WHERE GROUNDING ACCURACY LIVES ON THE IOU CURVE: LABEL-FREE INFERENCE-TIME BOUNDARY REFINEMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision–language models can identify the correct referent while returning an imprecise bounding box. We study whether a frozen direct-answer model can use its own prediction to allocate and safely exploit a localized re-observation without accessing target annotations at inference. Label-free precision refinement (LFPR) routes predicted-small regions to a higher-resolution pass, re-grounds the expression inside a context crop, admits a candidate only under fixed geometric guards, and returns a fixed coordinate-wise midpoint. We report results across three evidence tiers. On 31,921 Ref-L4 expressions – retrospective evidence, since the published test distribution informed method development – LFPR raises our primary endpoint, $\mathrm{mAcc}_{0.5:0.95}$, from 72.947% to 76.013% (Acc@0.5 88.531%→89.725%, Acc@0.9 55.788%→61.142%). A frozen, no-retuning transfer to 30,969 RefCOCO/RefCOCO+/RefCOCOg expressions – a later content-hash audit found substantial image overlap with Ref-L4 – improves every dataset individually at Acc@0.5, mAcc, and mean IoU (mAcc $+0.645$, Acc@0.5 $+0.817$, mean IoU $+0.552$ points pooled), while Acc@0.9 is not distinguishable from zero on any dataset or pooled – a stage-wise audit shows routing alone gains $+1.162$ points at Acc@0.9 while crop, guards, and fusion give back -1.192 points relative to that routed arm, so the two stages substantially offset rather than the operator having no strict-IoU effect. A prospective, image-disjoint Flickr30K Entities evaluation improves every endpoint under an MDETR-style merged-box protocol (mAcc $+0.973$, Acc@0.9 $+1.022$) and more strongly under a historical single-box selection on the same images (mAcc $+2.575$, Acc@0.9 $+3.689$). Applying the identical operator to two released grounding specialists improves every reported endpoint (Acc@0.9 $+1.569$ for EGM-4B, $+6.716$ for EGM-8B) at roughly twice the inference latency, showing that the refinement composes with specialist training rather than replacing it. On the untuned transfer and prospective tiers, a genuine unguarded control – the same candidates with the guard removed – underperforms the incumbent, showing the guard is load-bearing. Together, these results show that referent selection and boundary precision are partially separable, and that different inference components move different, sometimes opposing, regions of the IoU curve – behavior a single accuracy threshold cannot reveal.

1 INTRODUCTION

Referring expression grounding asks a vision-language model (VLM) to map a natural-language description to the image region it denotes. Modern VLMs are often strong enough to identify the correct object while still returning a box that is too coarse, especially for long expressions, small objects, and images whose native resolution falls below the model's effective visual input floor. The distinction matters because a small coordinate error can leave an Acc@0.5 outcome unchanged while causing a failure at stricter thresholds: a method improving only headline Acc@0.5 may hide a useful localization change, and one improving only strict thresholds should not be sold as a general accuracy gain.

The usual responses – train a grounding head, add a detector, or spend more computation on every image – each change the comparison or incur a substantial deployment cost. We study a narrower question: can a frozen direct-answer VLM use its own first prediction to protect a localized second observation from degrading an already strong incumbent? The question is deliberately operational,

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.

since the routing signal must be available before the ground-truth box is known. We therefore stratify our evidence into three tiers: a retrospective Ref-L4 analysis whose published test distribution informed method development, a frozen cross-dataset transfer to the RefCOCO family that retuned nothing, and an image-disjoint Flickr30K confirmation whose images and annotations played no role in method development.

Our procedure, label-free precision refinement (LFPR), has two components. First, it applies pixel-area size buckets to the incumbent's *predicted* box; predictions below the small-object threshold are re-evaluated at a higher image-pixel budget, and all other rows keep the default-resolution answer. Second, the resulting incumbent is re-grounded inside a context-expanded crop, where a candidate is admitted only if zoom-consistency and geometric guards pass, and the final box is the coordinate-wise midpoint of incumbent and candidate. That midpoint is a fixed convex combination chosen before the reported run: it lets the crop adjust boundaries without letting one noisy second answer replace a strong first one.

The main result is positive but bounded. On the analyzed Ref-L4 test split LFPR improves the frozen Qwen3-VL-8B (Qwen Team, 2025) baseline from 88.531% to 89.725% at Acc@0.5, with the largest differences at Acc@0.75 and Acc@0.9. Because the same published test distribution informed diagnostics and method development, we treat that estimate and its interval as retrospective; it now exceeds a cross-protocol 89.5% comparison point, but the comparison remains cross-protocol and parameter-asymmetric (Section 6.1), so we make neither a confirmatory nor an absolute state-of-the-art claim.

This paper makes three contributions within that evidence boundary:

1. We formulate a label-free, inference-only precision-refinement pipeline whose accept/reject decision uses only geometry computed from the model's own outputs, and show it improves a frozen generalist, two released specialists, and an image-disjoint prospective benchmark without any weight update.
2. We show that referent selection and boundary placement are separable capabilities that trade off along the IoU threshold axis: among released specialists on identical rows, the best-Acc@0.5 system has the worst Acc@0.9 and mean accuracy, so a permissive-threshold ranking can invert the localization-quality ordering.
3. We decompose the frozen operator on identical rows: resolution routing gives the most transferable strict-IoU gain, while local re-grounding, admission, and fusion have dataset-dependent marginal effects. Removing the admission guard makes the matched crop candidates noncompetitive, showing protected re-observation is essential.

2 Related Work

Referring-expression benchmarks such as RefCOCO and RefCOCOg established the standard single-region grounding setting (Yu et al., 2016; Mao et al., 2016); Ref-L4 revisits it for contemporary multimodal models with longer expressions and a stricter re-annotated protocol (Chen et al., 2025a). We use the official Ref-L4 evaluator and report the complete threshold profile rather than only Acc@0.5.

Supervised REC ranges from end-to-end text-conditioned detection (Kamath et al., 2021) to large-scale grounded pretraining (Li et al., 2022), with recent work on dynamic visual routing (Chen et al., 2025c) and zero-shot REC via debiased attention and programmatic relation reasoning (Chen et al., 2025b). Language-guided open-set detectors (Liu et al., 2024; Minderer et al., 2023) offer another route, but adopting one adds a second model, a proposal cache, and a selection policy. All of these change supervision or add machinery; LFPR keeps one direct-answer checkpoint frozen and changes only its inference path.

LFPR is structurally a coarse-to-fine cascade (Dai et al., 2025), but every stage reuses the same frozen weights and consistency between stages is enforced by fixed midpoint fusion and geometric guards rather than a learned loss (a directly trained adaptation did not improve on the frozen baseline; Appendix 5.11).

The closest family is test-time self-correction. Iterative loops that ask a model to revise its own box can show gains that depend on an oracle stopping rule and shrink or reverse without ground-truth access – the "self-correction mirage" (Tripathy and Krishnan, 2026); related work scales test-time perception through explicit zoom (Jiang et al., 2026) or an entropy-guided crop-and-reground loop (Gröpl et al., 2026). The closest training-free method on the same benchmark family is Chain-of-Caption, which supplies additional visual and textual context through tool use (Pang and Oh, 2026);

it improves context rather than the box itself, so it is complementary rather than a substitute for a geometry-based accept/reject step. Its official release provides an NVILA-8B/RefCOCO path, but not a released path for our Qwen3-VL backbone; we therefore treat cross-paper numbers as non-matched and reserve a matched NVILA comparison for the same-backbone evaluation described in Appendix 2.17. LFPR differs in three respects: at most one crop attempt rather than a loop, an accept/reject decision computed only from the geometry of the two candidate boxes, and a fixed conservative midpoint rather than outright replacement. A matched zero-shot self-correction control on the same incumbent regresses the headline metric (Appendix 5). A same-backbone mechanism control substitutes a frozen, label-free confidence threshold for the geometry guard on the same admission point – not a reproduction of EGG or TTSP (different backbone and closed-set interface respectively; Appendix 2.17) – and the confidence-gated arm does not recover even the guard's modest Acc@0.5 gain over routing (Appendix 3), evidence for the accept/reject step's specifically geometric criterion rather than any label-free filter.

Efficient Grounding Models (EGM) instead scale generated reasoning tokens for small VLMs and report strong accuracy–latency results on the same backbone family (Zhan et al., 2026); they are our released-specialist comparison in Section 5.3.

3 Problem Formulation and Method

3.1 Direct-answer incumbent

Let I be an image and q its referring expression. A frozen VLM f_θ returns a normalized box $b_0 = f_\theta(I, q) \in [0, 1]^4$, or an invalid output that is retained as an incumbent miss. For a ground-truth box y, the threshold utility is

$$u_t(b, y) = \mathbb{1}\{\mathrm{IoU}(b, y) > t\}. \tag{1}$$

Averaged over a population, $A(t) = \mathbb{E}[u_t] = \Pr(\mathrm{IoU}(b, y) > t)$ is the survival function of the IoU distribution, so $\mathbb{E}[\mathrm{IoU}(b, y)] = \int_0^1 A(t)\, dt$: an intervention can raise $A(t)$ at one threshold while lowering it at another, which is the lens we use throughout Section 5 and Section 5.2 rather than reporting a single operating point. The strict inequality is the official Ref-L4 convention. We use $\mathrm{mAcc}_{0.5:0.95}$, the mean of the ten values $100\,\mathbb{E}[u_t]$ over $t \in \{0.50, 0.55, \ldots, 0.95\}$, as the manuscript's primary summary because it captures the reported IoU profile. This designation is retrospective and does not confer confirmatory status. We report $\mathrm{Acc}_{0.5} = 100\,\mathbb{E}[u_{0.5}]$ as the standard benchmark operating point, with $\mathrm{Acc}_{0.75}$ and $\mathrm{Acc}_{0.9}$ as strict-IoU diagnostics.

3.2 Label-free resolution routing

Given an incumbent box and the original image dimensions (W, H), convert the box to pixel coordinates and define

$$s(b_0) = \sqrt{\max(0, x_2 - x_1)\max(0, y_2 - y_1)}. \tag{2}$$

The frozen bucket function is small if $s < 128$ pixels, medium if $128 \le s \le 256$, and large otherwise. LFPR routes only the small bucket to a second full-image inference with the already validated minimum-pixel budget of 4,194,304. Let b_0' denote this second answer for flagged rows and $b_0' = b_0$ otherwise. Crucially, the bucket is computed from b_0, not y; the ground-truth box is used only for evaluation.

This router is a label-free proxy for a resolution deficit, not an estimate of correctness: it can route a wrong large prediction or miss a small target hidden by a large incumbent box. The predicted-box router agrees with the annotation-derived bucket on 88.85% of pilot rows and 89.83% of official rows – a reasonable, though imperfect, substitute for the oracle diagnostic it is designed to approximate without reading a label.

The routing decision raises the pixel floor rather than lowering a ceiling: almost no small-bucket image already exceeds the higher pixel budget, so an upper-limit policy would be mechanically inert here, and raising the minimum pixel budget is the only lever that changes the visual evidence for these rows.

3.3 Local re-grounding and conservative fusion

For $b_0' = (x_1, y_1, x_2, y_2)$, let $c_x = (x_1 + x_2)/2$ and $c_y = (y_1 + y_2)/2$. The context crop $C_\gamma(b_0')$ has width $\min\{1, \gamma(x_2 - x_1)\}$ and height $\min\{1, \gamma(y_2 - y_1)\}$ with $\gamma = 2.0$; its center is shifted only as needed to keep the window inside $[0, 1]^2$. Pixel crop boundaries use floor for the upper-left and

ceiling for the lower-right before clipping to the original image. The frozen VLM is asked to locate the expression inside this crop and emit crop-normalized coordinates. The candidate is mapped affinely back to the original image, producing b_c.

Let a be the center of b'_0 after mapping the incumbent into crop coordinates, and let $c(b_c^C)$ be the center of the crop-coordinate candidate. The zoom-consistency distance is

$$z(b_c, b'_0) = \frac{\|c(b_c^C) - a\|_2}{\sqrt{1/2}}. \tag{3}$$

The crop candidate is accepted only if all three guards pass:

$$z(b_c, b'_0) \leq 0.35, \tag{4}$$

$$\mathrm{IoU}(b_c, b'_0) \geq 0.25, \tag{5}$$

$$0.5 \leq A(b_c)/A(b'_0) \leq 2.0, \tag{6}$$

where A is box area. If a guard or parser check fails, b_c is discarded and the pre-crop box is retained; an invalid first pass is retained as a miss. No coordinate is synthesized and no row is removed at any stage. The final prediction is

$$\widehat{b} = \begin{cases} \frac{1}{2}b'_0 + \frac{1}{2}b_c, & \text{if all guards pass,} \\ b'_0, & \text{otherwise.} \end{cases} \tag{7}$$

No score or target box is used in the routing or fusion decision, and every model output is parsed and canonicalized before use: coordinates must be finite, ordered, inside $[0, 1]^2$, and define positive area. Appendix 1 gives the exact prompts, pixel budgets, and clipping conventions. Historical records state that the fusion weight, crop context, resolution setting, and guards were fixed before the final Ref-L4 run, but the broader Ref-L4 development chronology is not independently auditable, so we treat the Ref-L4 intervals as retrospective (Section 4); RefCOCO-family is frozen no-retuning transfer with audited overlap, and Flickr30K is the prospective image-disjoint confirmation.

3.4 Complete procedure

Appendix 4 states the full pipeline in pseudocode and Figure 6 as a data-flow diagram. Every decision after the first incumbent prediction is a deterministic function of already-observed model outputs and image geometry; no step consults the ground-truth box y, which appears only in the evaluator. For a valid incumbent, LFPR uses two model calls on the roughly four in five rows the router does not flag (the initial full-image incumbent inference and one crop re-grounding inference), plus a third call for the higher-resolution full-image pass on routed rows before the crop stage. Invalid first-pass outputs (an incumbent miss) short-circuit the procedure to a single forward pass.

4 Experimental Protocol

4.1 Benchmark, model, and splits

We use the complete Ref-L4 test split: 31,921 expressions over 9,467 unique images, with COCO and Objects365 (Shao et al., 2019) source labels retained for stratified analysis. The baseline and all treatment predictions use the same frozen Qwen3-VL-8B checkpoint, prompt family, parser, coordinate convention, and official evaluator. The method is not trained on Ref-L4.

The shared 2,000-row crop/refinement pilot (1,713 images) was deterministically selected from the published Ref-L4 pool after excluding images in named earlier pilots, then reused first for conservative crop fusion and then for complete LFPR, so it is not an untouched confirmation set. Annotation-size and source diagnostics on the complete published test distribution also influenced the research program before the final 31,921-row run; "one read per method" does not undo this program-level adaptation. We accordingly call both the pilot and full-split results retrospective and use the image-cluster bootstrap to describe sampling variability under the frozen comparison, not a confirmatory error rate (Appendix 1.4 records what can and cannot be reconstructed from the archived manifests).

4.2 Implementation details

All arms use the same frozen checkpoint, prompt, parser, and deterministic decoding, so a difference between arms cannot be attributed to a prompt or parsing change; no weights are updated and no adapter is merged. Missing dimensions, malformed boxes, and invalid first-pass outputs are never

Evidence tier	n (expr / img)	B Acc@0.5	CRG Acc@0.5	ΔAcc@0.5 [95% CI]	ΔmAcc [95% CI]
Ref-L4 (retrospective)	31,921 / 9,467	88.531	89.725	+1.194 [0.944, 1.439]	+3.066 [2.818, 3.311]
RefCOCO family (frozen transfer)	30,969 / 3,982	89.212	90.029	+0.817 [0.639, 1.001]	+0.645 [0.469, 0.827]
Flickr30K (prospective, merged-box)	14,481 / 999	74.864	75.678	+0.815 [0.299, 1.324]	+0.973 [0.648, 1.310]

Table 1: Evidence-tier summary: CRG against the frozen pre-resolution incumbent B, same rows, image-cluster bootstrap intervals. Ref-L4 is retrospective; RefCOCO family is frozen no-retuning transfer with source-image overlap, while Flickr30K is prospective and image-disjoint. Full metric detail is in Table 2 and Appendix 2.1.

converted into a synthetic small box: they remain baseline misses and pass through the router, crop, and fusion stages unchanged. This preserves matched denominators. Appendix 1 lists the exact prompt, parser grammar, preprocessing constants, and the environment fields that were not recoverable.

4.3 Baselines and controls

We compare the frozen default-resolution answer, the annotation-stratified resolution diagnostic, conservative crop fusion without the label-free router, and the complete composition. The annotation-stratified arm is a diagnostic, not a deployable baseline: it uses ground-truth target size only to quantify resolution headroom, even though it reaches 89.22%. The crop/fusion-only arm tests whether the router contributes beyond conservative crop fusion. Ten further closed controls – including direct supervised adaptation, a released third-party transfer checkpoint, a full-parameter fine-tune, iterative unguarded zooming, and placement/self-consistency controls matched to the crop arm's call count – are reported in Appendix 3, ruling out that the crop's gain comes merely from a second look rather than an informed crop location. Every comparison is paired on identical expressions and resampled by image cluster (Appendix 6). We present $\mathrm{mAcc}_{0.5:0.95}$ as the manuscript's primary summary and Acc@0.5/0.75/0.9 alongside it. This hierarchy is a retrospective presentation choice for Ref-L4; for frozen transfer analyses, each table states its own Holm correction family.

4.4 Statistical analysis

For every arm we compute paired row-level differences while resampling whole images with replacement. For an image j carrying expressions $i \in \mathcal{I}_j$, let m_j^* be its multiplicity in a bootstrap draw. The treatment effect is

$$\widehat{\Delta}_t^* = \frac{\sum_j m_j^* \sum_{i \in \mathcal{I}_j} [u_t(\widehat{b}_i, y_i) - u_t(b_{0i}, y_i)]}{\sum_j m_j^* |\mathcal{I}_j|}, \tag{8}$$

which estimates the mean effect for a randomly chosen expression while using image clusters as the resampling units. We use 10,000 bootstrap replicates of (8) and report two-sided percentile 95% intervals on $100\widehat{\Delta}_t$. For the guard×fusion factorial, the per-dataset RefCOCO-family decomposition, and the IoU-threshold curve, every metric within one comparison is computed from the same seeded resample at each iteration, so cross-metric covariance is preserved; the remaining per-arm interval tables use an independent, deterministic, metric-specific seed per table. The intervals reflect image clustering but not upstream selection over candidate pipelines; secondary endpoints are exploratory unless a table explicitly reports Holm adjustment. Runtime and invalid outputs are counted rather than dropped, with row-level reports retaining paired win/loss/tie counts. Appendix 6 gives the complete metric definitions.

5 Results

Table 1 collects the deployed policy's effect ($CRG - B$) across the three evidence tiers before the per-evaluation detail below.

5.1 Retrospective full-split matched comparison

Table 1 above and Table 26 in the appendix report this population's full metric set and one paired denominator. The annotation-routed resolution diagnostic, reported separately because it uses target size, reaches 89.220% (+0.689 points); it is neither deployable nor an upper bound, and the

Dataset	Rows	Images	Acc@0.5 Δ	Acc@0.75 Δ	Acc@0.9 Δ	mAcc Δ
RefCOCO	10,752	1,500	**+0.772** [0.516, 1.037]	+0.456 [0.037, 0.881]	−0.530 [−1.301, 0.234]	**+0.343** [0.083, 0.594]
RefCOCO+	10,615	1,500	**+0.641** [0.381, 0.908]	**+0.895** [0.476, 1.319]	+0.377 [−0.360, 1.110]	**+0.644** [0.385, 0.911]
RefCOCOg	9,602	2,600	**+1.062** [0.751, 1.391]	**+1.250** [0.804, 1.705]	+0.083 [−0.587, 0.773]	**+0.984** [0.698, 1.281]

Table 2: Per-dataset deployed-policy effects, $CRG - B$ (final policy against the original pre-routing incumbent B), image-clustered 10,000-iteration bootstrap, Holm-corrected within the 12-comparison (3 datasets $\times$ 4 metrics) family; bold marks adjusted $p < 0.05$. Row-weighted deltas equal the pooled $CRG - B = (+0.817, +0.852, -0.029, +0.645)$ reported above, verified by a machine-checked weighted-identity assertion (Appendix 2.3). Image counts overlap across RefCOCO/RefCOCO+; RefCOCOg draws on a largely disjoint image set.

full deployable system now exceeds it, since crop re-grounding and fusion contribute accuracy the routing-only diagnostic cannot.

The larger differences at stricter thresholds are consistent with improved boundary placement, but do not by themselves distinguish tighter boxes from changes in target selection, annotation convention, or domain mix. The archived report records one invalid output per arm, 6,641 routed rows (20.8%), and 29,523 accepted-and-fused crop candidates of 31,920 attempts. Appendix 4.2 decomposes this into a guard $\times$ fusion factorial: the guard, not the midpoint rule, is what protects Acc@0.5.

5.2 Cross-dataset transfer to the RefCOCO family

After the primary Ref-L4 analysis was frozen, we applied the same checkpoint, prompt, parser, router, resolution floor, crop guards, and midpoint fusion to the five official RefCOCO/RefCOCO+/RefCOCOg test partitions (30,969 expressions over 3,982 images), retuning nothing. Every threshold and weight was fixed before any RefCOCO-family row was scored, so this comparison is a frozen no-retuning transfer, unlike the retrospective Ref-L4 estimate of Section 5. A later content-hash audit found that 3,828 of the RefCOCO-family's 3,982 unique source images also occur in the Ref-L4 population (40.44% of Ref-L4's unique hashes), so this is frozen transfer evidence rather than an image-disjoint confirmation; the audit does not change the fact that no RefCOCO-family labels were used to retune the operator (full counts and pairwise fractions are in Appendix 1.5).

