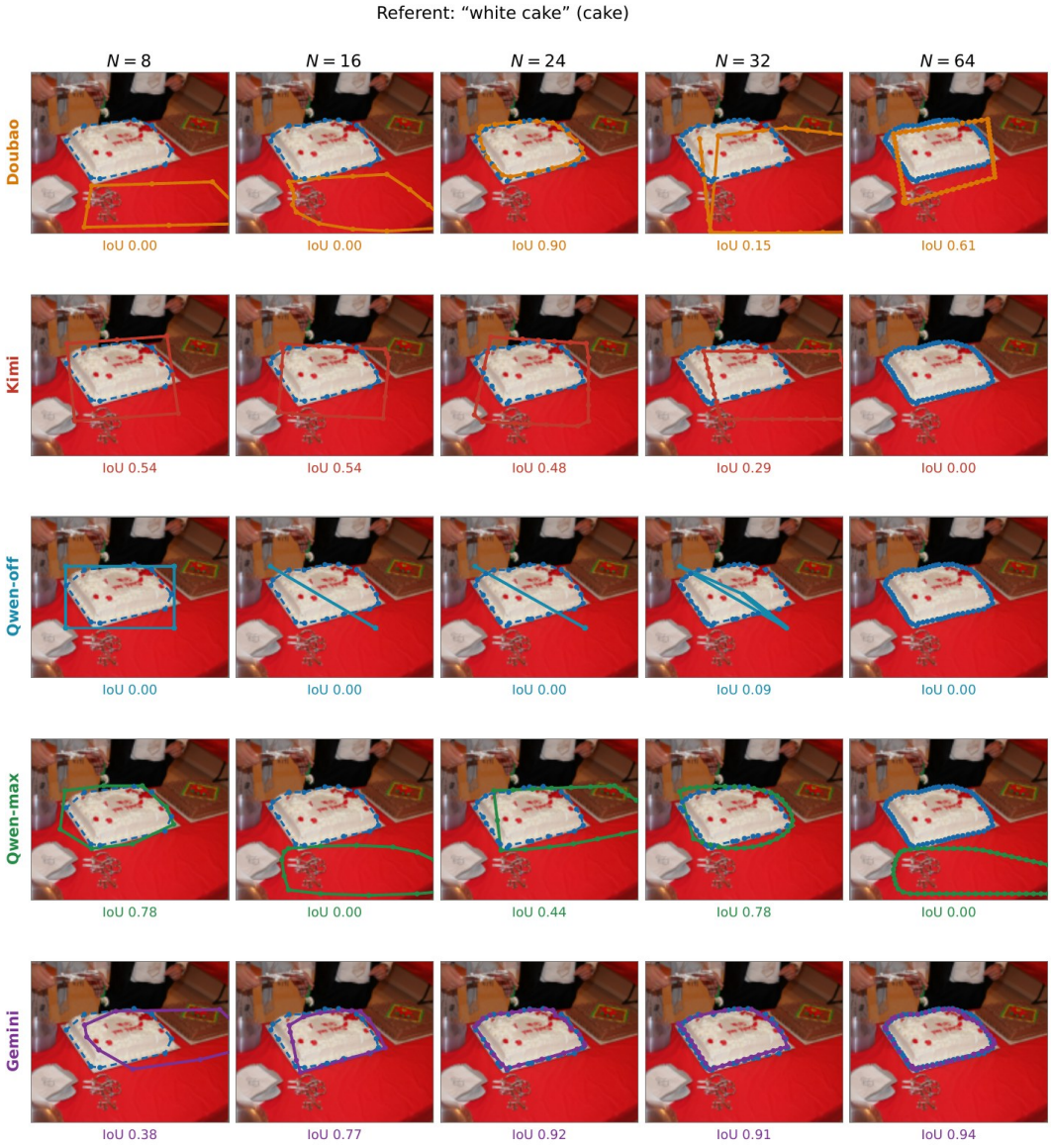

Figure S18: Five-configuration comparison on a cake referent. Gemini remains close to the target across most budgets, while the other configurations fluctuate across N. The example is illustrative; aggregate comparisons use the complete frozen evaluation cohorts.

