FOUR CELLS ARE NOT THE SAME DOSE: SIZE-CONDITIONED SPATIAL LESIONS IN DETECTION TRANSFORMERS

Anonymous authors
Paper under double-blind review

ABSTRACT

A spatial lesion that removes four feature cells can erase most of a small object's footprint while barely affecting a large one. We test whether this mismatch changes conclusions drawn from decoder cross-attention in frozen DETR and Conditional DETR checkpoints. On a preregistered 4,000-image COCO holdout, the small-minus-large local-quality contrasts for four in-box versus four out-of-box cells are 0.02136 and 0.02325. When we instead remove the same 0.02 clean mean-head attention mass from each eligible query, both 95% image-bootstrap intervals fall inside the prespecified ± 0.005 equivalence region. This low-dose result applies only to eligible GT-matched objects: the four-cell gate excludes 60.3% of candidate small objects. At an exploratory dose of 0.40, size contrasts rise again, so equal mass is not a general explanation of size sensitivity. Prediction-selected detector AP effects are small and differ from the GT-matched local estimand. An exploratory deformable sampling-footprint experiment shows a positive three-seed development-set comparison, but fails its prespecified baseline-feasibility gate. The practical contribution is a dose-aware spatial-ablation protocol that reports eligibility and the detector-level readout alongside query-level effects.

1 INTRODUCTION

The decoder in a detection transformer uses object queries to retrieve spatial information from image features. Its cross-attention weights invite an intuitive interpretation: cells with larger weights carry more evidence for the resulting object. Yet a perturbation of a fixed number of cells need not be a comparable intervention across objects. A small object's box may contain only a few feature-grid cells, whereas a large box may contain hundreds. Removing four cells can therefore differ both in the fraction of the box affected and in the clean attention mass removed. When a lesion study reports stronger small-object effects, these doses must be measured before the effect is attributed to object size or attention ranking.

Existing studies leave this comparison unresolved. EIVE derives instance-level visual explanations from decoder cross-attention (Xiang et al., 2026). Local Gains perturbs decoder query-to-query attention and demonstrates that a selected query's improvement can disagree with fixed-assignment and native set readouts (Zhang & Zhang, 2026). Neither directly supplies a dose-controlled spatial-memory lesion with both query-level and detector-level readouts. This is a narrow methodological gap; the present work does not propose a new detector or seek a higher AP benchmark number.

We intervene on two public, frozen R50 dense-attention detectors and compare a fixed cell count with fixed clean attention mass. Our registered question has a direct answer: a four-cell lesion shows a size interaction, while a low-dose equal-mass lesion meets the prespecified equivalence rule on eligible objects. We then test how this answer depends on dose, grid, layer, and readout. The design exposes both a population boundary (many small objects fail the four-cell gate) and a measurement boundary (local GT-matched effects do not directly determine prediction-selected AP). A separate FFOS extension tests whether a sampling-footprint adjustment transfers to trained sparse attention; its feasibility gate limits that conclusion.

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.

2 METHOD

Frozen detectors and attention intervention. We use public, frozen DETR-R50 and Conditional DETR-R50 checkpoints and their original validation transforms and postprocessors (Carion et al., 2020; Meng et al., 2021). The primary intervention is on the third of six decoder cross-attention layers. At that layer, let a_{qhj} be the *clean* softmax attention weight from query q, head h, to encoder-memory cell j. A set S of cells is selected before intervention. A full lesion adds $-\infty$ to the corresponding attention logit; a fractional lesion $f_j \in (0, 1)$ adds $\log(1 - f_j)$. Attention then renormalizes normally. The accounting dose for a clean anchor query is

$$m(q, S) = \sum_{j \in S} f_j \bar{a}_{qj}, \qquad \bar{a}_{qj} = H^{-1} \sum_{h=1}^{H} a_{qhj}. \tag{1}$$

The same spatial mask can be applied either to the anchor query alone (query-specific) or to every decoder query (all-query); both outputs are recorded separately. The main analysis uses all-query lesions. Equation equation 1 matches the *mean* clean attention mass for the selected anchor; it does not match each head's mass, the number of cells, their positions, or their feature content.