The pooled result (Table 1) is positive on the primary endpoint and the operating threshold, flat at the strictest one: Acc@0.75 gains $+0.853$pp ([0.566, 1.147]) and mean IoU gains $+0.552$pp ([0.438, 0.668], $p < 0.001$), each with a Holm-adjusted $p \approx 0.003$ where reported. Acc@0.9 does not move: -0.029 points ([-0.554, 0.497], adjusted $p = 0.93$). The paired report records 13,925 IoU wins, 12,357 losses, and 4,687 ties, with two invalid rows in each arm. Table 2 breaks the deployed policy's own effect, $CRG - B$, down by dataset: every dataset individually improves on Acc@0.5, mAcc, and mean IoU, eight of the twelve dataset-by-metric comparisons clear Holm-adjusted significance, and Acc@0.9 is indistinguishable from zero on every dataset, matching the pooled null rather than concealing a per-dataset regression. A separate stage decomposition (Appendix 2.3) explains why the pooled Acc@0.9 null is not simply "no effect": routing alone gains $+1.162$ points at Acc@0.9, while the marginal effect of crop, guard, and fusion after routing is -1.192 points relative to the routed arm, so the two stages substantially offset. Appendix 2.3 also gives the matching failure taxonomy.

The resolution router acts on small targets, and only on them. The label-free router flagged 2,223 of the 30,969 rows (7.18%) for a higher-resolution second pass, and changed the predicted box on 2,213 of them (99.6%). Routing alone is significant on every metric: $+0.691$ points at Acc@0.5 (95% CI [0.554, 0.836]), $+0.807$ at Acc@0.75, $+1.162$ at Acc@0.9, and $+0.861$ at mAcc, all with Holm-adjusted $p \approx 0.003$, from 1,444 IoU wins against 651 losses on the 2,095 rows whose accuracy classification flips at some threshold (28,874 of the 30,969 rows tie; the remaining 118 of the 2,213 coordinate changes move the box without crossing any scoring threshold); Appendix 4.5 documents an earlier resolution-router defect and its correction.

The size decomposition shows the effect is concentrated where the router targets it: against the incumbent, routing alone moves Acc@0.9 by $+10.13$ points on the small bucket (1,647 rows, 41.96% to 52.09%), by $+0.93$ on the medium bucket, and negligibly on the large bucket (9,567 rows).

The guard is load-bearing, not a trade against a competitive alternative. Because the crop candidate is retained before the guards and fusion are applied, the stages separate on identical rows (Table 9). Crop re-grounding alone, with the guard still active, is a boundary-precision mechanism: it gains $+1.844$ points at Acc@0.9 (95% CI [1.349, 2.350]) for only $+0.236$ at Acc@0.5. Removing

System	Params.	Acc@0.5	Acc@0.75	Acc@0.9	mAcc	Mean IoU	Invalid
Frozen incumbent	8B	89.212	80.790	60.955	75.925	82.262	2
+ resolution routing	8B	89.903	81.597	**62.117**	**76.786**	82.911	2
+ LFPR	8B	90.029	81.643	60.926	76.570	82.814	2
Released specialist	4B	90.723	81.856	61.558	76.722	83.171	115
+ LFPR	4B	91.007	82.311	**63.128**	**77.508**	**83.592**	115
Released specialist	8B	91.311	82.027	53.786	74.705	82.365	0
+ LFPR	8B	**91.688**	**83.257**	60.502	77.106	83.583	0
+ crop, routing, *no guard* ($CR^{\dagger}$, negative control)	8B	86.974	77.632	56.382	72.707	79.974	2

Table 3: Zero-shot released specialists against the frozen incumbent and LFPR, each also shown with LFPR applied to its own box (every specialist-plus-LFPR pair here improves significantly; CIs in Appendix 5.1.1). LFPR improves every specialist on every metric including Mean IoU, most sharply the 8B specialist's Acc@0.9 ($+6.72$). Among the base systems, best-Acc@0.5 (8B specialist) has worst Acc@0.9/mAcc; the trade persists after refinement. The bottom row is a negative control: the guard gate removed from CRG's own stored candidates. It underperforms every system above it on every metric. Invalid predictions remain in every denominator.

the guard entirely – a genuine ablation, recomputed directly from CRG's own stored candidate and guard fields with the guard gate removed – does not compound that gain: it underperforms the incumbent on every metric, -2.926 points at Acc@0.5, -3.969 at Acc@0.75, and -5.748 at Acc@0.9 (Table 10 gives the full intervals). An archived run under the same "CR" label uses different upstream predictions and is not a guard ablation; we retain it only as an archival artifact (Table 9).

The deployed policy's pooled Acc@0.9 point estimate is slightly negative (-0.029 points, 95% CI $[-0.554, 0.497]$, adjusted $p = 0.93$) – statistically indistinguishable from zero rather than a real strict-IoU regression – in exchange for the largest Acc@0.5 gain of any arm ($+0.817$), since guarding protects rows the incumbent already had approximately right. Across the full 30,969-row evaluation, CRG yields 13,925 IoU wins, 12,357 losses, and 4,687 ties relative to the matched pre-resolution incumbent, with Mean IoU increasing from 82.262% to 82.814% ($+0.552$ points) – many small, mostly favorable boundary shifts outweighing a smaller number of larger losses, not one mechanism dominating the other (Appendices 2.3 and 4.4).

We ship one deployment policy, named explicitly: **LFPR** (guards plus midpoint fusion, arm CRG), reported in the abstract's headline numbers. Removing the guard is not a viable second policy: it is a zero-cost accept/reject decision on an already-computed crop candidate (so $CR^{\dagger}$ and CRG issue identical model calls, Table 17), yet it is not competitive on any accuracy metric (Table 9).

The guard×fusion factorial (Appendix 4.2) shows guarded replacement (an arm we did not ship) significantly beats CRG at Acc@0.9 and $\mathrm{mAcc}_{0.5:0.95}$ at the cost of Acc@0.5; we retain midpoint fusion as the policy fixed before this factorial, with the direct paired contrast in Appendix 4.2.

5.3 Comparison against released specialists: accuracy is not ordered, it is redistributed

We also evaluated two released referring-expression specialists (4B and 8B parameters) zero-shot on the same rows, clusters, and official evaluator (Zhan et al., 2026); both emit a single answer with no refinement. Since the specialists are supervised on this benchmark family while our incumbent is zero-shot, this is not a like-for-like contest but a test of whether refinement remains useful once the incumbent is already specialised. Table 3 reports the result and Figure 1 the full threshold profile.

The headline reading is unfavourable to our method: a released 4B model doing one forward pass exceeds the refined 8B pipeline by 0.694 points at Acc@0.5 (95% CI $[0.234, 1.148]$, adjusted $p = 0.036$), and the 8B specialist exceeds it by 1.282 points ($[0.883, 1.698]$). Refining a frozen generalist narrows but does not close the gap to a task-specialised model at the permissive threshold that dominates published comparisons. The strict-threshold reading reverses, but only for the 8B specialist: the guarded system is indistinguishable from the 4B specialist at Acc@0.9 (-0.633 points, 95% CI $[-1.545, 0.228]$, adjusted $p = 0.96$) and beats the 8B specialist by $+7.139$ points ($[6.169, 8.125]$) with $+1.865$ points of mAcc.

The same negative control (Table 3, bottom row) trails both specialists at Acc@0.5 by a wider margin than the guarded system and no longer beats either convincingly at Acc@0.9, confirming the 8B-specialist reversal is a property of the guarded system specifically. Scaling the specialist from

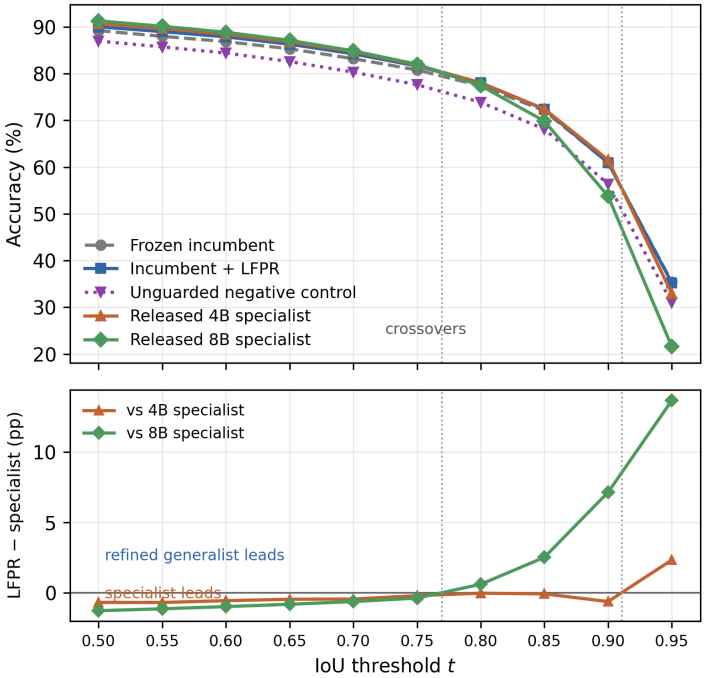


Figure 1: Accuracy against IoU threshold on the 30,969-row transfer split (top) and the paired difference between the guarded system and each specialist (bottom). Specialists lead at permissive thresholds; LFPR crosses the 8B specialist near $t \approx 0.78$ and widens thereafter, but tracks the incumbent and the 4B specialist at strict thresholds until $t = 0.95$. The unguarded negative control trails every system across the full range, reaching only 30.99% at $t = 0.95$ against 35.25% for LFPR, 35.43% for the incumbent, and 32.93%/21.60% for the specialists.

4B to 8B sharpens the same trade-off: Acc@0.5 rises $+0.588$ points while Acc@0.9 falls 7.772 and mAcc falls 2.017 (Appendix 5 gives the full row-level detail). An image-level firewall audit finds zero training-image overlap between the specialists and any evaluation partition, so their Acc@0.5 advantage is genuine rather than memorization, though benchmark-specific: the direction reverses on long expressions (Appendix 5).

Refinement composes with specialisation. Running the complete frozen procedure, unchanged, on each specialist's own predictions over the same 30,969 rows improves every reported endpoint for both (full intervals in Table 31), with a substantially larger strict-threshold gain for EGM-8B than EGM-4B here – not evidence of a general scaling law across model size.

5.4 A PROSPECTIVE FLICKR30K ENTITIES CONFIRMATION UNDER AN MDETR-STYLE MERGED-BOX PROTOCOL

A stronger test of generalisation is a dataset the method never saw during development. We use the official Flickr30K Entities test split (Plummer et al., 2015) under an MDETR-style merged-box protocol (Kamath et al., 2021): every visible phrase with at least one annotated box is represented by the enclosing box of all its annotations, one of several conventions used in prior work; Appendix 2.5 confirms the same arm ordering under any-associated-box scoring. The resulting 14,481-row, 999-image denominator (Table 1; Appendix 2.4 gives the manifest and firewall detail) improves every endpoint: Acc@0.75 rises $+1.326$ ($p = 0.0008$), and Acc@0.9 rises $+1.022$ ($p = 0.0052$), with zero invalid predictions.

The stage decomposition is less uniformly favorable than the endpoint comparison. The baseline (B) incumbent reaches mAcc 56.975 and Acc@0.9 37.753; routing alone (R) reaches mAcc 58.585 and Acc@0.9 39.749; full LFPR (CRG) reaches mAcc 57.948 and Acc@0.9 38.775. Thus routing transfers prospectively, but crop/guard/fusion does not add accuracy beyond routing on this dataset; its marginal value is dataset-dependent. The guard nevertheless remains load-bearing because the matched unguarded arm falls below the original incumbent (Appendix 2.6).

As a sensitivity analysis on the same images, we also retain the historical single-box selection (excluding 2,907 multi-box and 3,038 non-visual phrases; 11,574 rows over 979 images): it yields larger changes (mAcc $+2.575$, Acc@0.5 $+1.175$, Acc@0.75 $+3.145$, Acc@0.9 $+3.689$) but is scoring-protocol robustness on the same image population, not an independent confirmation. Both protocols were frozen before scoring, with zero image-hash overlap with historical development data.

6 Analysis and Discussion

6.1 Why the stages can oppose one another, and what a single threshold hides

The resolution stage changes the visual evidence available to the model; the crop stage changes the coordinate frame in which it localizes; their effects are separable and sometimes opposing (Table 2; Table 9). Resolution routing alone earns a clean, positive strict-IoU effect on every tier, accounting for most of the deployed policy's Acc@0.9 effect (Appendix 4.1). Crop re-grounding, the guard, and fusion protect against an unfiltered second opinion rather than guarantee further accuracy: their marginal effect beyond routing does not monotonically improve the routed incumbent. On the RefCOCO family (Appendix 2.3), they retain a small additional Acc@0.5 gain while giving back part of routing's strict-IoU and mAcc improvement; on Flickr30K (Appendix 2.6), their marginal effect is negative on every reported endpoint. Ref-L4 is the exception, but it is also the most heavily tuned tier, so only the untuned tiers speak without a tuning confound. The robust claim is narrower than "components compose": selective resolution transfers, the policy improves broadly, and the guard prevents an unfiltered crop answer from becoming competitive. CRG is the historically frozen policy evaluated across every tier, not retrospectively selected per dataset: on Flickr30K routing alone lies above it on the accuracy–cost frontier, yet we retain CRG because a fixed policy's purpose is to expose how components redistribute quality, not declare a per-benchmark winner.

The specialist comparison makes a broader methodological point: reporting Acc@0.5 alone can invert system rankings, since two systems can differ by seven points at Acc@0.9 while differing by one point the other way at Acc@0.5. We recommend reporting referring-expression results as a threshold profile, or at minimum with mean accuracy alongside Acc@0.5.

6.2 Limitations

Our evidence has five boundaries.

First, the Ref-L4 estimate is retrospective: the published test distribution shaped diagnostics and selection, so its intervals describe sampling variability under a frozen comparison, not selection across the pipelines and variants explored. The RefCOCO-family result is a frozen no-retuning transfer with substantial source-image overlap, not an image-disjoint confirmation.

Second, the specialist comparison is only partly compute-normalised (Appendix 5 gives same-host latency ratios); exact hardware FLOPs/energy were not measured, and the synchronized audit reports a parameter/token operation-count proxy, not a hardware counter (Appendix 2.9).

Third, the specialists are supervised on the RefCOCO family while our incumbent is zero-shot; a firewall audit finds no training-image overlap, so the comparison is uncontaminated but not like-for-like.

Fourth, the RefCOCO-family transfer is not content-disjoint from Ref-L4 (3,828 shared image hashes), so we reserve image-disjoint language for Flickr30K, whose 999 images had zero overlap with either audited population. Our operator-effect evidence is positive across the evaluated Qwen family and released specialists, but cross-family transfer is mixed: the complete GLM-4.1V-9B-Thinking run is negative on mAcc (Appendix 2.16), while the matched NVILA-8B pilot is inconclusive (Appendix 2.17).

Fifth, the size threshold, guards, and midpoint weight are fixed constants: a local grid sweep finds no configuration that Pareto-dominates the shipped rule (Appendix 3), and we did not learn them, since that would change the contribution from a frozen operator to a learned router.

7 Conclusion

We studied whether a frozen grounding-capable VLM can use its own incumbent prediction to obtain a more precise box without accessing target annotations at inference. LFPR's lesson: components move different, opposing parts of the curve, and the guard is load-bearing – a complement to stronger grounding models, not a replacement. The benefit is conditional on an already-roughly-correct incumbent, not a universal property of repeated visual inference.

AI USE STATEMENT

We used generative-AI tools (Claude Code, GitHub Copilot/Copilot CLI, and GitHub Copilot in VS Code) for research assistance, software development, and manuscript preparation.

Research ideation and execution. AI tools provided feedback during hypothesis refinement and methodology/experiment design; assisted with implementation, debugging, testing, data-parsing and schema-validation scripts; and helped organize qualitative failure analyses and preliminary interpretations. Authors made all final research decisions, manually verified data transformations and category counts, and audited and re-tested critical evaluation code, including metric computation, bootstrap resampling, guard logic, and dataset joins.

Theory, mathematics, and results. AI tools were not used to develop the paper's theoretical models or conceptual framework, formulate mathematical claims, provide ingredients for proofs, or write proofs. They were used only for latex formatting and notation-consistency checks. All reported numerical results were computed on real evaluation data using controlled, author-reviewed scripts; AI was not used to generate synthetic evaluation data, conduct benchmark inference, or replace official evaluation.

Writing and retrieval. AI tools assisted with drafting, translation, and editing of parts of the manuscript, and with related-work discovery and positioning checks. All AI-assisted text, claims, code, and analyses were reviewed, validated, and approved by the authors. The authors take full responsibility for the final content of this work.

Verification and reproducibility: The authors wrote and executed all experiment scripts, reviewed AI-assisted code before deployment, manually tested critical functions against hand-computed examples, and re-verified headline numbers and confidence intervals against versioned prediction artifacts. The anonymized supplementary artifacts, versioned run logs, and repository history retained for this project document the implementation provenance associated with the reported results. The authors take full responsibility for all claims, statistical conclusions, code correctness, and reported results.

ETHICS STATEMENT

This work introduces no new human-subject data collection and evaluates only on established public grounding benchmarks under their existing research-use terms. The source datasets may contain images of people and inherit the privacy, representation, and annotation limitations documented by their creators; our experiments neither infer sensitive attributes nor release new personal information. LFPR increases inference cost, which we report explicitly, but does not add identity-recognition or surveillance-specific training.

REPRODUCIBILITY STATEMENT

The complete source code implementing LFPR and the analysis scripts used to produce every table and figure in this paper are included as anonymized supplementary material submitted alongside this paper on OpenReview. All prompts, pixel budgets, clipping conventions, guard thresholds, and decoding settings used by LFPR are given in Appendix 1, together with the preprocessing constants and the environment fields that could not be recovered from archived logs. The image-cluster bootstrap procedure and, where applicable, the Holm multiplicity correction used for reported comparison families are specified in Appendix 6; tables state when multiplicity adjustment applies, and retrospective or exploratory intervals are not implicitly Holm-adjusted. The complete label-free procedure is stated as pseudocode in Appendix 4. Dataset splits and source labels are described in Section 4 and Appendix 1.4. The content-hash audit (Appendix 1.5) quantifies the substantial source-image overlap between Ref-L4 and the RefCOCO-family evaluation and separately verifies zero audited content-hash overlap between the prospective Flickr30K population and either the Ref-L4 or RefCOCO-family populations. Appendix 1.5 gives the SHA-256 checksum of the exact per-row prediction file behind every headline result (Table 3); the anonymized supplementary code includes the full-length hashes and generation commands in a machine-readable manifest.

REFERENCES

Jierun Chen, Fangyun Wei, Jinjing Zhao, Sizhe Song, Bohuai Wu, Zhuoxuan Peng, S.-H. Gary Chan, and Hongyang Zhang. Revisiting referring expression comprehension evaluation in the era of large multimodal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 513–524, 2025a.

Juntao Chen, Wen Shen, Zhihua Wei, Lijun Sun, and Hongyun Zhang. Leveraging debiased cross-modal attention maps and code-based reasoning for zero-shot referring expression comprehension. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20413–20424, 2025b.

Xiaofu Chen, Yaxin Luo, Gen Luo, Jiayi Ji, Henghui Ding, and Yiyi Zhou. DViN: Dynamic visual routing network for weakly supervised referring expression comprehension. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14347–14357, 2025c.

Ming Dai, Jian Li, Jiedong Zhuang, Xian Zhang, and Wankou Yang. Multi-task visual grounding with coarse-to-fine consistency constraints. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 2618–2626, 2025. doi: 10.1609/aaai.v39i3.32265.

Marcel Gröpl, Jaewoo Jung, Seungryong Kim, Marc Pollefeys, and Sunghwan Hong. Entropy-gradient grounding: Training-free evidence retrieval in vision-language models. *arXiv preprint arXiv:2604.08456*, 2026.

Zheng Jiang, Yiming Chen, Nan He, Jiahui Chen, Chaoyang Li, Houde Qian, and Lifeng Sun. Test-time scaling over perception: Resolving the grounding paradox in thinking with images. *arXiv preprint arXiv:2604.11025*, 2026.

Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. MDETR: Modulated detection for end-to-end multi-modal understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1780–1790, 2021.

Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, Kai-Wei Chang, and Jianfeng Gao. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10965–10975, 2022.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection. In *European Conference on Computer Vision*, 2024.

Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 11–20, 2016. doi: 10.1109/CVPR.2016.9.

Matthias Minderer, Alexey Gritsenko, Austin Stone, Maxim Neumann, Dirk Weissenborn, Alexey Dosovitskiy, Aravindh Mahendran, Anurag Arnab, Mostafa Dehghani, Zhuoran Shen, Xiao Wang, Xiaohua Zhai, Thomas Kipf, and Neil Houlsby. Scaling open-vocabulary object detection. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Yik Lung Pang and Changjae Oh. Chain-of-caption: Training-free improvement of multimodal large language model on referring expression comprehension. *arXiv preprint arXiv:2602.08211*, 2026.

Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2641–2649, 2015. doi: 10.1109/ICCV.2015.303.

Qwen Team. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Xiangyu Zhang, Jing Li, and Jian Sun. Objects365: A large-scale, high-quality dataset for object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8430–8439, 2019. doi: 10.1109/ICCV.2019.00852.

Animesh Tripathy and Aswanth Krishnan. Iterative visual thinking and the self-correction mirage in VLM grounding. *arXiv preprint arXiv:2606.13156*, 2026.

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

Guanqi Zhan, Changye Li, Zhijian Liu, Yao Lu, Yi Wu, Song Han, and Ligeng Zhu. Scaling test-time inference for visual grounding. *arXiv preprint arXiv:2601.13633*, 2026.