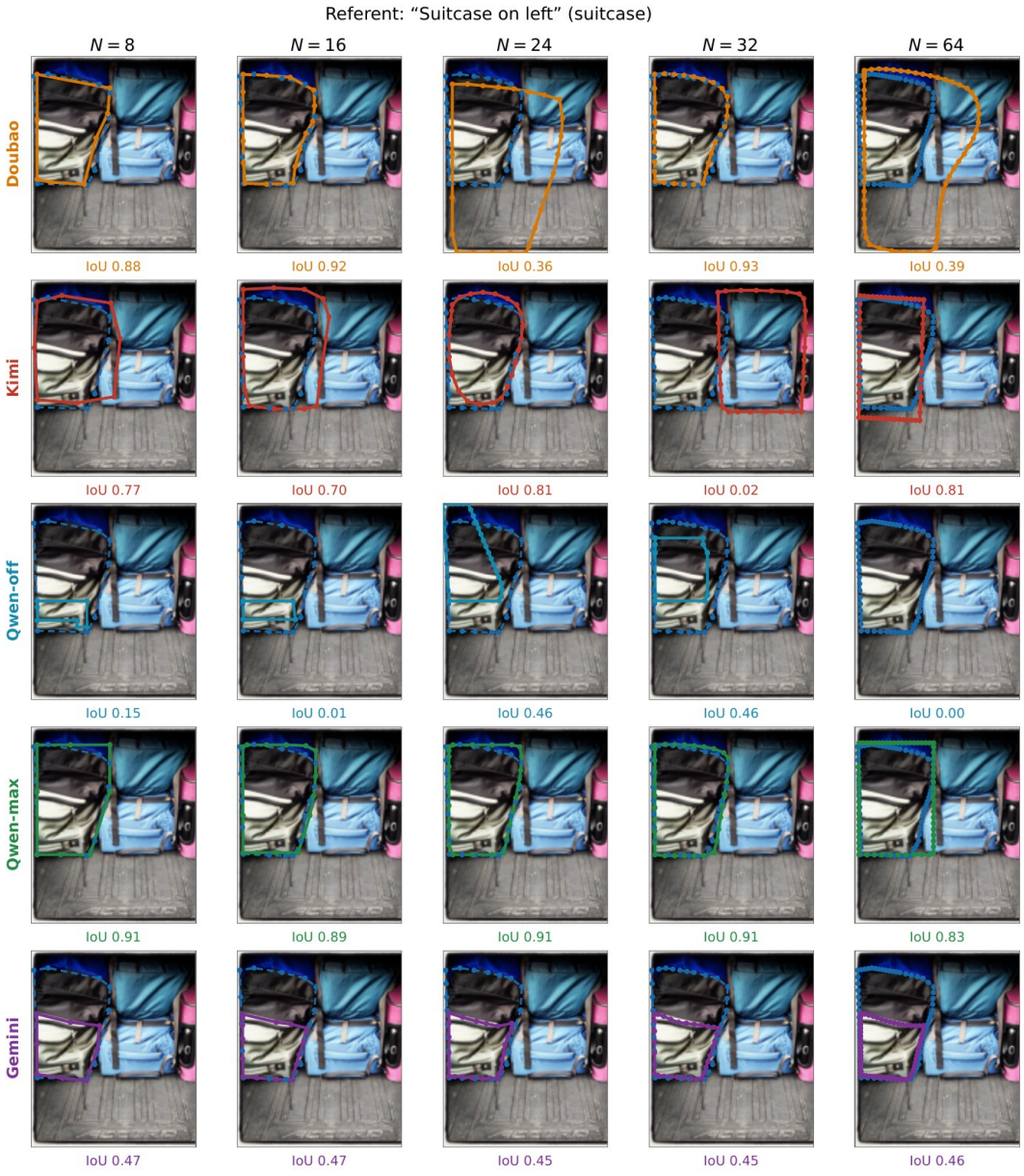

Figure S19: Five-configuration comparison on a suitcase referent. Qwen-max follows the same target consistently across the five budgets, whereas the remaining configurations show different levels of budget sensitivity. IoU values use the same evaluator as the corresponding frozen qualitative records.