Anchor selection and readouts. We run the model's original Hungarian matcher between clean predictions and non-crowd COCO ground truth (GT). Within each image and official COCO size bin, we retain at most one matched GT instance, chosen by smallest annotation ID. A spatial cell is in-box when its feature-grid center lies inside that GT box. The fixed-count comparison requires at least four in-box and four out-of-box cells. For the selected clean query q and GT g, local quality is $Q(q, g) = p_q(c_g)\,\mathrm{IoU}(b_q, b_g)$, using the detector's raw query-aligned class probability and box. We also record clean-assignment set utility, set utility after rematching, and the number of assignment switches. These latter readouts distinguish a local query effect from how the complete prediction set responds (Zhang & Zhang, 2026).

Three controlled comparisons. First, $k = 4$ randomly selected in-box cells are fully deleted and compared with four randomly selected out-of-box cells. The per-anchor contrast is $D_4 = Q_{\mathrm{outside}} - Q_{\mathrm{inside}}$. Second, we permute in-box cells with a fixed image/annotation-derived seed and delete full cells plus at most one fractional cell until $m(q, S) = 0.02$ with absolute error at most 10^{-4}. Anchors whose total clean in-box mass is below 0.02 are excluded from this comparison. The decrement is $L_{.02} = Q_{\mathrm{clean}} - Q_{\mathrm{inside},.02}$. Third, preregistered sensitivity analyses use $m = 0.05$ and a deletion of approximately 25% of in-box cells with an equal-count out-of-box arm. The frozen primary analysis uses token-selection seed 20270930; a post-freeze change reduced the originally planned three seeds to one before intervention outcomes. Later seed checks are exploratory.

Detector-level AP policy. To avoid GT selection in AP, we choose the clean highest-scoring non-background raw query and its *predicted* box in every holdout image. We then apply all-query four-cell random in-box or out-of-box deletion, an attention-ranked in-box reference, and a whole-grid random control. When either box region has fewer than four available cells, both random arms use the whole grid in that image. We evaluate each perturbed native prediction set with the original COCO postprocessor and official COCO AP evaluator. The AP and GT-anchor local analyses therefore have different policies and do not identify a numerical local-to-AP mediation effect.

3 EXPERIMENTAL DESIGN

Data lock and baselines. We permuted the 5,000 official COCO 2017 validation image IDs using NumPy PCG64 seed 20270925. Positions 0–499 and 500–999 supported earlier pilots; positions 1000–4999 define the 4,000-image primary holdout, untouched before the primary analysis. Its identifier hash is recorded in Appendix A. No weights or hyperparameters were tuned on these 4,000 images. Unperturbed full-val AP was 42.082 for DETR and 40.936 for Conditional DETR, within the preregistered two-AP-point tolerance of their respective official checkpoint descriptions (42.0 and 40.9). The baseline hook had zero numerical error under a zero-dose mask. Both source commits and checkpoint SHA256 values are in the reproducibility manifest.

Table 1: Small-minus-large local-quality contrasts on the frozen 4,000-image holdout. Entries are estimates and 95% whole-image bootstrap intervals; the unit is $p(c_{GT}) \times \mathrm{IoU}$ on $[0, 1]$. Fixed-count and 25% arms subtract the equal-count outside-box control. Only $m = .02$ has a frozen practical-equivalence decision. One token-selection seed is used.

Checkpoint	Intervention	Estimate	95% lower	95% upper
DETR-R50	Four cells	+0.021364	+0.017395	+0.025528
	Equal mass .02	+0.000147	+0.000009	+0.000286
	25% of box cells	+0.002967	+0.001543	+0.004509
Conditional DETR-R50	Four cells	+0.023249	+0.019712	+0.027121
	Equal mass .02	-0.000024	-0.000294	+0.000238
	25% of box cells	+0.003631	+0.001657	+0.005887

Estimands and uncertainty. The fixed-count size interaction is $\Delta_4 = \mathbb{E}[D_4 \mid \mathrm{small}] - \mathbb{E}[D_4 \mid \mathrm{large}]$. The *single primary* frozen question is the equal-mass interaction $\Delta_{.02} = \mathbb{E}[L_{.02} \mid \mathrm{small}] - \mathbb{E}[L_{.02} \mid \mathrm{large}]$, evaluated independently in each detector. The predeclared operational practical-equivalence rule is a 95% image-bootstrap interval wholly inside $[-0.005, 0.005]$ in *both* detectors. This margin was specified in local-quality units without an independently validated perceptual or AP anchor; it therefore defines a protocol decision rather than a universal practical threshold. A wholly outside interval would instead support persistence; an interval crossing a boundary would be inconclusive. We use 5,000 whole-image percentile bootstrap resamples with PCG64 seed 20271001; anchors within an image are not resampled independently. Fixed-count and crowding secondary comparisons use the logged Holm adjustment. Image-bootstrap intervals condition on the one token-selection seed and do not estimate seed-to-seed variation.