SUPPLEMENTARY MATERIAL

WHERE GROUNDING ACCURACY LIVES ON THE IOU CURVE: LABEL-FREE INFERENCE-TIME BOUNDARY REFINEMENT

Anonymous authors
Paper under double-blind review

This supplement gives the recoverable implementation specification, data-use audit, detailed result intervals, and disposition of the matched control experiments for LFPR. Missing provenance fields are stated explicitly rather than reconstructed from unversioned memory.

1 REPRODUCIBILITY SPECIFICATION

The complete source code implementing the LFPR pipeline – the resolution router, crop re-grounding, geometric guards, midpoint fusion, and the analysis scripts used to produce every table and figure in this paper – is included as anonymized supplementary material submitted alongside this paper on OpenReview.

1.1 MODEL AND DECODING

The generalist arms use a local copy of `Qwen/Qwen3-VL-8B-Instruct` at a single pinned revision, verified identical across every downloaded file's cache metadata (the exact revision hash is recorded in the released run manifest rather than restated here), in bfloat16, SDPA attention, evaluation mode, and deterministic, non-sampling greedy decoding, at most 64 generated tokens for the full-split accuracy runs. The synchronized cost protocol raises the ceiling to 4,096 new tokens and reports the realized generated-token count. The full-split Ref-L4, RefCOCO-family, and Flickr30K accuracy runs, and the policy-matched cost audit, used Python 3.10.12, PyTorch 2.8.0 (CUDA 12.8 build), Transformers 4.57.1, Pillow 11.1.0, NumPy 1.26.4, NVIDIA driver 595.84, Ubuntu 22.04.5 LTS, kernel 5.15.0, on a four-GPU host (NVIDIA GeForce RTX 3090, 24 GiB each). The full-parameter fine-tune control ran separately on an NVIDIA RTX PRO 6000 Blackwell Server Edition with approximately 96 GiB of memory; earlier development-phase pilots not reported as final numbers in this paper used a mix of rented cloud GPU instances, documented per-experiment in the released run logs rather than restated here.

The specialist cost audit uses local EGM-4B and EGM-8B checkpoints in bfloat16 with SDPA attention and greedy decoding, allowing at most 4,096 new tokens. It ran on a host with four physical NVIDIA RTX 3090 GPUs (24 GiB each). Each baseline/refinement arm was divided into two non-overlapping row shards, with one process per GPU; no GPU hosted more than one worker.

1.2 DETERMINISM CHECK

Since LFPR uses frozen weights and greedy decoding, training-seed replication is not meaningful; instead we verify output-level determinism directly. A fixed 128-row deterministic slice of the direct-answer incumbent (arm B) was run in three independent processes launched in parallel on three separate physical GPUs on the same host (rather than three sequential repeats on one card, which additionally rules out cross-device, not just cross-process, non-determinism). All three runs' predicted boxes were byte-identical on $128/128$ rows: a 100% reproducibility rate under greedy decoding on this hardware.

The full-image prompt is, verbatim:

> *Locate the region that matches: <expression>. Reply only with its bounding box as [x1,y1,x2,y2] in coordinates normalized to 0–1000. No explanation.*

The crop-pass prompt replaces the final two sentences:

Reply only with its bounding box as [x1,y1,x2,y2] in coordinates normalized to 0–1000 within this crop.

In both cases the checkpoint's native chat template is applied with an image followed by the user text, with the chat template's generation-prompt formatting enabled; no system message or stop string is added.

The default first-pass image-processing configuration is loaded from the checkpoint. The routed pass overrides the minimum-pixel budget to 4,194,304 and leaves the checkpoint's maximum unchanged. The crop pass reuses the same 4,194,304-pixel minimum override and also leaves the checkpoint maximum unchanged, which resolves to an effective processed-image size bounded between 4,194,304 and 16,777,216 pixels for both passes; this was confirmed by reading the image processor's own logged effective configuration from the run that produced every crop-based result in this paper, not from the smaller, crop-specific pixel range (approximately 200,704 to 1,003,520 pixels) that the runner accepts as an argument but never actually applies. Source images are opened with PIL, honor PIL's decoded orientation as stored, and are converted to RGB; explicit EXIF transposition was not applied. Crops use PIL's default resampling path through the Qwen processor. No weights are updated or adapter merged.

The parser strips the completion, accepts exactly one JSON/bare four-number box (a singleton wrapper is permitted), and rejects zero or multiple boxes. Coordinates must satisfy $0 \le x_1 < x_2 \le 1000$ and $0 \le y_1 < y_2 \le 1000$ before division by 1000. Invalid, non-finite, zero-area, or inverted boxes score as misses; no row is dropped. Crop pixel bounds use floor/ceiling, are clipped to the source image, and the accepted candidate is mapped back without integer rounding.

1.3 Router and crop specification

The normalized incumbent box is converted to original-image pixel coordinates. If $s = \sqrt{wh} < 128$ pixels, the row is routed to the high-resolution pass; $128 \le s \le 256$ is medium and is not routed; all larger boxes are not routed. Missing dimensions, malformed boxes, and invalid first-pass outputs are never converted into a synthetic small box. They remain baseline misses and are recorded in the error counters.

For every valid post-resolution incumbent, the crop width and height are each twice the incumbent width and height, capped at one and shifted to remain inside the image. Zoom consistency is the Euclidean distance between the crop candidate center and the incumbent center expressed in crop coordinates, divided by $\sqrt{1/2}$. The candidate is admitted only when this value is at most 0.35, candidate–incumbent IoU is at least 0.25, and the area ratio lies in $[0.5, 2.0]$. The output is the coordinate-wise midpoint. Invalid incumbents short-circuit as misses; invalid high-resolution or crop answers, zero-area denominators, degenerate crops, and runtime errors never trigger fusion. Available history says these constants preceded the full-split run, but not that they preceded all access to the published test distribution.

1.4 Data firewall

The original "test firewall" claim is withdrawn. The 2,000-row pilot is a deterministic image-disjoint holdout relative to named earlier pilots, but it was sampled from the already-scored Ref-L4 test prediction file and reused by the crop/fusion-only and full LFPR arms. The complete published test split was also used for annotation-size and source diagnostics that influenced later choices. Thus the holdout is a useful same-row engineering check, not an untouched confirmation benchmark.

1.5 Content-hash overlap audit

To separate filename conventions from actual image identity, we performed a CPU-only streaming SHA-256 audit over unique source-image manifests derived from the frozen Ref-L4, RefCOCO-family, and Flickr30K merged-box prediction artifacts. The audit resolved every one of the 14,448 manifest rows to a readable image file: 9,467 Ref-L4 images, 3,982 RefCOCO-family images, and 999 Flickr images. Ref-L4 contains one within-population duplicate content hash, so it has 9,466 unique hashes despite 9,467 image IDs. No declared-only or missing-file fallback was used.

Artifact	Rows	Images	Label access	First recoverable access	Recorded uses and decisions
Ref-L4 published test	31,921	9,467	Yes	2026-08-05 or earlier	Baseline scoring; source/size diagnostics; resolution arm; final LFPR scoring and manuscript framing.
Frozen RefCOCO-family transfer	30,969	3,982	Yes	2026-08-10	Frozen LFPR pooled audit; descriptive dataset slices; no target-label tuning.
Shared crop/refinement holdout	2,000	1,713	Labels present in rows	2026-08-08	Selected from the full-test baseline after excluding named prior-pilot images; reused for crop/fusion-only and full LFPR decisions.
Resolution/prompt pilots	800 each	Not fully exported	Yes	2026-08-07–08	Floor, prompt, and stopping choices; some samples were reused and the rung summaries used different references.
Supervised-adaptation split	3,314	Not fully exported	Yes	2026-08-08	Supervised-adaptation control and stopping decision; relation to all other image pools is not reconstructable from the compact export.
Backbone-scale pilot	380	344	Yes	2026-08-09	Same-family 32B direct-model control; no complete LFPR run.

Table 1: Minimum data-governance record recoverable from compact manifests and logs. "Not fully exported" is a provenance gap, not evidence of separation.

Population pair	Shared hashes	Fraction of first	Fraction of second
Ref-L4 / RefCOCO family	3,828	40.44%	96.13%
Ref-L4 / Flickr30K	0	0.00%	0.00%
RefCOCO family / Flickr30K	0	0.00%	0.00%

Table 2: Content-hash overlap over unique source-image manifests. Hashes were computed from file bytes with streaming SHA-256; paths and image IDs were never used as overlap keys. The full conflict list, manifest hashes, and file-resolution diagnostics are retained in the machine-readable audit artifact.

The result changes the evidence qualification in an important but narrow way. RefCOCO-family is not image-disjoint from Ref-L4: 3,828 unique content hashes are shared, corresponding to 40.44% of Ref-L4's unique hashes and 96.13% of the RefCOCO-family's unique hashes. We therefore describe RefCOCO-family as frozen transfer with no label-based retuning, not as an independent image-disjoint confirmation. Flickr30K remains the prospective image-disjoint confirmation relative to both audited populations: both pairwise intersections are exactly zero.

Reproducible future evaluations require stable record keys, content hashes rather than filenames, artifact checksums, and an access ledger with actor, purpose, split, label availability, and decision influence. Until that ledger and an untouched benchmark/model combination are available, intervals for the Ref-L4 retrospective analysis remain retrospective; the frozen-transfer intervals reported below are identified separately by their evidence tier.

Artifact hash manifest. Table 3 lists the checksum of the exact per-row prediction file behind each of this paper's headline evaluation results, spanning the retrospective Ref-L4 analysis, frozen RefCOCO-family transfer, and prospective Flickr30K confirmation (including its 512-row matched

Dataset	Arm	Rows	SHA-256 (12 hex)
Ref-L4	B	31,921	a31497f9f3cb
Ref-L4	R (diagnostic)	31,921	c4d83a2c7800
Ref-L4	CRG	31,921	3bee48fa7c3f
RefCOCO family	B	30,969	0b0e6a9b508d
RefCOCO family	CRG	30,969	fe4508dff83b
Flickr merged-box	B	14,481	24170b518493
Flickr merged-box	R	14,481	1044ef57c8b0
Flickr merged-box	crop-on-R (raw)	14,481	ed88edbf7f8c
Flickr merged-box	CRG	14,481	f3be68709a53
Flickr merged-box, 512-row matched	B	512	cc29a5e2f446
Flickr merged-box, 512-row matched	crop-on-R (raw)	512	129cd8e0cb17
Flickr merged-box, 512-row matched	R_{ALL}	512	8ccaa327e65f
Flickr merged-box, 512-row matched	CR_{ALL}	512	e536db5e0514
Flickr merged-box, full population	R_{ALL}	14,481	41df2c56899c
Flickr merged-box, full population	CR_{ALL}	14,481	1db4562f197e

Table 3: Headline artifact hash manifest. Each row is one frozen per-row prediction file (record_key, gt_box, pred_box, and arm-specific fields); every headline table or figure that reports the corresponding arm's accuracy or cost derives from exactly this file, not a superseded or differently-scored copy of it.

subset). Every downstream table that reports one of these arms' numbers is traceable to exactly one of these hashes – in particular, the RefCOCO family's B row is the hash-verified pre-resolution artifact used throughout the reported analysis; differently named alternative execution artifacts are excluded from all headline results. Hashes are truncated to 12 hex characters here for space; the full 64-character digests, the exact generation commands, and the artifacts still outside this manifest's coverage (the RefCOCO-family cost audit and several secondary diagnostic tables) are listed in the code supplement's machine-readable manifest referenced in the Reproducibility Statement.

2 Detailed Results

Arm glossary. Throughout this appendix, B denotes the frozen incumbent's pre-resolution prediction; R denotes the prediction after label-free resolution routing alone; C denotes a guarded crop pass without routing; $CR^{\dagger}$ denotes the unguarded negative control (routing plus crop with the geometric guard removed); and CRG denotes the full shipped policy (routing, guarded crop, and midpoint fusion). A contrast written $X{-}Y$ always means arm X minus arm Y on the same rows; tables and prose state their reference arm explicitly because $CRG{-}B$ (the deployed policy's effect) and $CRG{-}R$ (the post-routing stage effect) answer different questions and are not interchangeable, as Table 7 and its surrounding discussion make explicit.

2.1 Evidence-tier summary

Table 4 collects the treatment-minus-incumbent effect on every primary and secondary endpoint across the paper's three evidence tiers, before the per-evaluation breakdowns below. Effects are differences against each evaluation's own matched incumbent; the Ref-L4 and RefCOCO-family rows draw from Tables 26 and 6, and the Flickr rows from Section 5.4 of the main text and Appendix 2.4. Mean-IoU deltas are in percentage points and are given only where an archived report computed them; a blank cell means the quantity was not computed, not that it was zero.

2.2 Pilot and full-split intervals

2.3 Frozen transfer to the RefCOCO family

The transfer evaluation applies the frozen procedure without retuning to five official RefCOCO/RefCOCO+/RefCOCOg test partitions. The pooled paired audit uses 30,969 expressions, 3,982 image clusters, exact record-key joins, and 10,000 image-cluster bootstrap resamples. These partitions influenced no design decision: every threshold, guard, resolution floor, and fusion weight

Evaluation	Evidence tier	n (expr / img)	ΔAcc@.5	ΔAcc@.75	ΔAcc@.9	ΔmAcc	ΔMean IoU
Ref-L4 (full split)	Retrospective development	31,921 / 9,467	+1.194	+3.173	+5.354	+3.066	+1.704
RefCOCO/+/g (pooled)	Frozen transfer (content overlap)	30,969 / 3,982	+0.817	+0.852	−0.029	+0.645	+0.552
Flickr30K (merged-box)	Prospective, image-disjoint	14,481 / 999	+0.815	+1.326	+1.022	+0.973	+0.712
Flickr30K (single-box)	Sensitivity analysis (same prospective population)	11,574 / 979	+1.175	+3.145	+3.689	+2.575	–

Table 4: Evidence-tier summary of the three headline evaluations, in percentage points against each row's own matched incumbent. Ref-L4 is retrospective development evidence and should not be read as confirmatory; the RefCOCO-family pooled row is frozen transfer that retuned nothing but is not image-disjoint from Ref-L4 (Appendix 1.5); Table 7 decomposes this same $CRG - B$ effect by dataset. The merged-box row is the paper's prospective, image-disjoint confirmation; the single-box row reuses the same 999/979 source images under a different phrase-to-box protocol (Appendix 2.4) and therefore provides protocol-sensitivity evidence on the same population, not a second independent confirmation.

Arm	n	Acc@0.5	Acc@0.75	Acc@0.9	mAcc	Objects365 Δ
Crop/fusion reused pilot	2,000	+0.70	–	–	–	–
LFPR reused pilot	2,000	+0.85	+2.65	+2.65	+2.055	+1.30
LFPR full test	31,921	+1.194	+3.173	+5.354	+3.066	+1.720

Table 5: Historical treatment-minus-baseline effects in percentage points. The two pilot rows are reused and are not independent confirmation. LFPR pilot one-sided LCBs are +0.45, +1.85, +1.61, +1.69, and +0.68 for Acc@0.5, Acc@0.75, Acc@0.9, mAcc, and Objects365 Acc@0.5, respectively. The full test Acc@0.5 paired image-level 95% interval is [+0.944,+1.439]. LCBs explain historical selection and are not confirmatory.

was fixed on Ref-L4 before any row here was scored. The content-hash audit in Appendix 1.5 nevertheless finds 3,828 shared source images with the Ref-L4 population. We therefore retain the no-retuning and frozen-transfer qualification, but do not call this tier image-disjoint.

The paired report records 13,925 IoU wins, 12,357 losses, and 4,687 ties. Table 7 decomposes this pooled full-policy result, $CRG - B$ (the deployed policy against the original pre-routing incumbent B), by dataset, using the same exact-key join and image-cluster bootstrap as the pooled table, verified by a machine-checked assertion to reproduce the pooled estimate under row weighting to within 10^{-10} (Appendix 6).

The deployed policy improves every dataset individually at Acc@0.5, mAcc, and mean IoU (not shown in the table; $+0.323$/$+0.518$/$+0.847$ points for RefCOCO/RefCOCO+/RefCOCOg respectively, each $p < 0.001$), and at Acc@0.75 for RefCOCO+ and RefCOCOg; eight of the twelve dataset-by-metric comparisons in the table clear Holm-adjusted significance. Acc@0.9 is indistinguishable from zero on every dataset (RefCOCO -0.530, RefCOCO+ $+0.377$, RefCOCOg $+0.083$, none significant), matching the pooled null rather than concealing a per-dataset regression – no dataset shows a significant $CRG - B$ regression on any reported metric.

A separate, secondary decomposition asks a different question: relative to the *post-routing* incumbent R rather than B, does crop/guard/fusion preserve the strict-IoU gain that routing alone earns? Reusing the same per-row artifact, $CRG - R$ (not shown as a full table) is -1.423 ($[-2.206, -0.673]$, Holm-significant), -0.537 ($[-1.272, 0.171]$, not significant), and -1.656 ($[-2.288, -1.019]$, Holm-significant) at Acc@0.9 for RefCOCO/RefCOCO+/RefCOCOg, row-weighting to the pooled $CRG - R = (+0.126, +0.046, -1.192, -0.216)$ for Acc@0.5/0.75/0.9/mAcc (Table 9). This explains the mechanism behind the pooled $CRG - B$ Acc@0.9 null in Table 6 – routing gains strict-IoU accuracy that the later crop/guard/fusion stage partly returns – without changing the per-dataset $CRG - B$ conclusion above: $CRG - B$ and $CRG - R$ answer different questions (deployed-policy effect vs. post-routing marginal effect) and are not interchangeable.

Metric	Incumbent	LFPR	Δ (pp)	Two-sided 95% CI	Holm p
Acc@0.5	89.212	90.029	+0.817	[0.639, 1.001]	0.0032
Acc@0.75	80.790	81.643	+0.852	[0.566, 1.147]	0.0032
Acc@0.9	60.955	60.925	−0.029	[−0.554, 0.497]	0.9311
mAcc@0.5:0.95	75.925	76.570	+0.645	[0.469, 0.827]	0.0032
Mean IoU	82.262	82.814	+0.552	[0.438, 0.668]	—

Table 6: Pooled transfer audit on the RefCOCO/RefCOCO+/RefCOCOg family, after the resolution floor was verified to take effect. Intervals are image-cluster bootstrap intervals from the archived paired report; the Holm value is the multiplicity-adjusted primary-family value. Acc@0.9 is the one metric that does not move: the guards and midpoint return the strict-IoU accuracy that the crop and floor earn (Table 9). Mean IoU is outside the reported four-metric Holm family (no Holm-adjusted p reported here, though its own bootstrap $p < 0.001$) and moves in the *same* direction as Acc@0.5/mAcc, unlike the strict-IoU metric – consistent with the deployed policy broadly improving boundary precision even though its strict-threshold effect nets to zero after the crop stage gives back part of routing's Acc@0.9 gain (Table 9). Each arm contains two invalid rows, which remain in the denominator.

Dataset	Rows	Images	Acc@0.5 Δ	Acc@0.75 Δ	Acc@0.9 Δ	mAcc Δ
RefCOCO	10,752	1,500	**+0.772** [0.516, 1.037]	+0.456 [0.037, 0.881]	−0.530 [−1.301, 0.234]	**+0.343** [0.083, 0.594]
RefCOCO+	10,615	1,500	**+0.641** [0.381, 0.908]	**+0.895** [0.476, 1.319]	+0.377 [−0.360, 1.110]	**+0.644** [0.385, 0.911]
RefCOCOg	9,602	2,600	**+1.062** [0.751, 1.391]	**+1.250** [0.804, 1.705]	+0.083 [−0.587, 0.773]	**+0.984** [0.698, 1.281]

Table 7: Per-dataset deployed-policy effects, $CRG - B$, image-clustered 10,000-iteration bootstrap, Holm-corrected within the 12-comparison (3 datasets × 4 metrics) family. Bold marks Holm-adjusted $p < 0.05$. Row-weighted deltas equal the pooled $CRG - B = (+0.817, +0.852, -0.029, +0.645)$ in Table 6 to within 10^{-10}. Image counts overlap across RefCOCO/RefCOCO+ (both built on the same COCO subset with different expression styles); RefCOCOg draws on a largely disjoint image set, which is why the three counts do not sum to the pooled 3,982.

What kind of harm this is. We classified every harmed row on the RefCOCO-family transfer split (final box lower IoU-vs-ground-truth than the true pre-resolution incumbent B; 12,357 of 30,969 rows, matching the paired loss count above) into one of six geometric categories – router-caused, crop-boundary truncation, referent switch (final box barely overlaps its own pre-crop anchor), over-shrink, over-expand, or a residual modest-perturbation category – checked in that priority order per row, using each row's already-recorded box geometry (no new inference). The result is dominated by, but not exclusively, modest perturbation: 94.58% of harmed rows ($11{,}687/12{,}357$) are boundary shifts that still overlap their pre-crop anchor substantially, 0.78% are over-expansions, and 0.11% are over-shrinkages, with zero crop-boundary truncations or referent switches. A non-trivial remainder, 4.53% ($560/12{,}357$), is router-caused: the resolution router changed the incumbent's box and that change alone, before the crop stage ever ran, was already worse than the original pre-routing prediction. The Acc@0.9 null documented above is therefore mostly the aggregate of many small coordinate perturbations from midpoint fusion crossing the strict threshold (the same mechanism Table 28 isolates as the fusion rule's own cost), plus a smaller contribution from the router itself occasionally routing a row to a worse prediction – not evidence of crop-boundary truncation or referent switching.