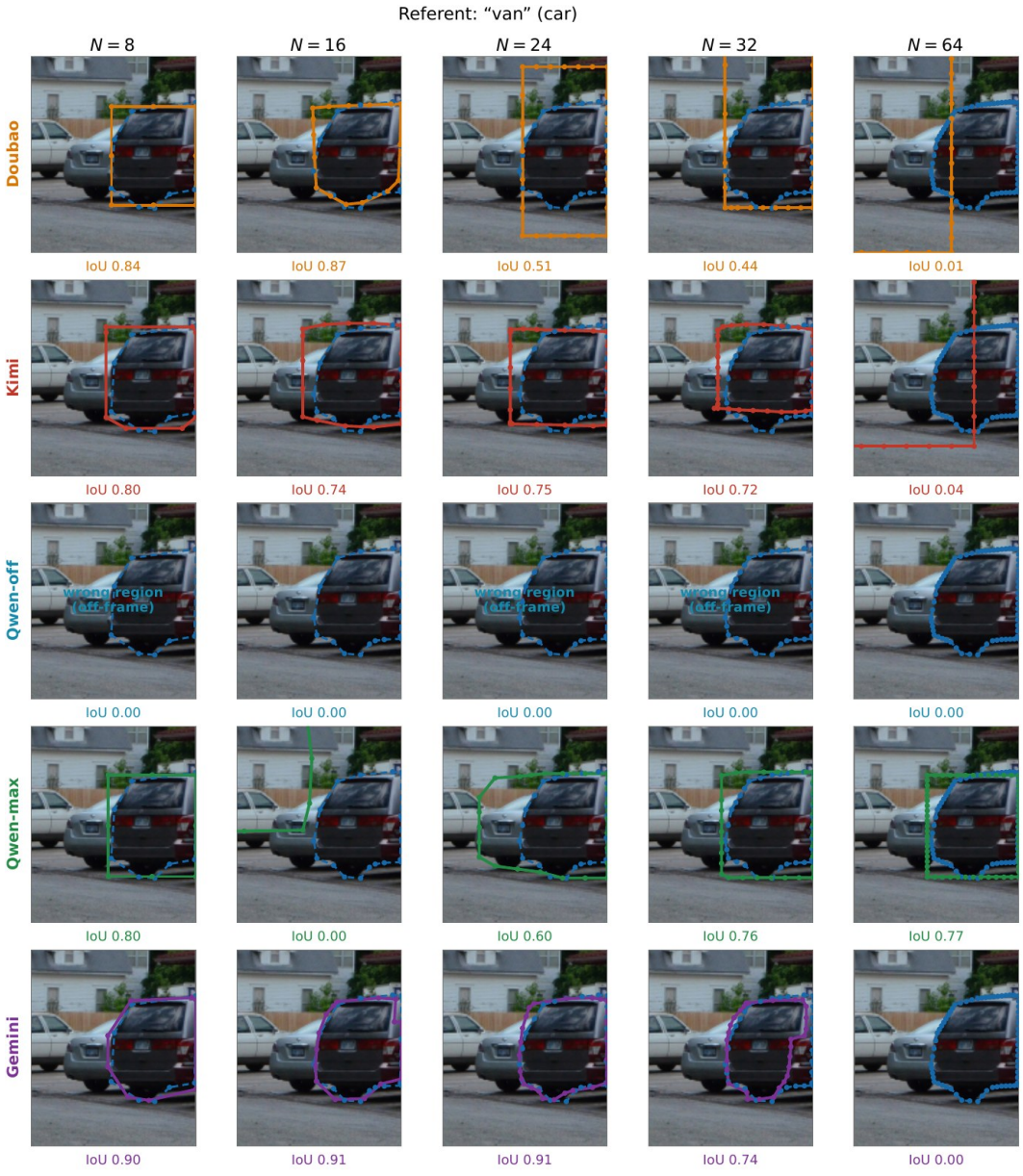

Figure S20: Five-configuration comparison on a vehicle referent. The same van is evaluated at all five budgets. Several configurations trace it well at moderate N, while their $N = 64$ outcomes differ sharply.



Figure S21: Five-configuration comparison on an elephant referent. The example contrasts stable target following with wrong-region and low-overlap polygons under the same referring expression and crop.