The pooled read above could in principle hide a dataset-specific failure mode that a pooled classification averages away. Table 8 repeats both the taxonomy and the strict-IoU transition-matrix cells on each of RefCOCO/RefCOCO+/RefCOCOg individually rather than pooled. The boundary-shift majority and near-zero crop-truncation/referent-switch rates hold on every sub-dataset, but the router-caused share varies more than the pooled figure suggests – from 3.1% on RefCOCO+ to 7.4% on RefCOCOg – and strict-IoU harm rates cluster in a narrower 5.7–6.5% range. The pooled numbers are therefore a fair summary of the dominant boundary-shift mechanism, but they average over real dataset-level variation in how often the router itself is the source of harm.

Dataset	Boundary shift	Router-caused	Over-expand	Over-shrink	Strict harm rate†
RefCOCO	95.76%	3.41%	0.72%	0.11%	6.48%
RefCOCO+	95.96%	3.10%	0.80%	0.15%	5.70%
RefCOCOg	91.66%	7.44%	0.82%	0.08%	5.71%

Table 8: Failure taxonomy and strict-IoU harm rate computed independently on each RefCOCO-family sub-dataset (not pooled), against the true pre-resolution incumbent B. Crop-truncation and referent-switch categories are each 0.00% on all three and omitted from the columns. †Fraction of $[.9, 1]$-baseline rows that fall to $[.75, .9)$ after LFPR (the same transition-matrix cell Figure 7 plots for Ref-L4/Flickr), computed per dataset. Zero new inference; reuses the same per-row predictions as Table 7.

Arm	Acc@0.5	Acc@0.75	Acc@0.9	mAcc	IoU W/L/T vs incumbent
B incumbent	89.212	80.790	60.955	75.925	—
R routing	89.903	81.597	62.117	76.786	1,444 / 651 / 28,874
C crop	89.448	81.462	62.799	76.723	15,725 / 10,608 / 4,636
$CR^{\dagger}$ unguarded negative control	86.974	77.632	56.382	72.707	11,683 / 16,010 / 3,276
CRG full (LFPR)	**90.029**	81.643	**60.925**	**76.570**	13,925 / 12,357 / 4,687

Table 9: Five-arm same-row decomposition on the 30,969-row transfer split, all against the same incumbent with 10,000-iteration image-cluster bootstrap and Holm correction. All reported arm-versus-incumbent contrasts are significant at Holm-adjusted $p \approx 0.003$ or better except the full CRG system at Acc@0.9, whose effect is statistically indistinguishable from zero ($p = 0.93$). Routing changes few rows (2,095 of 30,969) but nearly always improves them. †: CR **here is a genuine unguarded negative control**, recomputed directly from CRG's own stored per-row fields (incumbent box, crop candidate, guard decision) with the guard gate removed and every row unconditionally replaced – it uses the exact same crop candidates CRG uses, so the comparison is a true ablation of the guard alone. It underperforms the incumbent on every metric, confirming the guard is load-bearing rather than a trade against a competitive alternative. An archived, independently-generated CR run (Acc@0.5/0.75/0.9/mAcc $= 89.919/82.311/64.604/77.727$) exists under the same label; that artifact uses an identical guard/fallback/fusion mechanism to CRG (verified: its per-row admission decision reproduces CRG's frozen guard thresholds on 30,969/30,969 rows, 100.00% exact) but different upstream baseline and crop-candidate predictions from a separate execution, so it was never a guard ablation at all. That artifact is retained only as provenance of the mislabeling, not as a result.

2.3.1 Component decomposition on identical rows

Because the crop candidate is retained before the guards and the midpoint are applied, the crop, guard, and fusion stages separate on the same 30,969 rows without any additional inference. Table 10 reports the paired contrasts.

A processor audit identified a prior execution in which the requested minimum-pixel floor was silently discarded, causing the routed prediction to reproduce the default-resolution prediction byte-for-byte. That execution is excluded from the reported routing result. With the minimum-pixel floor verified in effect, the router flags 2,223 of 30,969 rows and changes the predicted box on 2,213 of them.

2.3.2 Released-specialist comparison

Two released referring-expression specialists were evaluated zero-shot on the identical rows, clusters, and evaluator. Paired contrasts against the frozen incumbent are $+1.511$ points at Acc@0.5 for the 4B specialist ($[1.029, 2.002]$) and $+2.099$ for the 8B specialist ($[1.663, 2.508]$); at Acc@0.9 the same contrasts are $+0.604$ ($[-0.317, 1.538]$, not distinguishable from zero) and -7.168 ($[-8.178, -6.185]$). In mean accuracy the 4B specialist gains $+0.798$ points and the 8B specialist loses 1.220. Paired IoU outcomes against the incumbent are 15,318 wins / 15,311 losses / 340 ties for the 4B specialist and 11,428 wins / 19,195 losses / 346 ties for the 8B specialist.

Contrast and metric	Δ (pp)	95% CI
Resolution routing minus incumbent		
Acc@0.5	$+0.691$	$[0.554, 0.836]$
Acc@0.75	$+0.807$	$[0.637, 0.984]$
Acc@0.9	$+1.162$	$[0.927, 1.402]$
Crop alone (guarded) minus incumbent		
Acc@0.5	$+0.236$	$[0.125, 0.356]$
Acc@0.75	$+0.672$	$[0.402, 0.940]$
Acc@0.9	$+1.844$	$[1.349, 2.350]$
Unguarded crop+routing minus incumbent (negative control)		
Acc@0.5	-2.926	$[-3.264, -2.587]$
Acc@0.75	-3.969	$[-4.461, -3.477]$
Acc@0.9	-5.748	$[-6.494, -5.002]$
Full system (LFPR) minus incumbent		
Acc@0.5	$+0.817$	$[0.639, 1.001]$
Acc@0.75	$+0.852$	$[0.566, 1.147]$
Acc@0.9	-0.029	$[-0.554, 0.497]$

Table 10: Same-row component decomposition, 10,000 image-cluster bootstrap resamples, every contrast against the same incumbent. "Crop alone" keeps the frozen guard active (guard and routing are independent axes; this row isolates the crop-only pipeline path, not a guard ablation). The "unguarded crop+routing" row is a true guard ablation, recomputed directly from CRG's own stored candidate and guard fields with the guard gate removed: it underperforms the incumbent on every metric, not merely on Acc@0.5, so cropping does not buy strict-IoU accuracy for free once the guard stops filtering bad candidates. The guards and midpoint in the full system then return a small further Acc@0.9 concession in exchange for the largest Acc@0.5 gain. All contrasts are significant at Holm-adjusted $p \approx 0.003$ or better except the full system at Acc@0.9 ($p = 0.93$).

The deployed guarded arm (LFPR/CRG) trails the 4B specialist by -0.694 points at Acc@0.5 ($[-1.148, -0.234]$) and is statistically indistinguishable from it at Acc@0.9 (-0.633, $[-1.545, 0.228]$); against the 8B specialist it trails by -1.282 at Acc@0.5 ($[-1.698, -0.883]$) and leads by $+7.139$ at Acc@0.9 ($[6.169, 8.125]$). Scaling the specialist from 4B to 8B gives $+0.588$ points at Acc@0.5 ($[0.171, 1.011]$), -7.772 at Acc@0.9 ($[-8.816, -6.730]$), and -2.017 in mean accuracy.

As a negative control, the same comparison was run against the genuine unguarded arm from Table 9 (guard gate removed, same underlying candidates as CRG). It trails both specialists at Acc@0.5 by a wider margin than the guarded arm (-3.749 against the 4B specialist, -4.337 against the 8B; point estimates, since the specialists were not independently re-scored against this recomputed arm) and no longer leads either specialist convincingly at Acc@0.9: -5.176 against the 4B specialist and $+2.596$ against the 8B, versus the guarded arm's -0.633 and $+7.139$ respectively. Removing the guard therefore does not trade Acc@0.5 for a stronger strict-IoU position relative to the specialists; it loses ground at Acc@0.5 without gaining a comparably strong position at Acc@0.9.

Two controls bound this comparison. An image-level firewall audit compared 17,978 unique training images from the specialists' published training manifest against the 3,982 unique evaluation images, partition by partition, and found zero overlap in all five partitions. This rules out exact-image memorization only; it does not rule out semantic overlap, near-duplicate images, shared class priors, or shared annotation style between the specialists' training data and our evaluation split, any of which could still inflate a specialist's advantage over a model that never trained on this distribution at all. We narrow the claim accordingly: the firewall supports a claim of no exact-image leakage, not a claim that the specialist comparison isolates model quality from training-distribution similarity. Separately, because midpoint fusion produces continuous coordinates while all three single-pass systems emit coordinates on a 0.001 grid, we requantized our predictions to that grid and rescored: every reported metric moved by less than 0.03 points. This check was run on a superseded prediction artifact and was not repeated on the final frozen predictions, so we treat it only as historical evidence that coordinate granularity is unlikely to explain the observed differences, rather than as a formal bound on the final results.

2.4 FLICKR30K ENTITIES STANDARD-PROTOCOL (MERGED-BOX) CONFIRMATION

The single-box subset used in the main text excludes every multi-box phrase, so it is not comparable to standard Flickr30K Entities phrase-grounding numbers, which merge a phrase's boxes into one enclosing target rather than discarding it. We therefore ran a second, separately-frozen confirmation that keeps every visible entity phrase with at least one annotated box – 14,481 rows over 999 images (11,574 single-box plus 2,907 multi-box phrases, each represented by the deterministic enclosing $xyxy$ box of its annotated boxes) – with its own denominator, image-hash manifest, and access ledger kept separate from the single-box subset so the two protocols cannot be confused. An image-level firewall audit against the historical Ref-L4 test and validation splits found zero overlapping images and zero overlapping record keys.

Full intervals: Acc@0.5 rises $+0.815$ points (95% CI $[0.299, 1.324]$, Holm-adjusted $p = 0.0052$), Acc@0.75 rises $+1.326$ ($[0.714, 1.933]$, $p = 0.0008$), Acc@0.9 rises $+1.022$ ($[0.322, 1.705]$, $p = 0.0052$), and mAcc rises $+0.973$ ($[0.648, 1.310]$, $p = 0.0008$), with zero invalid predictions in either arm. The gain pattern differs from the single-box result: there Acc@0.75 and Acc@0.9 lead Acc@0.5; here all four primary metrics move together by roughly the same amount. Taken together with the single-box result, LFPR improves both the historical single-box selection and the standard merged-box protocol on the same held-out test images, so the single-box result is not an artifact of excluding multi-box phrases.

2.5 ANY-ASSOCIATED-BOX ROBUSTNESS CHECK

The merged-box protocol above represents every multi-box phrase by the enclosing rectangle of its annotated regions, which is one of several conventions used in prior work and is not geometrically neutral: a boundary-refinement operator's effect on IoU could in principle differ between scoring against one enclosing box and scoring against the closest of several disjoint original regions. We therefore re-scored the same archived $B/R/CRG$ predictions – no new model inference – against each phrase's original, pre-merge region set, taking the maximum IoU to any associated region rather than IoU to the merged enclosing box. The original per-phrase regions were recovered deterministically by re-parsing the same Flickr30K Entities source XML already used to build the merged-box manifest, gated on an exact SHA-256 match against the hash recorded in that manifest for every row. Before reporting any any-box number, we additionally verified that these three prediction files reproduce the merged-box $B/R/CRG$ Acc@0.5/0.75/0.9/mAcc figures already published above and in Table 11 to within 0.01 percentage points under merged-box rescoring; both identity gates passed on all 14,481 rows.

Under any-box scoring, both R and CRG again improve on B with the paired 95% CI lower bound above zero on every primary endpoint (R: mAcc 54.719 vs. B's 53.201, $\Delta + 1.518$, CI $[1.189, 1.846]$; Acc@0.9 35.861 vs. 33.858, $\Delta + 2.003$, CI $[1.423, 2.582]$. CRG: mAcc 54.134, $\Delta + 0.934$, CI $[0.609, 1.263]$; Acc@0.9 34.915, $\Delta + 1.057$, CI $[0.384, 1.734]$). The ordering between the two treatment arms is also unchanged: R remains above CRG on every endpoint under any-box scoring, the same direction reported for merged-box scoring in §5.4 and §2.6 below. The merged-box gain is therefore not an artifact of the enclosing-box geometric convention: it survives, with the same qualitative arm ordering, under an alternative scoring rule that does not require multiple regions to be collapsed into one box. (10,000-iteration image-cluster bootstrap, same protocol as every other paired contrast in this paper.)

2.6 MECHANISM REPLICATION ON THE FLICKR30K MERGED-BOX CONFIRMATION

The transfer-split same-row decomposition (Table 9, Table 10) is the paper's most detailed mechanistic evidence for the guard and router on frozen transfer data, but it shares the RefCOCO-family evaluation used above. We therefore repeated the identical $B/R/C/CR^{\dagger}/CRG$ decomposition on the Flickr30K merged-box confirmation above, which never informed method development, reconstructing every arm from the frozen run's own archived per-row fields with no additional model inference.

The two unconditional resolution controls were also scored on the same full 14,481-row population, rather than only on the 512-row matched-cost subset. Their absolute scores and paired mean-IoU changes against the same pre-resolution incumbent are shown in Table 12.

Arm	Acc@0.5	Acc@0.75	Acc@0.9	mAcc	Mean IoU Δ vs B
B incumbent (pre-resolution)	74.864	59.720	37.753	56.975	—
R routing alone	75.893	61.784	39.749	58.585	+1.013
C guarded crop alone	74.926	59.050	37.200	56.463	-0.111
$CR^{\dagger}$ unguarded negative control	74.353	58.484	36.662	55.952	-0.652
CRG full (LFPR)	**75.678**	**61.046**	**38.775**	**57.948**	**+0.712**

Table 11: Same-row $B/R/C/CR^{\dagger}/CRG$ decomposition on the Flickr30K merged-box confirmation (14,481 rows, 999 images), reconstructed from the frozen run's own archived per-row fields with no new model inference. Deltas are against B, the pre-resolution incumbent, matching Table 9's convention. 10,000-iteration image-cluster bootstrap.

Arm	Acc@0.5	Acc@0.75	Acc@0.9	mAcc	Mean IoU Δ vs B
B incumbent (pre-resolution)	74.864	59.720	37.753	56.975	—
R_{ALL} unconditional floor	74.629	61.398	42.469	58.747	+0.692
CR_{ALL} unconditional floor+crop	74.774	61.149	41.551	58.461	+0.591

Table 12: Full-population unconditional-resolution controls on the Flickr30K merged-box confirmation (14,481 rows, 999 images). Mean-IoU deltas are paired against the pre-resolution incumbent B; no new inference was performed for the B row.

The pattern matches the Ref-L4 and RefCOCO-family decompositions: CRG beats B significantly at both the permissive and strict threshold (95% CI $[+0.293, +1.338]$ points at Acc@0.5, $[+0.319, +1.698]$ at Acc@0.9), and the guard remains load-bearing – $CR^{\dagger}$ is not merely worse than CRG, it is significantly *below* B at Acc@0.9 (CI $[-1.884, -0.299]$), the same asymmetry reported for Ref-L4 and the RefCOCO family. The router's own isolated contribution is positive and significant at every threshold (Acc@0.5 CI $[+0.526, +1.536]$, Acc@0.9 CI $[+1.417, +2.575]$) and larger in magnitude than on the RefCOCO family; this is not a new mechanism, it tracks a higher routing rate on this population – the router changes the predicted box on 4,318 of 14,481 rows (29.8%) here, against 2,213 of 30,969 (7.1%) on the RefCOCO family (Section 5.2), consistent with Flickr30K Entities' generally smaller annotated targets.

The same guard-by-fusion 2×2 factorial reported for Ref-L4 (Table 28) is computable on this dataset with the identical zero-new-inference method: the incumbent box, crop candidate, and guard decision are already stored per row.

A dense-grid extension of this comparison (46 thresholds, $t = 0.50, 0.51, \ldots, 0.95$, one shared cluster-bootstrap resample per iteration across the whole grid, and a simultaneous max-$|T|$ band rather than a pointwise one) finds the CRG-versus-B gain simultaneously significant on $t \in [0.50, 0.55] \cup [0.62, 0.76] \cup [0.78, 0.82]$ plus two isolated grid points at 0.84 and 0.89: the effect is not concentrated at one threshold, it holds across most of the permissive-to-strict range on this prospective evidence. Figure 2 plots the full curve. Appendix 2.8 extends this same dense-grid construction to all five evaluations using each evaluation's own pre-resolution and fused per-row prediction files.

The same dense-grid, simultaneous-band construction, applied to the R-versus-B and CRG-versus-R pairings on this same 14,481-row population, separates the routing stage's own effect from the marginal effect of crop, guard, and fusion at every threshold rather than only the four scalar endpoints already reported in §5.4. Routing alone (R minus B) is simultaneously significant and positive on $t \in [0.50, 0.93]$ plus the isolated grid point 0.95 – effectively the entire evaluated range – with representative deltas of $+1.029$ points at $t = 0.50$ (simultaneous CI $[0.262, 1.795]$), $+2.065$ at $t = 0.75$ (CI $[1.261, 2.869]$), and $+1.996$ at $t = 0.90$ (CI $[1.111, 2.880]$). The marginal effect of crop, guard, and fusion on top of routing (CRG minus R) is negative throughout the grid and simultaneously significant on $t \in [0.70, 0.88]$, with a representative delta of -0.739 at $t = 0.75$ (CI $[-1.459, -0.019]$); the point estimate remains negative at the permissive and strictest ends of the grid (-0.214 at $t = 0.50$, -0.974 at $t = 0.90$, -0.387 at $t = 0.95$) but those individual thresholds do not clear the simultaneous band. Together with the scalar comparison in §5.4, this confirms with the same statistical protocol used for the B-versus-CRG curve above that routing accounts for a broad,

	Ref-L4		Flickr30K merged-box	
	Midpoint	**Replacement**	**Midpoint**	**Replacement**
Guarded	Acc@0.5: 89.725 Acc@0.9: 61.142 mAcc: 76.013 Mean IoU: +0.691	Acc@0.5: 89.399 Acc@0.9: 63.143 mAcc: 76.166 Mean IoU: +0.712	Acc@0.5: 75.678 Acc@0.9: 38.775 mAcc: 57.948 Mean IoU: -0.301	Acc@0.5: 74.926 Acc@0.9: 37.200 mAcc: 56.463 Mean IoU: -1.123
Unguarded	Acc@0.5: 88.262 Acc@0.9: 59.350 mAcc: 73.864 Mean IoU: -0.454	Acc@0.5: 87.247 Acc@0.9: 62.059 mAcc: 74.491 Mean IoU: -0.884	Acc@0.5: 75.057 Acc@0.9: 38.071 mAcc: 57.003 Mean IoU: -0.726	Acc@0.5: 74.353 Acc@0.9: 36.662 mAcc: 55.952 Mean IoU: -1.665

Figure 1: Guard $\times$ fusion factorial on the retrospective Ref-L4 split and prospective Flickr30K merged-box confirmation. The guarded-midpoint cell (CRG) is the shipped policy in both cases; the unguarded cells lose accuracy on every reported threshold, showing that the guard, not the fusion rule alone, is the load-bearing component.

Cell	Acc@0.5	Acc@0.75	Acc@0.9	mAcc	Mean IoU Δ	IoU W/L/T
Guarded, midpoint (CRG, shipped)	**75.678**	61.046	38.775	57.948	-0.301	5,024 / 5,471 / 3,986
Guarded, replacement	74.926	59.050	37.200	56.463	-1.123	4,430 / 6,075 / 3,976
Unguarded, midpoint	75.057	59.747	38.071	57.003	-0.726	5,541 / 6,003 / 2,937
Unguarded, replacement ($CR^{\dagger}$)	74.353	58.484	36.662	55.952	-1.665	4,876 / 6,686 / 2,919

Table 13: Guard $\times$ fusion factorial on the Flickr30K merged-box confirmation (14,481 rows, 13,360 with an admitted crop candidate; 10,000-iteration image-cluster bootstrap per cell, no new inference). Mean IoU Δ is against the post-routing incumbent R (this factorial's own baseline), not against B as in Table 11 above – the two tables use different reference arms by construction, so their Δ columns are not directly comparable. The same pattern as Ref-L4 holds: the guard, not the fusion rule, is what protects accuracy here – both guarded cells beat both unguarded cells at every threshold, and only the unguarded cells fall below the shipped policy. Unlike Ref-L4, guarded replacement does not trade Acc@0.5 for a higher Acc@0.9 on this dataset (37.200 versus CRG's 38.775); on Flickr the guarded-midpoint cell dominates guarded-replacement on every metric, so the trade Ref-L4 exhibits is not a universal property of the replacement rule.

simultaneously-significant improvement over the incumbent on this prospective evidence, while the additional crop/guard/fusion stages give back part of that gain over a threshold band concentrated in the mid-strict range rather than at the extremes.

Table 14 extends the same synchronized cost audit already reported for the RefCOCO-family split (Table 17) to the Flickr30K merged-box confirmation, using the identical protocol (one untimed warm-up, three CUDA-synchronized timed repeats per row, same RTX 3090). B/R/C/CR/CRG were timed on this dataset only after the unconditional controls, and the five guarded arms' numbers were computed but not previously reported as a table. Zero runtime errors and zero invalid outputs on every arm and size.

Unlike the RefCOCO-family measurement, R_{ALL}/CR_{ALL} accuracy was also scored on Flickr, on the same frozen 512-row subset used for the cost audit above (single pass per row, no new inference beyond what the cost audit already ran). A matched-denominator readout for the other five arms was not initially available – B/R/C/ $CR^{\dagger}$/CRG had only been scored on the full 14,481-row population – so the two groups could only be compared qualitatively. We closed that gap by restricting the archived full-population B and crop-pass artifacts to the identical 512 rows (row selection matched by `record_key` against the cost audit's own frozen manifest, zero new model inference) and rescoring all seven arms on that one matched set, reported in Table 15.

The point estimates reproduce the qualitative pattern already suggested by the unmatched comparison – every treatment arm sits numerically above B, and the two unconditional controls sit above the four selective arms on Acc@0.9 and mAcc – but at 512 rows drawn from only 36 images, the image-cluster bootstrap resolves very little of it: only R's mAcc gain (95% CI $[+0.336, +3.374]$) and R_{ALL}'s mAcc gain (CI $[+0.116, +5.776]$) exclude zero; every other arm's interval, including CRG's (CI $[-0.427, +2.749]$ at mAcc) and CR_{ALL}'s (CI $[-0.591, +4.903]$ at mAcc), does not. This is a sample-size limitation of the 512-row subset, not a reversal of the full-population finding: the same CRG-versus-B contrast is significant at every threshold on the full 14,481-row population (Table 11). We therefore report Table 15 as what it is – a matched-denominator accuracy-cost read-

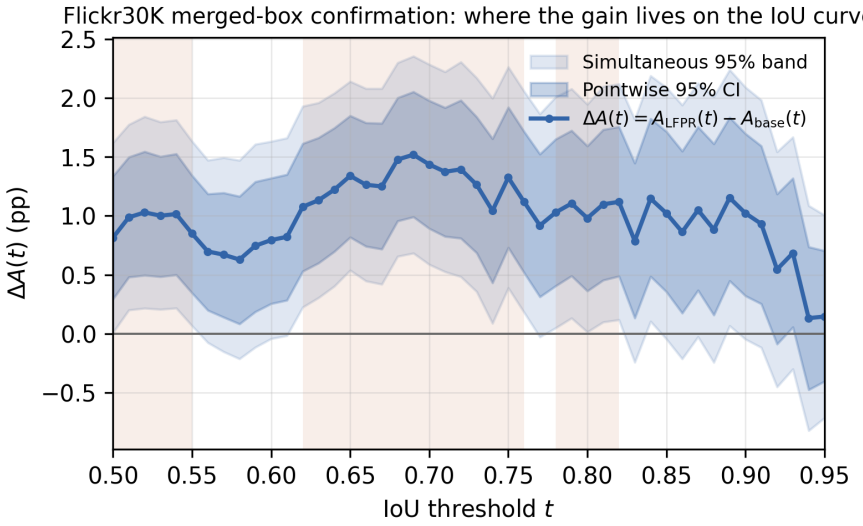


Figure 2: Treatment-minus-incumbent accuracy, $\Delta A(t) = A_{\mathrm{LFPR}}(t) - A_{\mathrm{base}}(t)$, as a function of the IoU threshold on the Flickr30K merged-box confirmation (14,481 rows, 999 images). The darker band is the pointwise 95% interval; the lighter band is the simultaneous (max-$|T|$) 95% band used to license concentration claims without cherry-picking a threshold. Shaded vertical strips mark the grid points where the simultaneous band excludes zero. The gain is positive across nearly the entire threshold range rather than concentrated at one operating point.

Rows	Arm	Calls	Mean s	P95 s	Max alloc. GiB	Invalid %
128	B	1.000	0.837	0.900	16.44	0.00
128	R	1.156	1.299	3.847	17.58	0.00
128	C	2.000	1.646	1.821	16.44	0.00
128	CR	2.156	2.160	4.717	17.58	0.00
128	CRG	2.156	2.244	4.734	17.58	0.00
128	R_{ALL}	1.000	3.072	3.225	17.59	0.00
128	CR_{ALL}	2.000	5.928	6.023	17.60	0.00
512	B	1.000	0.878	0.982	16.44	0.00
512	R	1.250	1.671	4.057	17.58	0.00
512	C	2.000	1.813	2.277	16.44	0.00
512	CR	2.250	2.515	4.895	17.58	0.00
512	CRG	2.250	2.417	4.693	17.58	0.00
512	R_{ALL}	1.000	3.003	3.075	17.59	0.00
512	CR_{ALL}	2.000	6.103	6.489	17.60	0.00

Table 14: Synchronized cost audit on the Flickr30K merged-box confirmation, all seven arms, same protocol and same frozen 128/512-row prefixes as Table 17. CR and CRG share the identical call budget and generation calls (they differ only in which box the already-generated crop candidate is turned into – hard replacement versus midpoint fusion – a CPU-only decision that costs nothing), so their latencies match within timing noise, mirroring the same pattern already reported for the RefCOCO-family audit. R's calls exceed 1.000 by exactly its label-free routing rate on this row prefix (e.g. 0.250 at 512 rows, i.e. 25%), consistent with the 29.8% population-level routing rate reported in Appendix 2.6 – the two figures differ because this is a smaller 512-row sample, not a second measurement of the same population. Peak allocated memory is 16.44 GiB for the two arms that never raise resolution (B, C) and 17.58–17.60 GiB for every arm that can trigger the resolution floor.

out, not an additional significance result – and continue to rely on the full-population table for the paper's confirmatory claims. Within this matched set, CR_{ALL} does not improve on R_{ALL} despite costing 2× the calls: the additional unconditional crop pass is a net negative on every metric at

Arm	Acc@0.5	Acc@0.75	Acc@0.9	mAcc	Mean IoU Δ vs B
B incumbent (pre-resolution)	74.219	64.648	42.773	59.648	—
R routing alone	75.586	66.602	44.531	61.387	+1.175
C guarded crop alone	74.805	63.086	42.578	59.668	+0.323
$CR^{\dagger}$ unguarded negative control	74.805	63.086	41.992	59.609	−0.082
CRG full (LFPR)	75.000	66.406	43.555	60.723	+0.932
R_{ALL} unconditional floor	75.586	67.773	48.047	62.617	+1.828
CR_{ALL} unconditional floor+crop	75.391	67.383	45.117	61.816	+1.590

Table 15: All seven arms scored on the identical frozen 512-row Flickr30K merged-box subset (36 images) used for the synchronized cost audit (Table 14), so accuracy and cost are now comparable on the same denominator throughout this section. B/R/C/ $CR^{\dagger}$/CRG are restricted from the archived full-population run (Table 11) by exact `record_key` match, not rerun; R_{ALL}/CR_{ALL} are the same predictions already used for the cost audit. Deltas and the 10,000-iteration image-cluster bootstrap are against this table's own 512-row B, not the 14,481-row B used elsewhere in this section.

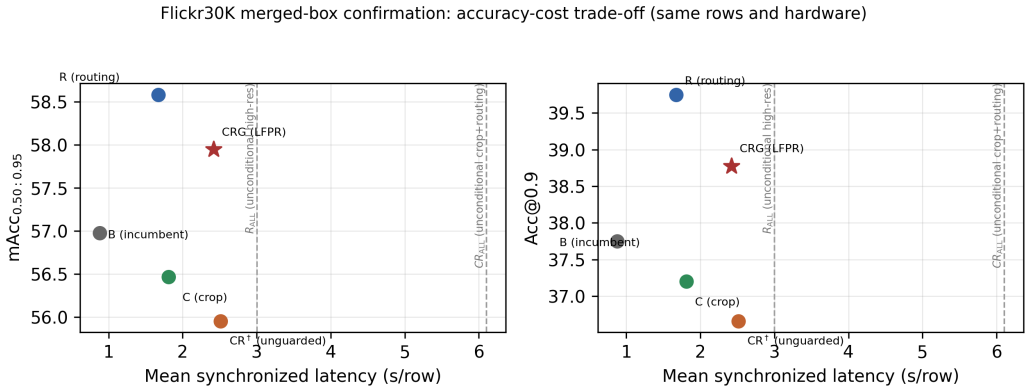


Figure 3: **Accuracy-cost trade-off on the Flickr30K merged-box confirmation.** Each point uses the same 14,481-row evaluation (accuracy, Table 11) and the same synchronized 512-row cost prefix on the same RTX 3090 (cost, this section); no new inference was run to build this figure. R_{ALL}/CR_{ALL} (dashed reference lines) cost 3.42×/6.95× B; their own matched-sample accuracy is reported in the surrounding text rather than plotted here, since it uses the smaller 512-row sample rather than the full population the other five points use. On this dataset, R alone (not CRG) is the accuracy-maximizing point on the frontier – see text.

this sample size, reinforcing why the deployed policy routes crop attempts selectively rather than applying them everywhere.

Figure 3 plots the accuracy-cost trade-off for this dataset, pairing each of the five B/R/C/$CR^{\dagger}$/CRG arms' own full-population accuracy (Table 11) with its own synchronized 512-row cost measurement above; no new inference was run to build this figure. Unlike the RefCOCO-family panel (Figure 5), where CRG is the best matched-cost mAcc point, here R alone reaches the highest mAcc and Acc@0.9 of any arm, including CRG, at lower cost (1.671 vs. 2.417s/row) – the crop/guard/fusion component contributes nothing beyond routing at the mAcc level on this dataset, consistent with the router's outsized isolated contribution already noted above (a 29.8% routing rate here against 7.1% on the RefCOCO family). CRG still beats B significantly at every threshold (Table 11) and the guard remains necessary – $CR^{\dagger}$ falls to the lowest mAcc and Acc@0.9 of any arm here – but on Flickr the accuracy-maximizing point on this frontier is R, not the full shipped policy. This is reported for completeness rather than folded into the main claim: the shipped policy is fixed across every evaluation in this paper, and this trade-off reflects Flickr's own routing rate, not evidence for changing the deployed rule.

Budget (k px)	Selected %	Acc@.5	Acc@.9	mAcc	ΔMean IoU	ΔmAcc
0	0.0	75.893	39.749	58.585	0.000	0.000
50	24.6	74.857	38.582	57.169	-0.890	-1.416
100	43.4	74.808	37.587	56.562	-1.206	-2.023
150	62.4	74.532	37.076	56.214	-1.463	-2.371
200	96.6	74.353	36.662	55.958	-1.665	-2.627
250	100.0	74.353	36.662	55.952	-1.665	-2.633

Table 16: CPU-only budgeted replay of the frozen Flickr raw crop candidate, relative to the post-routing R incumbent. Values are percentages except mean-IoU changes, which are percentage points. Candidate selection uses the stored rank and crop-pixel cost only; all non-zero-budget changes have image-cluster 95% intervals below zero for both mean IoU and mAcc.

2.7 BUDGETED RAW-CROP FRONTIER

The raw crop artifact also permits a cost sensitivity analysis without new model inference. We replayed its stored candidate in rank order, using the crop-window pixel area as the budget and retaining the post-routing R incumbent whenever the candidate did not fit. The policy is label-free: the ground-truth box is read only after selection, and no budget point or candidate is chosen using IoU. The predeclared budgets were 0, 50k, 100k, 150k, 200k, and 250k crop pixels over all 14,481 Flickr merged-box rows.

Table 16 shows a deliberately negative result. As more rows can afford the unguarded raw candidate, the candidate-selection rate increases to 100%, but every non-zero budget is worse than the R incumbent in mean IoU and mAcc. At the full 250k budget, the paired image-cluster 95% intervals for the mean-IoU and mAcc changes are $[-1.972, -1.355]$pp and $[-3.050, -2.210]$pp, respectively. This is a cost-frontier sensitivity of the cached crop candidate, not a new shipped policy or a claim that the guarded LFPR operator should be replaced.

2.8 IOU GAIN CURVE ACROSS ALL FIVE EVALUATIONS

Figure 2 above shows the dense-grid curve for the Flickr30K merged-box confirmation alone. Figure 4 overlays the same $\Delta A(t) = A_{\mathrm{LFPR}}(t) - A_{\mathrm{base}}(t)$ curve, computed with the identical protocol (pointwise and simultaneous max-$|T|$ image-cluster bootstrap, 10,000 iterations) and against each evaluation's own true pre-resolution incumbent B, for all five evaluations: Ref-L4 (retrospective), Flickr30K merged-box (prospective), and RefCOCO/RefCOCO+/RefCOCOg shown individually rather than pooled. Each curve is built from a separately-verified B/CRG per-row pair – Ref-L4 from `baseline_predictions.jsonl` and `fused.jsonl`, Flickr from the merged-box confirmation's own `baseline/merged.jsonl` and `lfpr_predictions.jsonl`, and RefCOCO/RefCOCO+/RefCOCOg from the true pre-resolution artifact used to correct Table 7 – and at $t \in \{0.50, 0.75, 0.90\}$ every curve reproduces its evaluation's own fixed-threshold point estimate to within 10^{-10}, the machine-checked consistency gate this figure requires.

All five curves start positive at $t = 0.50$ and stay positive through most of the permissive-to-mid range. Ref-L4 and Flickr grow *larger* at strict thresholds (Ref-L4 reaches $+5.35$ points at $t = 0.90$ and stays above $+5$ through $t = 0.94$); RefCOCO/RefCOCO+/RefCOCOg instead grow smaller and turn negative only close to the strictest end of the range – RefCOCO crosses zero between $t = 0.84$ and 0.85 (reaching -1.13 at $t = 0.93$), RefCOCO+ between $t = 0.91$ and 0.92 (reaching -0.86 at $t = 0.94$), and RefCOCOg between $t = 0.90$ and 0.91 (a shallow dip to -0.30 at $t = 0.91$ before recovering to slightly positive by $t = 0.95$). This is consistent with every fixed-threshold result already reported: the deployed policy improves Acc@0.5/0.75/mAcc/mean IoU broadly across the RefCOCO family (Table 7), and its strict-IoU effect softens toward the top of the threshold range rather than being uniformly positive like Ref-L4 and Flickr's.

2.9 SYNCHRONIZED COST AND ROBUSTNESS AUDIT

The generalist cost audit uses deterministic 128- and 512-row prefixes of the frozen transfer order. Every arm runs in a fresh process with batch size one, one untimed warm-up, and three CUDA-synchronized timed repeats per row. Both sizes passed the declared admission checks for complete

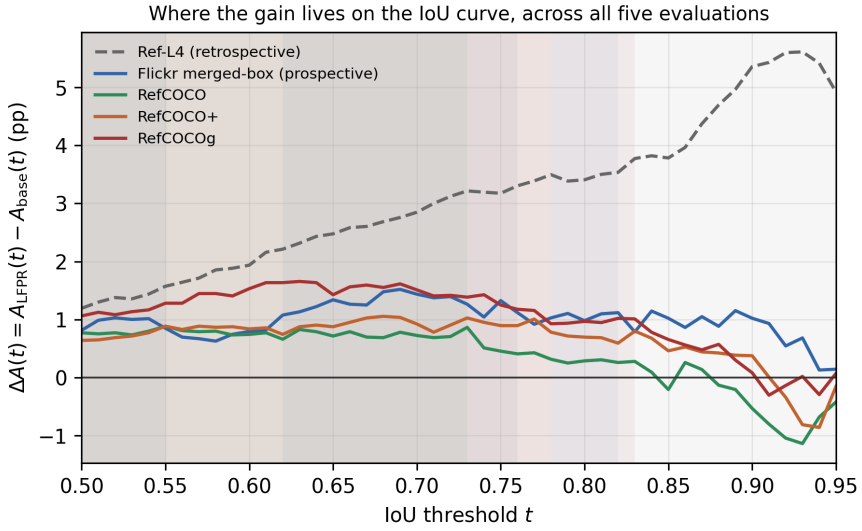


Figure 4: Treatment-minus-incumbent accuracy, $\Delta A(t)$, as a function of the IoU threshold, across all five evaluations, each against its own true pre-resolution incumbent B. Shaded vertical strips mark each series' own simultaneous-band significant intervals. Ref-L4 and Flickr (retrospective and prospective evidence respectively) grow larger at strict thresholds; RefCOCO/RefCOCO+/RefCOCOg (shown individually, not pooled) stay positive through most of the range and turn negative only close to $t = 0.85$–0.92, consistent with the deployed policy's broad per-dataset improvement and softening strict-IoU effect reported at fixed thresholds.

exact-key manifests, finite timings, and absence of OOM or unclassified runtime failures. A later processor audit showed that the requested high-resolution floor had not been applied in R, CR, or CRG; those timing values were withdrawn. The unaffected B and C measurements are retained below, and Table 17 restores R/CR/CRG with the floor verified in effect, alongside the two unconditional controls R_{ALL} and CR_{ALL}.

An earlier generalist timing table reported B/C mean latency of 343.7/683.5 ms (128 rows) and 347.6/677.1 ms (512 rows). That table is withdrawn: the canonical policy-matched audit below, run under the same declared protocol, measures B at 0.945–1.049 s and C at 1.876–2.315 s per row – roughly 2.7× higher. We could not reconcile the discrepancy from the archived environment record (neither audit's log fixes the generation-token count, software commit, or GPU clock state precisely enough to attribute the gap to a specific cause), so rather than report two mutually inconsistent latency figures under the same label, we retain only the canonical audit below as the paper's cost claim.

The same five policy arms were also rerun on an independent deterministic 128-row prefix with the realized multimodal prompt-token count recorded for each model call. We report an operation-count proxy, not hardware-measured FLOPs: each call is estimated as $2N_{\mathrm{params}}$ per input token using $N_{\mathrm{params}} = 8{,}767{,}123{,}696$ for Qwen3-VL-8B-Instruct, and the row value sums the calls issued by that arm. The resulting proxy is included alongside the canonical 512-row latency audit below; it should not be read as an exact hardware FLOP or energy measurement.

Figure 5 plots the accuracy-cost trade-off implied by Tables 9 and 17 directly, addressing the question of why LFPR does not simply spend its second forward pass on another unconditional high-resolution pass.

The specialist audit reuses the same row prefixes but records one timed repeat after a per-row warm-up because EGM generation is substantially slower. The 128- and 512-row gates both pass for all four arms: every expected key is present, every timing is finite, and there is no OOM or unclassified runtime failure.

Arm	Calls	Mean s	P95 s	Max alloc. GiB	Invalid %	Proxy TFLOPs/row
B	1.000	0.945	1.049	16.49	0.39	6.10
C	1.996	1.876	2.315	16.49	0.39	10.40
R	1.025	1.048	1.252	17.58	0.39	6.68
CR	2.021	1.977	2.347	17.58	0.39	10.98
CRG	2.021	1.865	1.985	17.58	0.39	10.98
R_{ALL}	1.000	3.291	3.510	17.59	0.00	—
CR_{ALL}	2.000	6.121	6.522	17.60	0.00	—

Table 17: Canonical policy-matched 512-row synchronized cost audit with the resolution floor verified in effect, including the unconditional R_{ALL}/CR_{ALL} controls. One untimed warm-up plus three CUDA-synchronized timed repeats per row, zero runtime errors on every arm. This table supersedes the earlier B/C figures withdrawn above; treat only these absolute latencies, not the earlier ones, as the paper's cost claim. Selective routing (R/CR/CRG) costs little beyond B/C because only the flagged small-object rows pay the high-resolution premium; forcing that premium onto every row (R_{ALL}/CR_{ALL}) costs 3.10–3.14× more than its selective counterpart at matched call count, quantifying why selective routing is the deployed policy rather than an always-high-resolution default. The final column is the independent 128-row operation proxy described above; the unconditional controls were not included in that five-arm rerun and are therefore marked as unavailable.

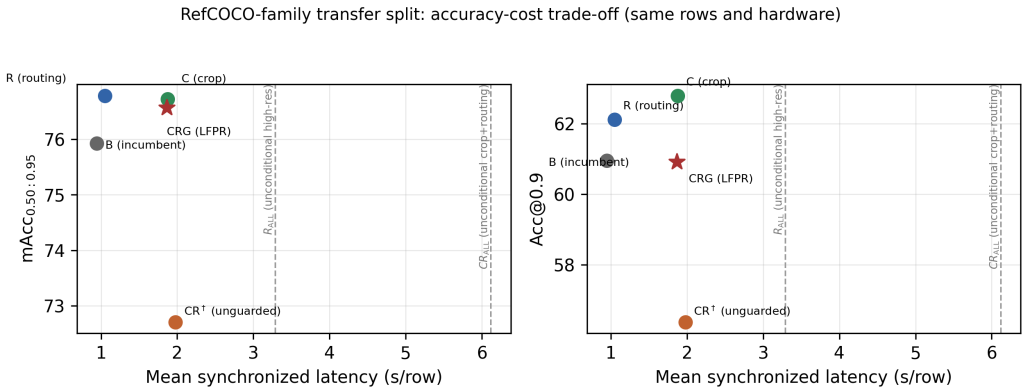


Figure 5: **Accuracy-cost trade-off on the RefCOCO-family transfer split.** Each point uses the same 30,969-row evaluation (accuracy, Table 9) and the same synchronized 512-row cost prefix on the same RTX 3090 (latency, Table 17); no new inference was run to build this figure. CRG (LFPR, starred) costs 1.865s/row against B's 0.945s/row, a 1.97× increase – consistent with the paper's "approximately doubles inference latency" characterization, since every admitted row pays one crop pass and router-flagged rows may pay an additional high-resolution pass. The two unconditional controls, R_{ALL} and CR_{ALL} (dashed reference lines), cost 3.48× and 6.48× B respectively – 3.10–3.14× more than their selective counterparts R and $CR^{\dagger}$ at matched call count (Table 17) – for a policy that no longer distinguishes rows worth the extra pass from rows that are not. Their accuracy is now measured on a matched 26,488-row RefCOCO validation split (Table 18): both beat the matched B arm across all reported metrics, while R_{ALL} is stricter-threshold better than CR_{ALL} (Acc@0.9/mAcc). Because that accuracy readout is on a validation split and this figure's selective-arm points remain on the frozen test partitions, this panel is still a cost-vs-cost picture rather than a same-split selective-versus-unconditional accuracy frontier.

At 512 rows, EGM-4B throughput changes from 0.216 to 0.110 rows/s and maximum allocated memory remains 8.45 GiB; EGM-8B changes from 0.218 to 0.109 rows/s and remains at 16.48 GiB. EGM-4B has one invalid row in both baseline and LFPR (0.20%); EGM-8B has none. The paired mean overhead of LFPR is 4.458 s and 99.0 generated tokens for EGM-4B, and 4.574 s and 98.7 tokens for EGM-8B.