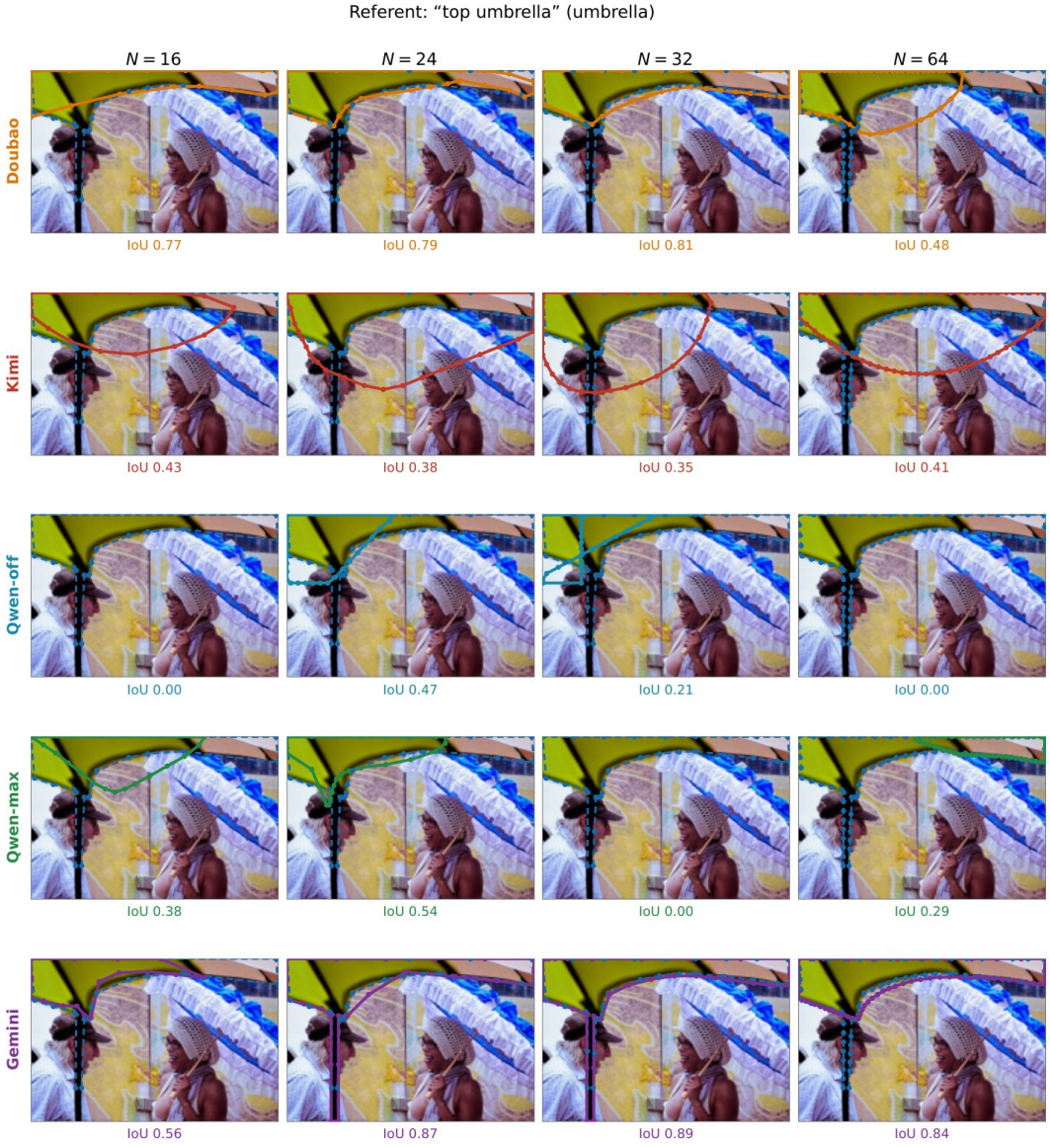

Figure S22: Five-configuration comparison on an umbrella referent. This example shows the four budgets jointly covered by all five frozen configuration cohorts, $N = 16, 24, 32, 64$. No unavailable configuration–budget output is imputed.