A separate same-host anchor reruns the generalist B and corrected CRG arms on GPU 0 of the RTX 3090 host over the exact 128-row prefix. It retains the generalist protocol of one warm-up plus

Arm	Acc@0.5	Acc@0.75	Acc@0.9	mAcc	Mean IoU
B (matched val baseline)	89.218	81.029	62.753	76.475	82.557
R_{ALL}	90.328	83.581	69.394	79.514	84.327
CR_{ALL}	90.464	83.513	68.065	79.275	84.233

Table 18: RefCOCO validation-split unconditional-resolution controls (26,488 rows; 2,738 images), scored on one matched row set with 10,000-iteration image-cluster bootstrap deltas against B. R_{ALL} gains $+1.102/+2.552/+6.641/+3.039$ points at Acc@0.5/0.75/0.9/mAcc (95% CIs $[+0.686, +1.519]/[+1.913, +3.181]/[+5.706, +7.574]/[+2.574, +3.508]$); CR_{ALL} gains $+1.246/+2.484/+5.312/+2.800$ (95% CIs $[+0.820, +1.656]/[+1.849, +3.136]/[+4.394, +6.240]/[+2.347, +3.261]$).

Rows	System	Calls	Gen. tokens	Mean s	P95 s
128	EGM-4B	1.000	98.5	4.798	6.055
128	+ LFPR	2.000	197.4	9.250	11.395
128	EGM-8B	1.000	98.7	4.553	5.148
128	+ LFPR	2.000	196.9	9.116	11.536
512	EGM-4B	1.000	98.9	4.622	5.503
512	+ LFPR	1.998	197.9	9.080	10.861
512	EGM-8B	1.000	97.8	4.581	5.764
512	+ LFPR	2.000	196.5	9.154	10.531

Table 19: Specialist cost stability across audit sizes. Generated tokens are summed across calls; each row has one synchronized timed repeat after a per-row warm-up.

three timed repeats, giving 384 timed observations per arm. B has mean/p95 latency 1.041/1.185 s, 0.961 rows/s, 1.000 calls, 17.5 generated tokens, and 16.49 GiB peak allocated memory. CRG has 1.948/2.284 s, 0.513 rows/s, 2.008 calls, 35.3 tokens, and 17.58 GiB. Both arms have zero invalid outputs and zero runtime errors. Only one of the 128 rows is routed to the high-resolution pass, so this anchor is a same-hardware cross-system comparison, not a substitute for the corrected full-size generalist cost grid. On these rows, EGM-4B/8B baseline latency is $4.61\times/4.37\times$ B, and EGM-4B/8B with LFPR is $4.75\times/4.68\times$ generalist CRG; generated-token ratios are about $5.6\times$ in both comparisons.

The retrospective full-split effects by source are:

The strict-IoU intervals for the full-split comparison are +3.173 points at IoU 0.75 (LCB +2.831), +5.354 points at IoU 0.9 (LCB +4.898), and +3.066 points for mAcc (LCB +2.857). They are exploratory because endpoint and pipeline selection are not reflected in those historical one-sided bounds.

2.10 Router diagnostics

The label-free router flagged 463 of 2,000 pilot rows (23.2%) and 6,641 of 31,921 full-split rows (20.8%). Where annotation boxes are available for diagnosis, predicted-box buckets agree with annotation-defined buckets on 88.85% of pilot rows and 89.83% of full-split rows. This statistic is not a classifier accuracy and is not used as a gate; it describes why the deployment proxy is a reasonable but imperfect substitute for the annotation-stratified diagnostic.

2.11 Malformed-row handling

The pilot has zero runtime and invalid-prediction rows. The full-split baseline contains one inverted-coordinate completion. The LFPR fusion implementation passes an invalid incumbent through unchanged rather than raising an exception or inventing a crop. The resulting full-split report therefore contains one inherited runtime/invalid count. This row is included in both matched arms and is not hidden from the error accounting. It is a protocol repair for a degenerate baseline case, not a model or hyperparameter change.

Source	Rows	Acc@0.5 Δ	One-sided LCB
COCO	19,889	+0.875	+0.693
Objects365	12,032	+1.720	+1.260

Table 20: Full-split source-stratified effects. Source rows use the same frozen parser and paired image bootstrap; no interaction interval was archived.

Bucket	Rows (%)	Acc@0.5	Gap to large
Small	6,665 (20.9%)	78.92	−13.81
Medium	16,311 (51.1%)	90.15	−2.58
Large	8,945 (28.0%)	92.73	–

Table 21: Baseline Acc@0.5 by annotation-derived target size, before any resolution intervention.

2.12 Size-bucket baseline accuracy

Table 21 gives the raw, pre-intervention accuracy by annotation-derived target size on the full 31,921-row test split; this is the diagnostic that originally motivated targeting the small bucket with additional resolution (Section 5.7 below).

2.13 Resolution-ladder and prompt-sensitivity pilot designs

All pilots in this subsection use 800 rows and the historical paired image-level bootstrap. The complete access chronology is not auditable, so these comparisons are exploratory regardless of when individual runs were launched. The staged floor-selection summaries (Section 5.8 below) report a moderate floor of $1{,}048{,}576$ pixels at $+0.75$ points relative to default (one-sided LCB -0.99) and the chosen $4{,}194{,}304$ floor at $+1.875$ points relative to default (LCB $+0.25$). The medium-bucket extension pilot applied the identical $4{,}194{,}304$-pixel policy to an 800-row medium-bucket sample and gave -0.125 points (95% CI $[-1.86, +1.62]$).

The later prompt pilots (Section 5.9 below) used an 800-row Objects365 sample paired against the same rows' recombined post-floor predictions. Two variants replaced the default prompt text without changing any other setting: a *tight-box* variant explicitly asking for the tightest enclosing box gave $+0.125$ points (95% CI $[-1.39, +1.65]$, LCB -1.13), and a *fine-grained* variant emphasizing instance-level detail gave $+0.25$ points (95% CI $[-1.25, +1.75]$, LCB -1.00). A further resolution rung at $8{,}388{,}608$ pixels ($\sim 2\times$ the chosen floor), on the same 800-row small-bucket sample used for the floor-selection ladder, gave 82.875% versus 81.750% at $4{,}194{,}304$: a $+1.125$-point stepwise difference (95% CI $[-0.38, +2.73]$, LCB -0.13). This value is not default-relative and therefore cannot establish that the ladder peaked. The former Figure 3 was removed; row-level common-reference rescoring is required before drawing a resolution-curve conclusion.

The zero-shot self-correction control (Section 5.10 below) used the same frozen incumbent with no fine-tuning: the model is shown the original image with its own first-pass box rendered as an overlay and asked to return one replacement box, with neither a guard nor a fusion step applied to the output. On its own 800-row pilot this gives Acc@0.5 $82.50 \rightarrow 80.50$ (-2.00 points, LCB -3.14), Acc@0.75 $72.625 \rightarrow 70.50$ (-2.125 points, LCB -3.78), Acc@0.9 $52.375 \rightarrow 54.625$ ($+2.25$ points, LCB $+0.12$), and mAcc $68.35 \rightarrow 67.29$ (-1.06 points, LCB -2.14).

2.14 Backbone-scaling pilot design

The backbone-scaling control in Table 25 uses the official bf16 checkpoint of the same model family at approximately $4\times$ the incumbent's parameter count, under the unchanged prompt, parser, resolution policy, and evaluator. A 500-row engineering pilot completed with zero errors and clean bare-box parsing, confirming the larger checkpoint is compatible with the frozen harness before spending a larger budget. The confirmation pilot used 380 rows (190 COCO, 190 Objects365, 344 unique images) rather than a larger round number: by this point in the project, the shared image-disjoint validation pool had already been drawn on by earlier pilots and related controls, and 380 was the largest balanced size the remaining unvisited pool supported without compromising image-

Arm	Acc@0.5	Acc@0.75	Acc@0.9	$\text{mAcc}_{0.50:0.95}$
B (direct answer)	72.933	57.500	36.633	54.703
+LFPR (CRG)	74.233	60.800	37.800	56.673
Δ (95% CI)	+1.300 [0.494, 2.122]	+3.300 [2.297, 4.349]	+1.167 [0.236, 2.085]	+1.970 [1.377, 2.572]

Table 22: Qwen3-VL-4B-Instruct, full guarded-fusion pipeline vs. its own direct-answer baseline, on the same 3,000-row stratified Flickr30K merged-box subset used for the cross-family screen. All four 95% paired image-cluster bootstrap CIs exclude zero.

disjointness from every prior pilot. On this confirmation pilot, Acc@0.9 regresses by -1.32 points (LCB -4.56), the Objects365 subset effect is -0.53 points (failing outright), and the COCO LCB is -2.07 points, outside the historical -0.5-point non-inferiority band. No full-test run was taken, and the FP8-quantized variant of the same checkpoint was not separately piloted, since quantization would only be expected to add noise in the same failing direction already observed at full precision.

The pilot above only tested the larger direction. To test the smaller direction of the same model family – a stronger test in one respect, since a smaller same-architecture checkpoint reuses the incumbent's frozen crop/guard/fusion policy directly rather than requiring new adapter engineering – we ran the complete $B/R/C/CR/CRG$ pipeline on Qwen3-VL-4B-Instruct (roughly $0.5\times$ the incumbent's parameter count, the identical `Qwen3VLForConditionalGeneration` architecture class), under the unchanged prompt, parser, resolution floor, crop context factor, and guard thresholds. It was evaluated on the same 3,000-row stratified Flickr30K merged-box subset used for the cross-family screen below, with zero runtime errors on both the direct-answer arm and the full guarded-fusion arm. Table 22 reports the paired result: every metric improves, and every 95% CI excludes zero, matching the direction (though not the magnitude) of the 8B incumbent's own LFPR gain – the mechanism is not specific to the incumbent's particular parameter count.

2.15 Cross-family checkpoint screening

The backbone-scaling control above stays inside the incumbent's own model family. To screen genuinely different architectures, we evaluated two candidates' own direct-answer (arm B) accuracy before committing to the substantially larger engineering effort of wiring a new checkpoint into the full router/crop/guard/fusion pipeline, following the same B-arm-first protocol already used elsewhere in this project to screen candidate checkpoints cheaply: a candidate whose own accuracy sits meaningfully below the incumbent offers nothing for LFPR to refine into a win.

GLM-4.1V-9B-Thinking (glm4v architecture, 9B) was evaluated on a 4,000-row proportionally stratified subset of the Flickr30K merged-box confirmation (stratified on referring-expression length and ground-truth box area, disjoint from no other evaluation in this paper). Table 23 reports its own Acc@0.5/0.75/0.9 and mAcc with an image-cluster bootstrap CI, against the incumbent's own Acc@0.5 on the full 14,481-row population as context (an unpaired reference, not a matched comparison). Its Acc@0.5 sits -4.69 points below the incumbent's, with the 95% CI excluding the incumbent's point estimate. Two further practical problems compound this: box-parser success is only 83.1% ($3{,}323/4{,}000$ rows), well below the incumbent's own 99.6% (Table 17's invalid-output rate), and per-row latency has a severe unpredictable tail (up to 153.4s on a single row against a 9.2s mean) driven by unbounded "thinking" trace length – a cost profile that would compound badly once the guard/crop pipeline's extra forward passes are layered on top.

Qwen2.5-VL-7B-Instruct (a different generation and architecture within the Qwen family) was evaluated earlier in this project as an independent harness-reconciliation anchor: on an identical 3,000-row random subsample of Ref-L4-test, it reaches Acc@0.5 $= 81.57\%$ ($2{,}910/3{,}000$ valid predictions, 97.0% parser success) against the incumbent's own Acc@0.5 $= 88.03\%$ on the same rows through the same harness – a -6.47-point paired, same-harness gap. Its own reproduction also lands closer to its publisher's reported Ref-L4-test number ($+0.77$ points offset) than GLM-4.1V-9B-Thinking's reproduction does (-2.90 points offset), indicating a more faithfully reproduced evaluation.

Neither candidate's own accuracy sits close enough to the incumbent's to justify treating cross-family transfer as a general positive result under the project's screening rule. We nevertheless completed the full $B/R/CRG$ pipeline for GLM-4.1V-9B-Thinking at scale (§2.16 below) to mea-

Checkpoint	Dataset (rows)	Acc@0.5	Parser success	Incumbent gap
GLM-4.1V-9B-Thinking	Flickr merged-box (4,000)	70.18	83.1%	−4.69pp
Qwen2.5-VL-7B-Instruct	Ref-L4-test (3,000, paired)	81.57	97.0%	−6.47pp

Table 23: Cross-family checkpoint screening summary. The two rows use different evaluation datasets (noted in column 2) and are not directly comparable to each other in absolute accuracy; each is compared only to the incumbent on its own dataset. Parser success is the fraction of rows producing a usable bare box.

Arm	Acc@0.5	Acc@0.75	Acc@0.9	$\text{mAcc}_{0.50:0.95}$	Mean IoU
B (direct answer)	70.47	58.46	39.19	55.65	63.96
R (resolution-only)	70.57	59.66	39.14	56.21	64.09
CRG (full LFPR)	69.97	57.41	38.09	54.90	63.46

Table 24: GLM-4.1V-9B-Thinking, full $B/R/CRG$ pipeline at scale (1,998 rows, 999 images, identity gate passed, 10,000-iteration paired image-cluster bootstrap). R improves significantly over B at Acc@0.75 only (+1.20pp, 95% CI $[0.10, 2.25]$); CRG is significantly below B on $\text{mAcc}_{0.50:0.95}$ (-0.75pp, CI $[-1.40, -0.11]$). All other paired deltas' 95% CIs include zero.

sure whether the frozen operator transfers beyond the Qwen family. Qwen2.5-VL-7B-Instruct – the more procedurally viable of the two under the screening above – was not similarly extended; the matched NVILA-8B pilot (§2.17) remains an inconclusive small-scale cross-family comparison. We report all screening attempts in full rather than omitting negative results, consistent with this paper's evidence-tier discipline of not repackaging screening pilots as confirmatory evidence.

2.16 Full cross-family pipeline: GLM-4.1V-9B-Thinking at scale

Despite the screening-stage concerns above (lower own-accuracy, weaker parser reliability, and a heavy latency tail), we completed the full $B \to R \to CRG$ pipeline for GLM-4.1V-9B-Thinking under the unchanged prompt, parser, resolution floor, crop context factor, and guard thresholds, on a 2,000-row stratified Flickr30K merged-box subset (999 distinct images; 1,998 valid rows after removing 2 rows with unusable ground truth). The CRG arm was computed as two 999-row shards processed on separate devices under an identical protocol and merged before scoring; the merge was verified free of duplicate or missing record keys before the identity-gated scorer ran. All three arms share identical record keys, image IDs, and ground-truth boxes (identity gate passed).

Table 24 reports the paired image-cluster bootstrap result. R improves significantly over B at Acc@0.75 only (+1.20pp, 95% CI $[0.10, 2.25]$); every other R-vs-B delta is not distinguishable from zero. CRG does not improve over B on any metric and is significantly *below* B on $\text{mAcc}_{0.50:0.95}$ (-0.75pp, CI $[-1.40, -0.11]$), with Acc@0.9 bordering significance (CI $[-2.20, 0.00]$). This is the same direction as the incumbent's own CRG-vs-R sensitivity on Flickr (Appendix 2.6): the guarded crop-refinement step is not uniformly beneficial across every threshold and every backbone, and here it does not even recover the small resolution-only gain that R alone captures. Combined with GLM-4.1V-9B-Thinking's weaker own-accuracy and parser reliability under this harness (§2.15), this result is consistent with – not contradictory to – the paper's screening rule: a backbone whose own localization is less reliable does not automatically benefit from an unmodified crop/guard/fusion policy tuned on a stronger incumbent.

This result is exploratory and optional. It does not change the paper's main claims, which concern the incumbent Qwen3-VL-8B-Instruct backbone; instead, it places an explicit negative boundary on cross-family transfer. We report it in full for transparency rather than treating either the screening-stage numbers or the small NVILA pilot as confirmatory evidence for a second backbone family.

2.17 COMPATIBILITY AND MATCHED-PILOT ANALYSIS OF RECENT TRAINING-FREE BASELINES

Section 2 notes that Chain-of-Caption is complementary to our geometry-based accept/reject step rather than a substitute for it. The official release is available and provides an NVILA-8B/RefCOCO evaluation path, but it does not provide a released path compatible with our Qwen3-VL backbone. We therefore do not treat cross-paper numbers as a same-backbone comparison; instead we ran a matched pilot using the official NVILA-8B release, Chain-of-Caption's own evaluation harness and prompt, and the same 100 rows (99 distinct images, seed-fixed) for direct NVILA-8B (B), Chain-of-Caption, and our full guarded, fused refinement policy (CRG).

At this scale, none of the three arms is distinguishable from the others: Acc@0.5/0.75/0.9 are numerically tied across all three (23.0/3.0/0.0%), and the CRG-vs-B effect on mAcc is -0.4pp with a 95% CI of $[-1.5, 0.8]$pp. This is not an artifact of CRG silently failing to engage: its predicted box differs from B's on 46 of the 100 rows, and the crop-refinement step is accepted (not guard-rejected) on 75 of them. Rather, direct NVILA-8B's zero-shot localization under this harness is frequently wrong by a wide margin rather than merely imprecise, and a local crop-and-refine step is not designed to recover from that – it corrects imprecision in an already-plausible box, not a wrong region entirely. We also checked whether the low absolute baseline reflects our prompt rather than the backbone: our prompt is word-for-word identical to Chain-of-Caption's own official RefCOCO prompt template, and two alternative phrasings tested on the same rows landed in the same low regime, ruling out prompt wording as the explanation. Consistent with this paper's evidence-tier discipline (§2.15), we report this $n = 100$ pilot as exactly that – a pilot, not a confirmatory cross-backbone result – rather than omitting a result that does not support the main claim. The same question arises for the two other closest training-free methods on the same benchmark family, Entropy-Gradient Grounding (EGG) and Test-time Scaling over Perception (TTSP); this appendix expands the compatibility check behind that decision rather than leaving it asserted.

EGG (Gröpl et al., 2026) back-propagates next-token entropy to visual token embeddings to produce a grounding signal without an auxiliary detector. We did not conduct a head-to-head comparison with EGG. Its official implementation supports only LLaVA-1.5/1.6, whereas our method is evaluated with Qwen3-VL-8B-Instruct – the authors' own repository lists Qwen 2.5 and InternVL 3.5 support as not-yet-released roadmap items, and Qwen3-VL is not on that roadmap at all – so porting or reimplementing EGG for our backbone would not constitute a faithful reproduction. The released repository also provides no license, an independent reason not to clone, modify, or run its code without the authors' permission; this applies equally to comparing on EGG's own supported backbone (LLaVA-1.5/1.6) instead of Qwen3-VL, since the missing license does not depend on which backbone hosts the comparison. We therefore report this compatibility limitation rather than presenting an unsupported experimental result. This assessment could change if the EGG authors release a Qwen3-VL-compatible implementation or grant explicit authorization and technical support.

TTSP (Jiang et al., 2026) generates multiple perceptual exploration traces, filters them by entropy-based confidence, and iteratively refines a structured evidence ledger. Its public Apache-2.0-licensed code does support Qwen3-VL-8B-Instruct directly – architecturally, this is the more compatible of the two candidates. The incompatibility is instead at the task-interface level: TTSP's released pipeline and benchmark CLI take a closed-set multiple-choice question (`question` plus an `options` list) and return a reliability-weighted vote over those options; it has no open bounding-box output mode. Adapting it to referring-expression grounding would mean replacing its core voting mechanism (which tallies votes over discrete options) with a continuous box-aggregation rule of our own design – at that point the result would be a new method inspired by TTSP's evidence-ledger idea, not a reproduction of TTSP itself, and reporting it as "TTSP's accuracy on this benchmark" would misattribute our own design choices to the original authors.

Both checks were run against the actual released code and documentation, not assumed from the papers alone, before concluding neither supports a faithful head-to-head comparison on this task without either changing the backbone (EGG) or replacing the core method (TTSP).

3 CONTROL EXPERIMENTS AND STOPPING RULES

The following table records the nearby routes that were evaluated on image-disjoint pilots or reserved evaluation data. "Closed" records the historical stopping decision under the then-used one-sided gate; it does not denote a multiplicity-adjusted confirmatory conclusion.

Table 25: Matched controls and dispositions.

Route	Main result	Disposition
Released EGM transfer	-7.45 points at Acc@0.5; 95% CI $[-9.56, -5.34]$	Closed as a zero-shot transfer; native EGM recipe still needs a matched-compute replication.
Supervised coarse-to-fine	-0.543 points at Acc@0.5; LCB -1.264	Closed before the planned larger run.
Transferred parallel box decoding	-15.80 points; LCB -17.64	Transfer route closed.
SAM3 projection	Positive strict-IoU effects but no headline gate	Closed as a standalone replacement.
SAM3 midpoint fusion	$+0.10$ points at Acc@0.5; LCB -0.20 on fresh data	Retrospective fusion signal did not replicate.
Full crop replacement	$+0.50$ points at Acc@0.5; LCB -0.10	Crop replacement closed; strict-IoU signal retained.
Standalone midpoint fusion	$+0.70$ points on the reused pilot; source-weighted projection $+0.65$	Useful component; insufficient alone.
LFPR	$+1.194$ points on full test; LCB $+0.984$	Strongest matched full-split result in the audit; retrospective.
Frozen-split RefCOCO-family transfer	$+0.817$ points on pooled RefCOCO/+/g; 95% CI $[+0.639, +1.001]$	Confirmed positive transfer at Acc@0.5; Acc@0.9 is unchanged, not because an unguarded arm banks the strict-IoU gain instead (the genuine unguarded negative control underperforms the incumbent on every metric, Table 9), but because the guard trades a small Acc@0.9 concession for the Acc@0.5 gain directly; a separate cross-family evaluation was subsequently completed with GLM-4.1V-9B-Thinking (§2.16) as an explicit negative boundary on cross-family generalization – resolution-only yields a limited benefit there, while the full CRG operator does not transfer and significantly decreases mAcc – and the released-specialist baseline comparison is reported separately in Section 5.3.
Backbone scaling (32B)	$+0.26$ points at Acc@0.5; LCB -2.45	Scaling closed; fails on both data sources and regresses Acc@0.9.
Backbone scaling (4B, full LFPR sweep)	$+1.970$ points $\text{mAcc}_{0.50:0.95}$; 95% CI $[+1.377, +2.572]$	Confirmed positive; the mechanism also transfers to a half-size same-family checkpoint (Qwen3-VL-4B-Instruct, $\approx 0.5\times$ the incumbent's parameter count; Appendix 2.14).

continued on next page

Table 25 continued from previous page

Route	Main result	Disposition
Guard/fusion threshold sweep	No configuration Pareto-dominates the shipped rule across a max-zoom-consistency grid of 0.15–1.00 and a fusion-weight grid of 0.5–1.0	Closed: every accuracy-improving configuration fails a Holm-adjusted significance check; the shipped threshold is not improvable within the swept grid.
Confidence-gate mechanism control (Ref-L4, R incumbent)	An independent crop-refinement pass over all 31,921 Ref-L4 R rows, recording both the shipped geometry guard's admission decision and the model's own greedy-decoding confidence (mean token log-probability) on every row, then scoring the same replacement/midpoint-fusion formulas gated by each rule in turn (image-clustered, 9,467 images). Geometry-gated fusion reproduces a small Acc@0.5 gain over R, $[+0.139, +0.399]$ (95% CI), with the same Acc@0.9/mAcc cost already reported for $CRG - R$ on the RefCOCO family; the confidence-gated analogue does not reproduce even that Acc@0.5 gain, $[-0.334, -0.098]$, and geometry-gated replacement-only is far worse than either fusion arm, mAcc $[-2.987, -2.437]$. The two admission rules agree on 52.5% of rows (15,411 both-admit, 1,355 both-reject)	Closed: on this independent replication, a frozen label-free confidence threshold is not a substitute for the shipped geometry guard – only the geometry-gated arm recovers the small Acc@0.5 gain over routing, and both share the Acc@0.9/mAcc trade-off already attributed to the post-routing marginal effect (Appendix 2.3; §5.4), so agreement on *which* rows to admit does not by itself explain the guard's contribution.

continued on next page

Table 25 continued from previous page

Route	Main result	Disposition
Cross-family incumbent candidate (zero-shot, pre-refinement)	Full 31,921-row result: Acc@0.5 53.933, Acc@0.75 39.958, Acc@0.9 22.22, and mean IoU 49.641, versus 88.531 Acc@0.5 for the Qwen3-VL-8B incumbent; 6,584 rows (20.6%) were unparseable or malformed and scored as zero IoU. The full split confirms the 500-row pilot's negative direction (pilot Acc@0.5 50.60).	Closed at the zero-shot stage; the pilot and full-split directions agree, so the candidate was not carried into the refinement pipeline.
Full-parameter fine-tune	Image-disjoint 2,000-row held-out evaluation: Acc@0.5 87.50 → 86.05 (−1.45), Acc@0.75 76.15 → 74.55 (−1.60), Acc@0.9 54.70 → 54.40 (−0.30)	Closed: underperforms the frozen incumbent at every threshold; the training loss plateaus after warmup, consistent with the negative held-out result.
Iterative zoom (2 unconditional rounds, no guard)	Acc@0.5 89.00 → 86.80 (−2.20), Acc@0.9 55.60 → 51.20 (−4.40) on a 500-row sample	Closed: repeated unguarded cropping compounds error rather than refining it.
Random- and center-crop placement controls (matched call count)	Acc@0.5 collapses to 42.00 (random) and 44.20 (center) vs. 89.00 incumbent on the same 500-row sample	Supports the mechanism claim: the crop's gain requires an incumbent-informed window, not merely a second look.
Self-consistency, matched call count (2 samples, temperature 0.7, midpoint-fused)	Acc@0.5 89.00 → 88.40, Acc@0.9 55.60 → 52.60 on the same 500-row sample	Supports the mechanism claim: re-sampling the identical question does not recover the crop's gain.

The controls are not presented as a universal impossibility theorem. They close the tested instantiations under the declared protocol and justify keeping the main claim narrow. A substantially larger backbone within the same model family (Qwen3-VL-32B, roughly $4\times$ the parameters) was also evaluated under the identical frozen prompt, parser, and evaluator; it failed every admission requirement, with a positive but non-significant Acc@0.5 point estimate (LCB negative), a regression at the stricter Acc@0.9 threshold, and a failing effect on the Objects365 subset (Table 25). Scaling the backbone within this family therefore does not reliably close the residual gap under this harness. LFPR's positive result does not imply that all spatial heads or proposal sources would fail; it shows that this specific label-free composition is useful without reaching absolute SOTA.

4 Procedure in pseudocode

The prose in Section 3 is normative; this restatement is provided for implementers.

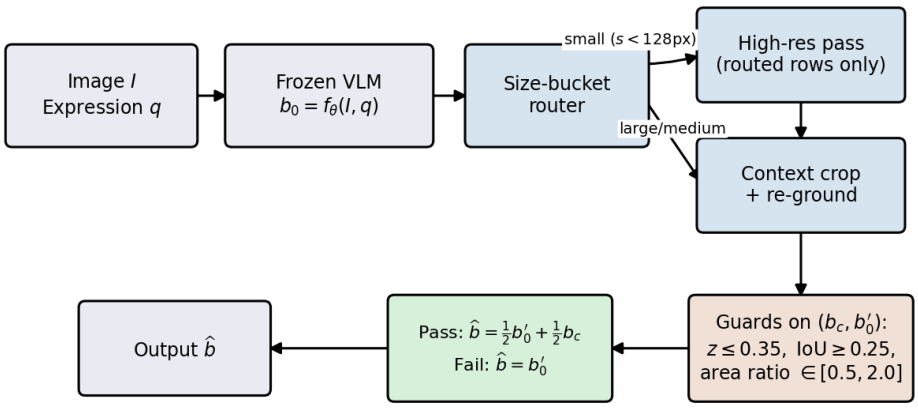


Figure 6: The LFPR pipeline. The router decision and the guard checks are both computed from already-observed model outputs and image geometry, never from the ground-truth box.

Algorithm 1 Label-free precision refinement

Input: image I, expression q, frozen VLM f_θ
$b_0 \leftarrow \mathrm{canon}(f_\theta(I, q))$ {parse, clip, order, validate}
if b_0 is invalid **then**
  return b_0 {retained as a miss, never synthesized}
end if
$s \leftarrow \sqrt{\max(0, x_2 - x_1)\max(0, y_2 - y_1)}$ for b_0
{label-free size-bucket router, decided from b_0 alone}
if $s < 128$ px **then**
  $b_r \leftarrow \mathrm{canon}(f_\theta(I, q \mid \text{min-pixel budget } 4{,}194{,}304))$
  $b_0' \leftarrow b_r$ if valid, otherwise b_0
else
  $b_0' \leftarrow b_0$
end if
$C \leftarrow$ integer-clipped context crop of b_0' with γ=2.0
$b_c \leftarrow \mathrm{canon}(\text{map } f_\theta(C, q) \text{ to image})$
if b_c is valid and $z(b_c, b_0') \le 0.35$ and $\mathrm{IoU}(b_c, b_0') \ge 0.25$ and $A(b_c)/A(b_0') \in [0.5, 2.0]$ **then**
  $\widehat{b} \leftarrow \mathrm{canon}(\frac{1}{2}b_0' + \frac{1}{2}b_c)$
  if $\widehat{b}$ is invalid **then**
    $\widehat{b} \leftarrow b_0'$
  end if
else
  $\widehat{b} \leftarrow b_0'$ {guard failure: keep the pre-crop answer}
end if
log route, parse outcomes, guard values, and final decision
return $\widehat{b}$

4.1 Routing-alone and post-routing marginal decomposition

RefCOCO-family (Table 10) and the Flickr30K merged-box confirmation (Table 11) both report a same-row decomposition that isolates label-free resolution routing's own effect (R-B) from the marginal effect of crop, guard, and fusion on top of it (CRG-R). Table 27 reports the identical decomposition for Ref-L4, computed with zero new inference from the same three already-archived, hash-verified per-row artifacts behind Table 26 (an independent integrity check cross-checks the

Metric	Incumbent	LFPR	Δ (pp)	Two-sided 95% CI
Acc@0.5	88.531	89.725	+1.194	[0.944, 1.439]
Acc@0.75	77.416	80.590	+3.173	[2.768, 3.584]
Acc@0.9	55.788	61.142	+5.354	[4.808, 5.906]
mAcc@0.5:0.95	72.947	76.013	+3.066	[2.818, 3.311]
Mean IoU	80.944	82.648	+1.704	[1.542, 1.868]

Table 26: Matched full-split results on 31,921 expressions from 9,467 images. Intervals use 10,000 image-cluster bootstrap resamples and describe this retrospective comparison. $\mathrm{mAcc}_{0.5:0.95}$ is the manuscript's primary summary, designated retrospectively rather than confirmed by an external pre-registration; Acc@0.5 is the standard benchmark operating point; Acc@0.75/0.9 are secondary diagnostics; Mean IoU is reported alongside as a threshold-free summary. All five are exploratory and unadjusted here because the comparison itself is retrospective, not because of their role in the metric hierarchy.

Contrast	Acc@0.5	Acc@0.75	Acc@0.9	mAcc
R-B (routing alone)	$+0.711$	$+1.920$	$+3.255$	$+1.773$
	$[0.498, 0.932]$	$[1.607, 2.244]$	$[2.862, 3.658]$	$[1.560, 1.994]$
CRG-R (marginal, after routing)	$+0.482$	$+1.253$	$+2.099$	$+1.293$
	$[0.347, 0.626]$	$[0.974, 1.535]$	$[1.669, 2.540]$	$[1.154, 1.431]$
CRG-B (deployed policy)	$+1.194$	$+3.173$	$+5.354$	$+3.066$
	$[0.944, 1.439]$	$[2.768, 3.584]$	$[4.808, 5.906]$	$[2.818, 3.311]$

Table 27: Ref-L4's routing-alone and post-routing marginal contrasts, 31,921 rows, 9,467 images, 10,000-iteration image-cluster bootstrap (95% CI below each point estimate). CRG-B reproduces Table 26 exactly, as it must since both are computed from the same artifacts. Unlike the RefCOCO family (where CRG-R's Acc@0.9 is negative, -1.192pp, and unlike Flickr30K (where crop/guard/fusion adds no accuracy beyond routing at any threshold, §5.4), Ref-L4 is the one evidence tier where crop/guard/fusion has a genuine positive marginal effect beyond routing at every reported threshold, not merely a small Acc@0.5-for-Acc@0.9 trade: CRG-R's Acc@0.9 is $+2.099$pp, not negative. Routing alone still accounts for the majority of the deployed policy's Acc@0.9 effect ($+3.255$ of $+5.354$pp, 61%), so the same qualitative pattern found on the other two tiers – routing does most of the strict-IoU work – still holds directionally here, but Ref-L4 is not simply a case of the two stages offsetting each other.

fused artifact's embedded R box against the standalone R artifact's own prediction on every row before scoring, and withholds any number unless the recomputed B/CRG Acc@0.5 reproduce Table 26 to within 0.01pp – both gates passed here).

4.2 Guard-by-fusion factorial decomposition

CRG crosses two binary choices: whether the guard gates the crop candidate at all, and whether an admitted candidate replaces the incumbent outright or is midpoint-fused with it. Because the incumbent box, the crop candidate, and the guard's own admit/reject decision are already stored per row, all four cells of this 2×2 design are computable offline from the same frozen artifact with no additional inference, at the exact input denominator (31,921 rows, 1 invalid incumbent, 29,523 rows with an admitted candidate).

This factorial resolves a concern that the two-arm CRG-versus-guard-off contrast alone cannot separate rejection from midpoint shrinkage as the source of the accuracy trade: with both binary factors held out, rejection (the guard) is the factor that protects Acc@0.5, and midpoint fusion (independent of the guard) is what gives up strict-IoU accuracy for stability rather than for its own accuracy gain, since guarded replacement's Acc@0.9 exceeds guarded midpoint's by 2.0 points at a comparable Acc@0.5.

Direct paired contrast between the two guarded cells. The two guarded-cell CIs above are each against the incumbent B, which does not by itself establish whether guarded replacement and guarded midpoint (CRG) differ from each other. We therefore computed the direct paired contrast

Cell	Acc@0.5	Acc@0.75	Acc@0.9	mAcc	Mean IoU Δ	IoU W/L/T
Guarded, midpoint (CRG, shipped)	**89.725**	80.590	61.142	76.013	+1.704	19,386 / 10,260 / 2,275
Guarded, replacement	89.399	79.947	**63.143**	76.166	+1.725	17,333 / 12,317 / 2,271
Unguarded, midpoint	88.262	77.739	59.350	73.864	+0.559	19,764 / 11,321 / 836
Unguarded, replacement	87.247	78.140	62.059	74.491	+0.129	17,612 / 13,490 / 819

Table 28: Guard × fusion factorial on the corrected full split (Appendix 4.5; 10,000-iteration image-cluster bootstrap per cell). Mean IoU Δ and IoU W/L/T are against the true pre-resolution incumbent B, cross-checked against the standalone B artifact (Appendix 4.6); an earlier version of this table computed both columns against the post-routing R box while this caption described them as against B, which understated every cell's Mean IoU gain and, for the two unguarded cells, had the wrong sign. The guard, not the fusion rule, is what protects Acc@0.5: both guarded cells beat both unguarded cells at Acc@0.5, and only the unguarded cells fall below the 88.531% incumbent. Midpoint fusion trades some of the guarded arm's Acc@0.9 for stability – guarded replacement alone reaches 63.143% at Acc@0.9 (its 95% CI is $[+3.512, +4.700]$pp over the incumbent, wider than CRG's $[+1.662, +2.545]$pp) at the cost of a slightly lower Acc@0.5 and more harmed rows (12,317 versus CRG's 10,260). CRG is the historically fixed shipped policy – its parameters were set before this factorial was computed, not selected by it; the other three cells are reported for mechanism decomposition, not as alternate deployment candidates.

between exactly these two cells on the same 31,921 rows (10,000 image-cluster bootstrap resamples; zero new model inference, reusing the same incumbent/candidate/guard fields already stored for the table above; the same identity-gate convention used throughout this appendix withholds any number unless the recomputed guarded-replacement and guarded-midpoint scores both reproduce this table's published figures to within 0.01pp – both gates passed here, matching to within 0.001pp). Guarded replacement minus CRG is -0.326 points at Acc@0.5 (95% CI $[-0.447, -0.208]$), -0.642 at Acc@0.75 (CI $[-0.933, -0.343]$), $+2.002$ at Acc@0.9 (CI $[1.600, 2.420]$), and $+0.153$ at mAcc (CI $[0.011, 0.299]$, $p = 0.034$). All four intervals exclude zero: guarded replacement is significantly better than the shipped policy at Acc@0.9 and at mAcc – the manuscript's own declared primary endpoint – and significantly worse at Acc@0.5/0.75. CRG remains the shipped policy because it was the historically fixed choice before this factorial was computed (§5.3), not because it dominates guarded replacement on every metric; it does not, and this direct contrast is reported so that a reader who weights strict-IoU accuracy or the primary endpoint above Acc@0.5 can see exactly where the two policies disagree and by how much.

4.3 IoU TRANSITION MATRIX

Figure 7 bins baseline (pre-LFPR) and treatment (post-LFPR) IoU into $[0, .5)$, $[.5, .75)$, $[.75, .9)$, $[.9, 1]$ and reports the row-normalized transition matrix for Ref-L4 and the Flickr merged-box confirmation, using the same paired per-row predictions underlying Table 9 and Table 11 – no new inference.

4.4 QUALITATIVE CASES

Rows below are selected by a rule fixed before inspection: the three largest per-row IoU gains and three largest losses among admitted-candidate rows, plus the rates of the two out-of-policy paths a reader might expect but that are not selected as favourable cases. Guard rejections (2,394 of 31,921 rows, 7.5%) and parser failures (3 rows, 0.009%) are reported by pool size rather than by example, since neither category is about the fused box's accuracy.

4.5 ERRATUM: THE SILENTLY DISCARDED PIXEL BUDGET

An earlier version of this work reported that the resolution router fired on 2,223 transfer rows without altering a single output, and attributed the null to RefCOCO-family images already exceeding the routed pixel budget. Both parts were wrong. The images are roughly fifteen times *smaller* than that budget, not larger, and the budget was never applied: the override was passed as a processor keyword argument that this checkpoint's fast image processor silently discards, leaving the checkpoint default in force. The routed pass therefore ran at default resolution and reproduced the incumbent byte for

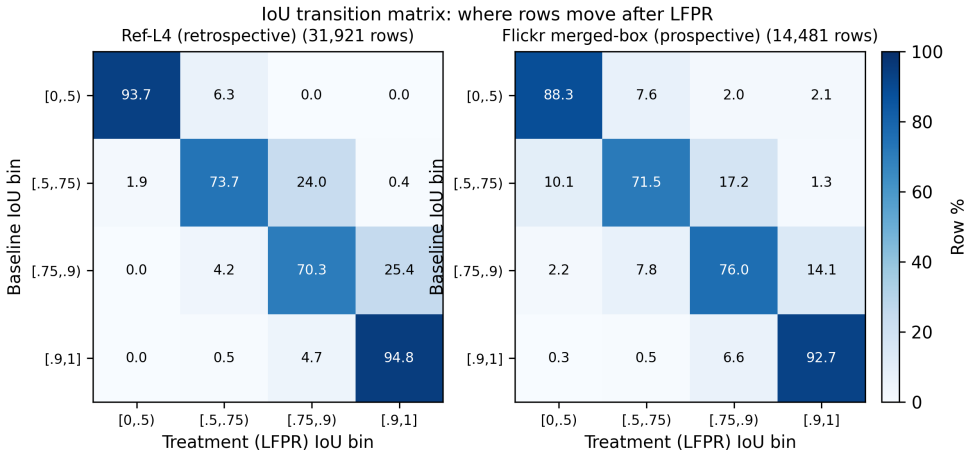


Figure 7: IoU transition matrix. Diagonal cells dominate on both datasets (70–95% of rows stay in their baseline bin), consistent with LFPR changing a minority of predictions (Section 5.7). The two off-diagonal cells of most interest: strict-IoU harm, $[.9, 1] \to [.75, .9)$, affects 4.7% of Ref-L4's near-perfect rows ($894/18{,}847$) and 6.6% of Flickr's ($359/5{,}467$) – a real, nontrivial cost, not a rounding artifact, consistent with the guard-versus-Acc@0.9 trade reported throughout this paper. Genuine boundary refinement, $[.5, .75) \to [.9, 1]$, is comparatively rare on both datasets (0.4% Ref-L4, 1.3% Flickr); most of LFPR's gain instead comes from the larger $[.75, .9) \to [.9, 1]$ cell (25.4% Ref-L4, 14.1% Flickr), i.e. rows that were already close and get pushed over the strict threshold, not rows recovered from a coarse miss.

Category	Expression (truncated)	Incumbent IoU	Fused IoU
Largest improvement	The middle white boat docked between two other vessels, distinguished…	0.561	0.891
Largest improvement	The table, adorned with a white tablecloth, has a man in a gray shirt…	0.657	0.958
Largest improvement	The table covered with a white tablecloth, at which the man wearing a…	0.644	0.933
Largest harm	An individual in a blue and black jersey is maneuvering the foremost b…	0.985	0.574
Largest harm	Within the cluttered area, a building instrument, namely a hammer, is…	0.940	0.530
Largest harm	A cluster of thin green onions in a blue container.	0.954	0.558

Table 29: Fixed-rule qualitative cases from the corrected full split (Appendix 4.5). The largest-harm rows share a pattern: the incumbent was already close to correct (IoU > 0.9), and the crop candidate pulled the box toward a plausible but wrong sub-region (a hand, a nearby object, a neighbour in a cluster) that the guard's geometric checks did not catch. The largest-improvement rows are the mirror case: a coarse incumbent box that the crop resolves to the correct sub-region within a larger scene. Neither pattern is unique to one expression style or object category in this sample.

byte, which is also why all five arms of the cost audit reported an identical 16.49 GiB peak, i.e. model weights only.

Five call sites carried the defect. All are repaired, and the budget is now written into the resolved image-processor configuration and verified, so a request that fails to take effect raises an error instead of silently producing a no-op arm. Every resolution-routed number in this paper was re-measured afterwards. On the transfer split the regenerated arm changes 2,213 of 2,223 flagged rows against zero before; on Ref-L4 it changes 6,591 of 6,641 against 627 rows of ± 1-unit numerical jitter before.