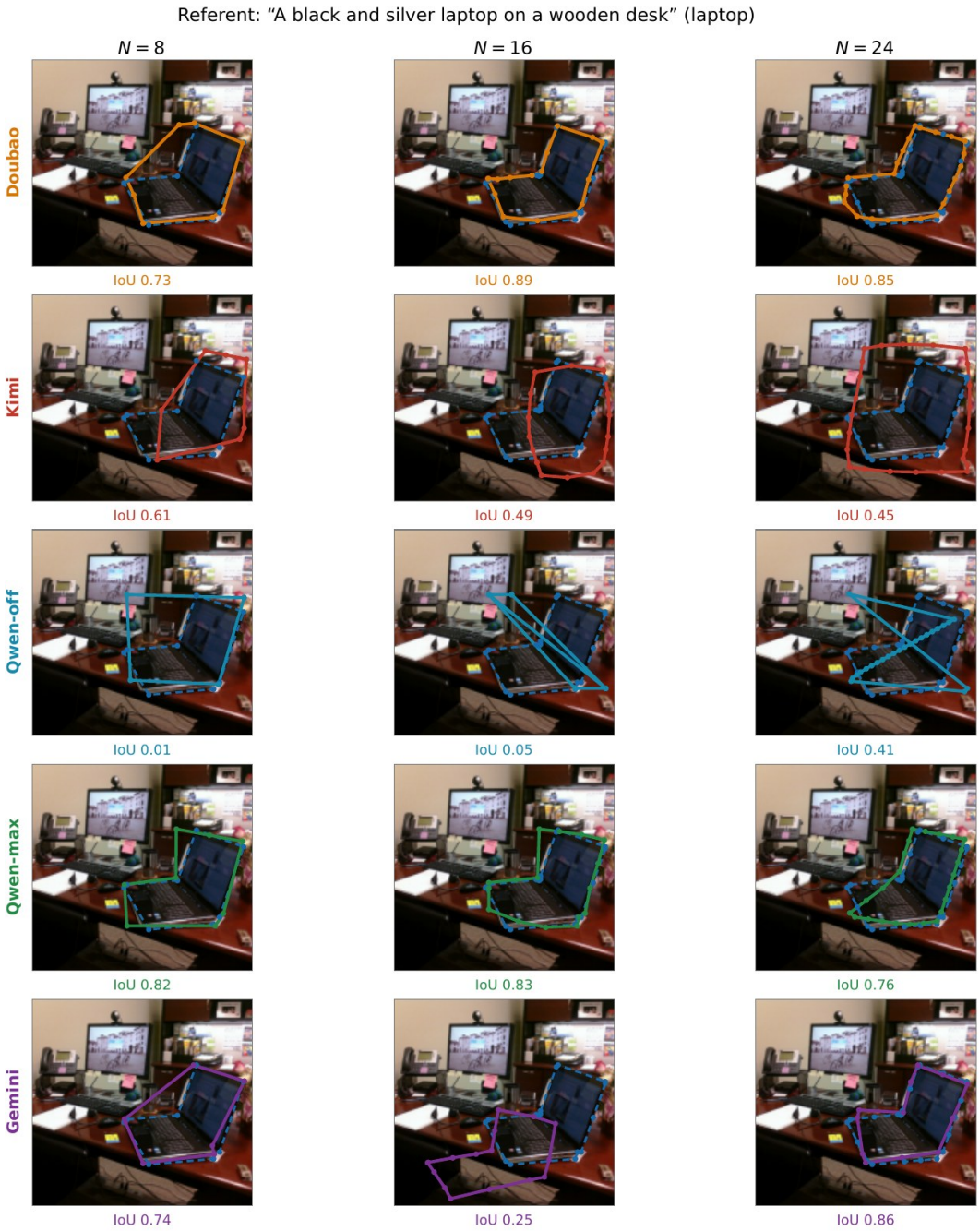

Figure S23: Five-configuration comparison on a laptop referent. The three budgets jointly available across all five frozen configurations, $N = 8, 16, 24$, are shown. Blue dashed contours are exact-N targets and coloured contours are observed predictions; no unavailable cell is imputed.



Figure S24: Five-configuration comparison on a bicycle referent. The articulated wheel-and-frame boundary is compared at the three jointly available budgets, $N = 16, 24, 64$. Blue dashed contours are exact-N targets and coloured contours are observed predictions; no unavailable cell is imputed.