4.6 Erratum: Ref-L4's Mean IoU was reported against R, not B

Two tables reported Ref-L4's Mean IoU delta as though it were the deployed policy's effect against the pre-resolution incumbent B (Table 26; Table 4's Ref-L4 row), and a third (Table 28) re-

System	Calls/row	Gen. tokens	Mean s	P95 s	Rows/s	Peak GiB	Invalid
EGM-4B	1.000	98.9	4.622	5.503	0.216	8.45	0.20%
+ LFPR	1.998	197.9	9.080	10.861	0.110	8.45	0.20%
EGM-8B	1.000	97.8	4.581	5.764	0.218	16.48	0.00%
+ LFPR	2.000	196.5	9.154	10.531	0.109	16.48	0.00%

Table 30: Synchronized 512-row specialist cost audit on physical RTX 3090 GPUs. Generated tokens are summed across calls. Peak memory is maximum allocated memory; invalid rows remain in the denominator. Each row has one timed repeat after a per-row warm-up.

ported both its Mean IoU delta and win/loss/tie columns the same way for all four factorial cells. In every case the reported "incumbent" Mean IoU was actually the post-routing R arm's Mean IoU (81.957%): a fresh, hash-verified audit against the standalone pre-resolution B artifact (run while adding the R-alone/CRG-R decomposition in Table 27 below) finds B's true Mean IoU is 80.944%, not 81.957%. The discrepancy was isolated to the Mean IoU column of three tables rather than a whole mislabeled file. Acc@0.5/0.75/0.9/mAcc were never affected: those columns were always computed from the correct B artifact, which is exactly how the discrepancy was caught – the audit script's identity gate reproduces every other Ref-L4 figure in Table 26 to within 0.001pp, so the artifact is not in question, only which arm's Mean IoU had been transcribed into the "incumbent" column. The corrected tables report the verified pre-resolution incumbent B's true 80.944% Mean IoU and the resulting deltas (Table 26: $+1.704$pp, not $+0.691$pp; Table 28: $+1.704$/$+1.725$/$+0.559$/$+0.129$pp for the four cells, not $+0.691$/$+0.712$/-0.454/-0.884pp – the two unguarded cells' sign flips from apparently harming Mean IoU relative to B to slightly helping it, though both still trail the guarded cells by a wide margin and neither changes the guard's role in protecting Acc@0.5). The separate $80.936\% \rightarrow 81.949\%$ routing-only figures in §5.8 are a distinct, already-flagged superseded diagnostic from before the pixel-budget fix above, not this erratum; that section's canonical cross-reference is corrected to 80.944% accordingly.

5 Detail relocated from the main text

The subsections below were condensed out of the main text to meet the page limit. They are retained verbatim because each is cited from the main text, but none is needed to follow the argument there: they record cost tables, protocol detail, exploratory diagnostics, and closed negative controls.

5.1 Specialist cost cross-check on RTX 3090

The generalist cost audit reported in Appendix 2.9 runs entirely on the same RTX 3090 host used throughout this paper; an earlier draft additionally measured the generalist arms on a single RTX PRO 6000 before the specialist checkpoints were available on that host, but that measurement used different hardware from every other cost number in this paper and has been withdrawn to avoid implying a same-hardware comparison that was never intended. The specialist cross-check below stays on RTX 3090 throughout, matching Appendix 2.9's protocol exactly.

A specialist audit used four physical RTX 3090 GPUs, with each arm split into two non-overlapping shards and one worker per GPU. EGM-4B and EGM-8B were each evaluated alone and with LFPR using batch size one, bfloat16, SDPA attention, greedy decoding, one per-row warm-up, and one synchronized timed repeat. All four arms passed the 128- and 512-row admission gates with complete keys, finite timings, no OOM, and no unclassified runtime failure.

For EGM-4B, LFPR adds 99.0 generated tokens and 4.458 s per paired row, giving $2.001\times$ the generated tokens and $1.965\times$ the latency of the specialist alone. For EGM-8B, the corresponding overhead is 98.7 tokens and 4.574 s, or $2.009\times$ the tokens and $1.999\times$ the latency. Throughput falls by approximately one half in both cases. Sequential refinement does not add model-resident memory: maximum allocated memory remains 8.45 GiB for EGM-4B and 16.48 GiB for EGM-8B. The 128-row measurements show the same pattern (Appendix 2.9).

System	Acc@0.5 Δ	Acc@0.75 Δ	Acc@0.9 Δ	mAcc Δ
EGM-4B	+0.284	+0.455	+1.569	+0.785
	[0.121, 0.449]	[0.147, 0.776]	[1.044, 2.099]	[0.626, 0.940]
EGM-8B	+0.378	+1.230	+6.716	+2.401
	[0.227, 0.537]	[0.915, 1.545]	[6.010, 7.406]	[2.226, 2.577]

Table 31: Accuracy delta from applying LFPR's crop/guard/fusion mechanism to each specialist's own predicted box, full 30,969-row transfer split, 95% paired image-cluster bootstrap intervals below each point estimate. Every interval excludes zero: LFPR improves both specialists on every reported metric, most sharply EGM-8B's comparatively weak Acc@0.9 ($+6.716$ points). Absolute values appear in Table 3 in the main text.

5.1.1 Specialist accuracy with LFPR applied

The cost audit above measures latency and memory only. Table 31 scores accuracy on the same mechanism (LFPR's crop/guard/fusion pipeline applied to each specialist's own predicted box) on the full 30,969-row transfer split, using an already-archived per-row artifact with no new inference: a paired image-cluster bootstrap (10,000 iterations, shared draws across metrics).

These measurements close the marginal cost question for adding LFPR to each specialist. To reduce the hardware mismatch in a cross-system comparison, we also reran the generalist baseline and full LFPR on the exact 128-row prefix on the same RTX 3090 host. The baseline uses 1.000 calls, 17.5 generated tokens, and 1.041 s per row; LFPR uses 2.008 calls, 35.3 tokens, and 1.948 s. Thus LFPR is $1.871\times$ slower and uses $2.011\times$ the generated tokens, with peak allocated memory increasing from 16.49 to 17.58 GiB. On the same rows, EGM-4B and EGM-8B alone are respectively $4.61\times$ and $4.37\times$ slower than the generalist baseline; with LFPR they are $4.75\times$ and $4.68\times$ slower than generalist LFPR, and generate approximately $5.6\times$ as many tokens.

This anchor is deliberately small and contains only one routed row. The generalist uses three timed repeats while EGM uses one, and the EGM shards ran on different physical 3090s of the same SKU. The comparison therefore places the observed inference burden on one host; it does not replace the corrected full-size generalist audit or establish a hardware-independent energy ranking.

5.2 Specialist scaling and firewall detail

This subsection was compressed in the main text's specialist-comparison discussion (§5.3) to save space; both facts below are cited from there. Against the frozen incumbent, the 8B specialist gains $+2.099$ points at Acc@0.5 yet loses 7.168 at Acc@0.9 and more paired rows than it wins on IoU (11,428 against 19,195) – a single permissive-threshold number would rank this model first and its localization quality last. The 4B→8B scaling contrast reported in the main text carries 95% CIs $[0.171, 1.011]$ at Acc@0.5 and $[-8.816, -6.730]$ at Acc@0.9. On the firewall audit (17,978 specialist-training images against 3,982 evaluation images, zero overlap), the Acc@0.5 advantage's benchmark-specificity is concrete: 79.60% against 87.05% for the matched direct prompt on long expressions. A coordinate-granularity check rules out quantization as an explanation (Appendix 2.3).

5.3 Historical selection and stopping rules

The historical workflow advanced a candidate when a one-sided 95% lower bound on Acc@0.5 exceeded zero and treated recomputation on reused rows as motivating evidence. The surviving records document at least fourteen named candidate routes, in addition to resolution, prompt, guard, and fusion choices. These rules reduced compute but were not registered in an external, timestamped archive and did not adjust for selection across the full family. Consequently, historical one-sided lower bounds are retained only to explain which experiments were run; they are not confirmatory tests. The primary descriptive contrast in this manuscript is full LFPR minus the frozen incumbent on the same 31,921 rows. Acc@0.75, Acc@0.9, mAcc, source effects, and all pilot results are exploratory secondary analyses.

System	Params.	Acc@0.5
Qwen3-VL-8B + LFPR (ours)	**8B**	**89.73**
GLM-4.5V (GLM-V Team, 2025)	106B (MoE)	89.5
GLM-4.6V (GLM-V Team, 2025)	–	88.9
Qwen3-VL-8B incumbent (ours)	8B	88.53
PLaMo 2.1-8B-VL (GLM-V Team, 2025)	8B	86.8
GLM-4.1V-9B-Thinking (GLM-V Team, 2025)	9B	86.8
Qwen2.5-VL-7B (Qwen Team, 2025)	7B	80.8

Table 32: Published Ref-L4 test-split Acc@0.5, collected from one leaderboard table for a shared reference point, against our two directly measured rows. Rows are not evaluated under one identical protocol; see text.

5.4 Exploratory source-stratified effects

The source-restricted point differences are positive on both subsets: +0.875 points for COCO and +1.720 for Objects365. One-sided lower bounds are +0.693 and +1.260, respectively, but no interaction interval for the difference between source effects was archived. We therefore do not claim heterogeneity or "source consistency." The method uses no source-specific threshold or prompt. The router flagged 6,641 of 31,921 rows (20.8%) and agreed with the annotation-derived bucket on 89.83% of rows; this is a descriptive diagnostic, not a training or selection target.

5.5 Pilot-to-full-split shift

The pilot's own source-stratified effects (+1.30 points on its Objects365 rows) do not arrive in the same 62.3%/37.7% COCO/Objects365 mixture as the official split, so a naive pilot-to-population extrapolation would overstate the effect if it simply reused the pilot's own composition. Re-weighting the pilot's per-source effects by the official split's composition instead projects the +0.85-point pilot effect to approximately +0.74 points, already below the +0.97-point gap to the working 89.5% comparison before the complete split is scored. The full-split difference turned out larger, not smaller, at +1.194 points once the resolution-routing step (initially executed against the unrouted incumbent by implementation error, corrected and re-run) was actually included. This shift, in both directions across the two corrections this project made to it, illustrates why a pilot projection should not become an absolute ranking claim. Because both datasets belong to the same adapted research program, the full-split interval is descriptive rather than an independent prospective confirmation.

5.6 Comparison to published Ref-L4 results

Table 32 places the measured incumbent and LFPR against published Ref-L4 test-split numbers collected from a single technical report so that every external entry shares one leaderboard's reconciled definitions (GLM-V Team, 2025). The frozen Qwen3-VL-8B incumbent (88.531%) already exceeds two published 7–9B-parameter systems and trails the strongest published entry, a substantially larger mixture-of-experts model, by 0.97 points. LFPR's official result (89.725%) exceeds that entry by 0.225 points without changing the backbone. This table is provided for context and is not the basis of any claim in this paper: the published figures come from a different evaluation protocol (prompt, parser, and in one case an asterisked benchmark condition), so a comparison across rows is cross-protocol and parameter-asymmetric, and Section 6.1 explains why we do not convert Table 32 into a ranking statement.

5.7 Where the headroom comes from: a source- and size-stratified diagnosis

Before describing how the resolution floor and crop context were chosen, we report the diagnostic that motivated targeting them at all. Splitting the 31,921-row test split by the annotation-derived target size gives a strictly monotone accuracy gap: small targets score 78.92% (6,665 rows, 20.9% of the split), medium targets 90.15% (16,311 rows, 51.1%), and large targets 92.73% (8,945 rows, 28.0%) – a 13.81-point gap between the smallest and largest bucket. This association motivated the resolution intervention, but target size is confounded with category, scene density, expression content, and annotation properties and therefore does not identify a resolution deficit by itself.

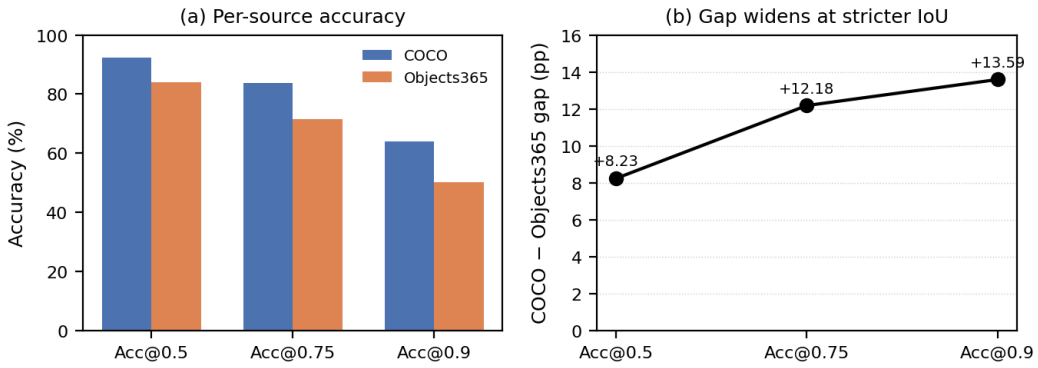


Figure 8: Source-stratified accuracy on the recombined predictions. The COCO–Objects365 gap widens as the IoU threshold tightens (+8.23, +12.18, +13.59 points), an association compatible with several localization and domain-shift explanations.

Source	Acc@0.5	Acc@0.75	Acc@0.9	mAcc
COCO (62.3%)	92.32	83.76	63.87	78.81
Objects365 (37.7%)	84.09	71.58	50.28	67.54
Gap	+8.23	+12.18	+13.59	+11.27

Table 33: Source-stratified accuracy on the recombined predictions (post-resolution-diagnostic). The gap widens at stricter thresholds and is present within every size bucket, showing that target-size composition alone does not account for the source association.

A second, source-based split shows a different and complementary pattern. Table 33 and Figure 8 report Acc@0.5, Acc@0.75, Acc@0.9, and mAcc separately for the COCO and Objects365 rows of the same predictions. The gap is present at every threshold and *widens* as the threshold becomes stricter (+8.23, +12.18, +13.59 points at Acc@0.5, Acc@0.75, and Acc@0.9), a pattern consistent with a localization-precision difference. The gap is also present within every size bucket (small +5.46, medium +6.25, large +7.69 points), so target-size composition alone does not explain it. These descriptive summaries cannot rule out source-specific category mix, crowding, expression complexity, semantic errors, or annotation policy. They motivate, but do not causally validate, the resolution and crop components.

5.8 Selecting and validating the resolution floor

The minimum pixel budget used in the router (Section 3) was chosen with a small staged pilot on the small-target bucket before committing to a full run, following the historical rule fixed at pilot time that a pilot without a positive signal does not earn a full-scale run. On an 800-row small-bucket sample, a moderate floor ($\sim 2.7\times$ the default) gave +0.75 points with a one-sided LCB of -0.99 (not distinguishable from zero); the floor used throughout this paper ($4{,}194{,}304$ pixels, $\sim 10.6\times$ the default) gave +1.875 points with LCB $+0.25$, clearing zero on the pilot sample and motivating the full run.

Those pilots, and the full Ref-L4 run they authorised, were later found to have requested the floor through a processor keyword this checkpoint discards (Section 5.2), so they measured default-resolution work. We therefore re-ran the routed pass on all 6,641 flagged Ref-L4 rows with the override verified in effect. It changes the predicted box on 6,591 of them (99.2%), against 627 rows of ± 1-unit numerical jitter before. Applying the floor to the routed rows and leaving every other row untouched moves the full 31,921-row split from 88.531% to 89.230% at Acc@0.5 ($+0.699$), from 77.410% to 79.321% at Acc@0.75 ($+1.911$), from 55.778% to 59.030% at Acc@0.9 ($+3.252$), and mean IoU from 80.936% to 81.949%. This routing-only measurement was itself later superseded: the crop stage had a second, independent bug at the time (it was cropping the unrouted baseline rather than the routed prediction), so the 89.230% figure and its 80.936% baseline predate that fix. The

canonical full-policy numbers, with both bugs fixed, are Table 26's 88.531%→89.725% at Acc@0.5 and 80.944%→82.648% mean IoU (Appendix 4.6); readers should treat Table 26 as the canonical baseline and the figures in this paragraph as a superseded intermediate diagnostic of the routing stage alone.

The direction of the original claim survives; its magnitude and its mechanism do not. The gain is larger than the withdrawn $+3.30$-point small-bucket figure and, more importantly, it is not confined to the smallest targets: the medium bucket gains $+16.51$ points at Acc@0.9 while the large bucket gains $+1.88$. The earlier conclusion that the floor does not generalize past the small bucket was an artifact of the same defect, since the medium-bucket extension pilot that produced -0.125 points also ran at default resolution and was therefore comparing the incumbent against itself. The router's restriction to the small bucket is consequently a conservative operating choice rather than an empirically forced one, and widening it is a concrete avenue this work leaves open.

5.9 Exploratory resolution checks do not establish a peak

Two prompt variants were piloted against the frozen default prompt. A *tight-box* instruction changed Acc@0.5 by +0.125 points (historical one-sided LCB -1.13), and a *fine-grained* instruction changed it by +0.25 points (LCB -1.00). These are exploratory null results.

The archived resolution summaries reveal a reference inconsistency in the earlier ladder figure, which is therefore removed. On an 800-row small-target sample, the $1{,}048{,}576$- and $4{,}194{,}304$-pixel summaries report default-relative changes of +0.750 and +1.875 points. The $8{,}388{,}608$-pixel summary instead reports +1.125 points relative to the $4{,}194{,}304$ arm on the same 800 rows (82.875% versus 81.750%). A stepwise increment cannot be plotted as a default-relative ordinate or used to conclude that the curve peaked. Until the row-level predictions are rescored against one common reference, the defensible conclusion is only that $4{,}194{,}304$ was the historically selected operating point; remaining resolution headroom is unresolved.

5.10 An unguarded self-correction baseline regresses the headline metric

Section 2 distinguishes LFPR from iterative visual self-correction methods by its single guarded attempt. To make this distinction an empirical rather than only a design point, we ran a matched, zero-shot self-correction control on an 800-row pilot: the same frozen incumbent is shown the original image with its own first box drawn as an overlay and asked to return one replacement box, with no guard and no fusion. This regresses Acc@0.5 by -2.00 points (one-sided LCB -3.14) and Acc@0.75 by -2.125 points (LCB -3.78); only Acc@0.9 improves (+2.25 points, LCB $+0.12$), and mAcc regresses by -1.06 points (LCB -2.14). An ungated second opinion is therefore not a substitute for LFPR's guarded crop: it can occasionally sharpen an already-close box (the Acc@0.9 effect) while making the incumbent's committed answer worse on average, which is the same asymmetry the self-correction mirage literature reports for oracle-free stopping (Tripathy and Krishnan, 2026). LFPR's guards are best understood as an explicit, non-learned stopping rule that keeps this same mechanism from being applied unconditionally.

5.11 Matched negative controls define the claim

The positive full LFPR result is interpretable because nearby alternatives were not silently promoted. A supervised coarse-to-fine adaptation of the incumbent, fine-tuned for 10,000 steps on external Objects365 boxes and evaluated on a reserved, never-previously-read 3,314-row confirmation split through the same frozen evaluation harness, decreased Acc@0.5 by 0.543 points, with LCB -1.264: adapting the weights directly, using the same external data source LFPR's own diagnostics draw on, did not transfer to a gain. A released third-party checkpoint built for parallel box decoding was evaluated zero-shot on the shared pilot with zero runtime errors and decreased the metric by 15.80 points; this checkpoint's own training-image overlap with Ref-L4 was never resolved by its authors, so it was treated as diagnostic-only regardless of its result. SAM3 boundary projection with the same midpoint fusion produced only +0.10 points on a fresh pilot, with LCB -0.20; a retrospective check that applied the same fusion rule to the projection pilot's own pilot rows had suggested +0.30 points, but this signal did not survive a prospective confirmation on a disjoint holdout, illustrating why this paper treats retrospective checks as motivating evidence only, never as confirmation. Full crop

replacement and standalone conservative fusion improved strict-IoU metrics but did not establish the required headline gain on their own. Finally, scaling the incumbent within its own model family to a $4\times$-larger 32B checkpoint, under the identical frozen prompt, parser, and evaluator, also closed negative: +0.26 points at Acc@0.5 with LCB -2.45, a regression at Acc@0.9 (-1.32 points, LCB -4.56), a failing effect on the Objects365 subset (-0.53 points), and a COCO LCB (-2.07) outside the historical -0.5-point band. These controls show that several nearby alternatives failed under their respective pilots; because populations differ and a full same-row factorial table is absent, they do not isolate the causal contribution of routing, guards, or fusion.

6 Additional Metric Definitions

For an image j carrying expressions $i \in \mathcal{I}_j$, the paired image-level bootstrap samples images rather than individual expressions. Let m_j^* be the multiplicity of image j in one bootstrap draw. The treatment effect is

$$\widehat{\Delta}_t^* = \frac{\sum_j m_j^* \sum_{i \in \mathcal{I}_j} [u_t(\widehat{b}_i, y_i) - u_t(b_{0i}, y_i)]}{\sum_j m_j^* |\mathcal{I}_j|}. \tag{1}$$

The reported percentage-point effect is $100\widehat{\Delta}_t$. This is an expression-weighted estimand with image-cluster resampling. mAcc is computed before taking the treatment difference so that each threshold has equal weight. Source effects use the same definition after restricting i to the source subset. No multiplicity correction was used historically; Acc@0.75, Acc@0.9, mAcc, source slices, and pilot comparisons are therefore reported as an exploratory family rather than confirmatory tests.

References

GLM-V Team. GLM-4.5V and GLM-4.1V-Thinking: Towards versatile multimodal reasoning with scalable reinforcement learning. *arXiv preprint arXiv:2507.01006*, 2025.

Marcel Gröpl, Jaewoo Jung, Seungryong Kim, Marc Pollefeys, and Sunghwan Hong. Entropy-gradient grounding: Training-free evidence retrieval in vision-language models. *arXiv preprint arXiv:2604.08456*, 2026.

Zheng Jiang, Yiming Chen, Nan He, Jiahui Chen, Chaoyang Li, Houde Qian, and Lifeng Sun. Test-time scaling over perception: Resolving the grounding paradox in thinking with images. *arXiv preprint arXiv:2604.11025*, 2026.

Qwen Team. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025.

Animesh Tripathy and Aswanth Krishnan. Iterative visual thinking and the self-correction mirage in VLM grounding. *arXiv preprint arXiv:2606.13156*, 2026.