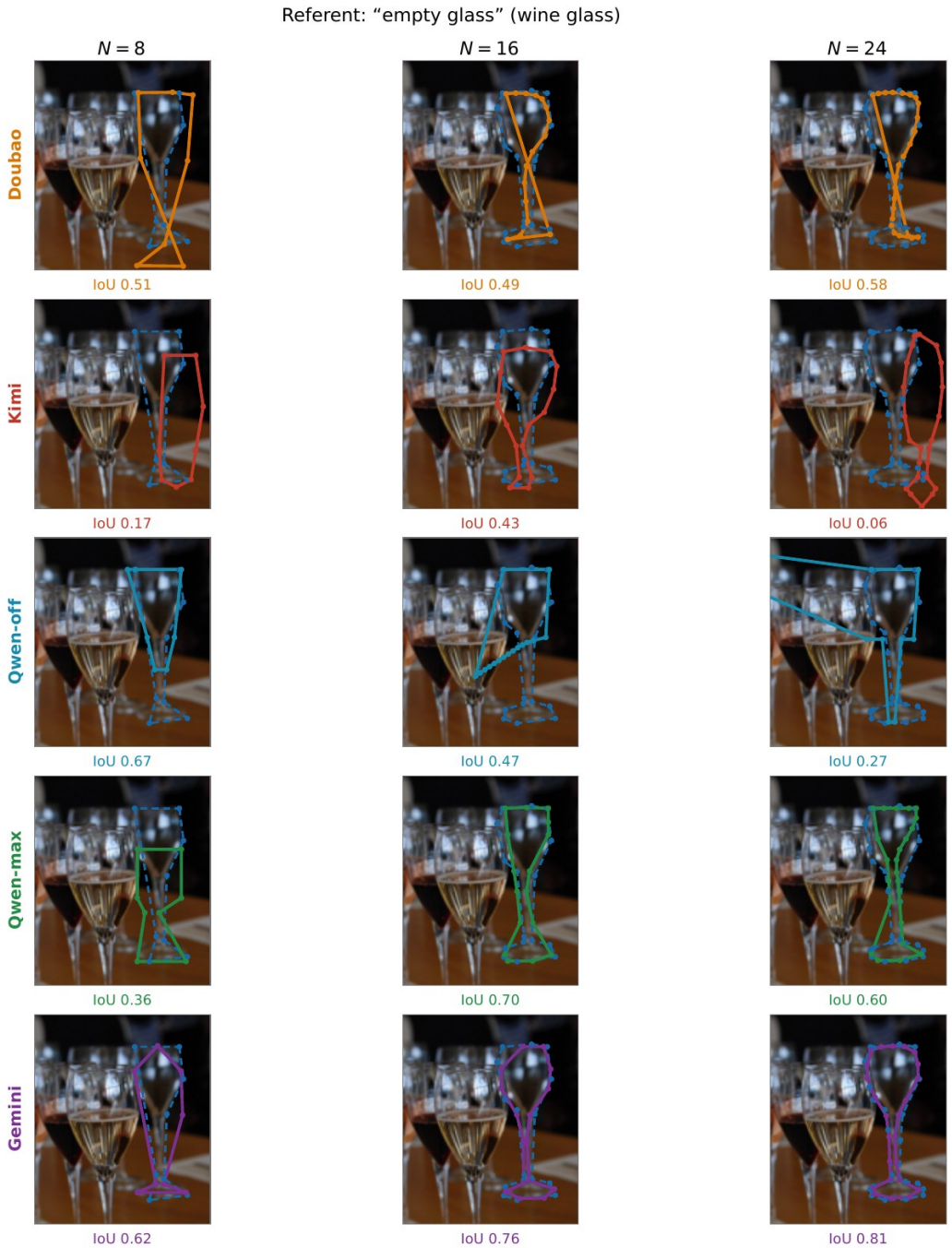

Figure S25: Five-configuration comparison on a wine-glass referent. The small transparent target is evaluated at the three jointly available budgets, $N = 8, 16, 24$, amid visually similar neighbouring glassware. Blue dashed contours are exact-N targets and coloured contours are observed predictions; no unavailable cell is imputed.



Figure S26: Five-configuration comparison on a dog referent. The collared dog is evaluated at the four jointly available budgets, $N = 8, 24, 32, 64$, under crate occlusion and same-category distraction. Blue dashed contours are exact-N targets and coloured contours are observed predictions; no unavailable cell is imputed.