Analysis sequence. The spatial-size question followed an unsuccessful pilot attention-ranking hypothesis. The primary protocol was frozen before holdout outcomes, with a pre-outcome amendment reducing three planned token-selection seeds to one. Per-head, high-dose, grid, layer, checkpoint, and AP-uncertainty analyses were added later. They reuse the same images and are exploratory; their nominal intervals do not account for adaptive analysis selection. Appendix A records this sequence.

4 Results and analysis

Fixed count and low equal mass tell different stories. Table 1 reports the frozen size interactions in local-quality units. The four-cell contrast exceeds the preregistered $+0.005$ threshold in both models: $+0.02136$ for DETR and $+0.02325$ for Conditional DETR; both bootstrap intervals are above the margin, with Holm-adjusted boundary $p = 0.0008$ each. At equal mean-head dose $m = .02$, the intervals for $\Delta_{.02}$ are $[0.000009, 0.000286]$ and $[-0.000294, 0.000238]$, respectively. Both satisfy the frozen equivalence rule. The largest absolute interval endpoint is 0.000294, so both intervals also fit inside any symmetric margin strictly wider than 0.000294. This post hoc sensitivity calculation quantifies dependence on the chosen margin; it does not externally validate practical equivalence. This is a statement about the selected, dose-eligible anchors at this low dose, not an absence of size sensitivity generally.

Dose and coverage are unequal under a four-cell lesion. Among 2,116 candidate small anchors, only 839 (39.7%) have at least four in-box cells; 826/836 remain eligible for $m = .02$ in DETR/Conditional. Corresponding large-object $m = .02$ counts are 3,193/3,194. The fixed four in-box cells remove mean clean attention mass 0.3962 for DETR small anchors versus 0.0229 for large anchors (Conditional: 0.4138 versus 0.0247). Thus the same four-cell operation removes roughly 17 times as much clean attention mass for an eligible small object as for an eligible large object in either detector. This observed dose difference explains why a fixed cell count is not a controlled amount of intervention; it does not establish that dose alone causes the local-quality difference. The 25%-coverage sensitivity produces a smaller contrast than fixed count in both models (Table 1), but Conditional DETR's interval reaches above $+0.005$; neither sensitivity identifies

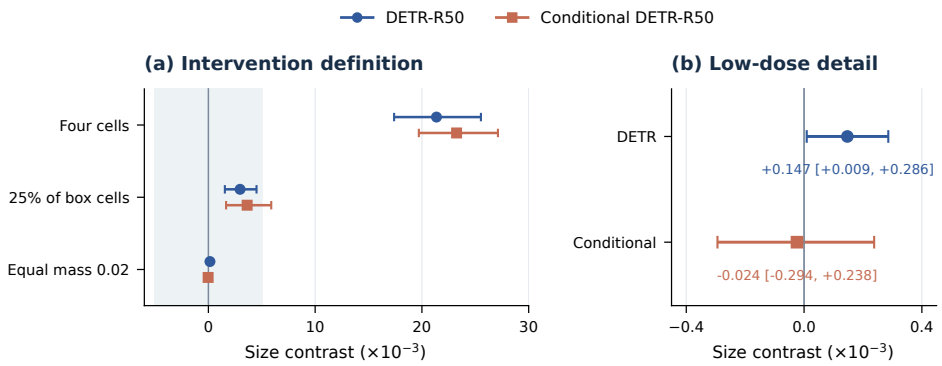

Figure 1: Size interactions and 95% whole-image bootstrap intervals, conditional on one token-selection seed. Panel (a) compares the three operators; the shaded region is the prespecified ± 0.005 margin for the $m = .02$ primary only. Panel (b) expands the low-dose axis and prints its estimates and intervals. Axes use units of 10^{-3}.

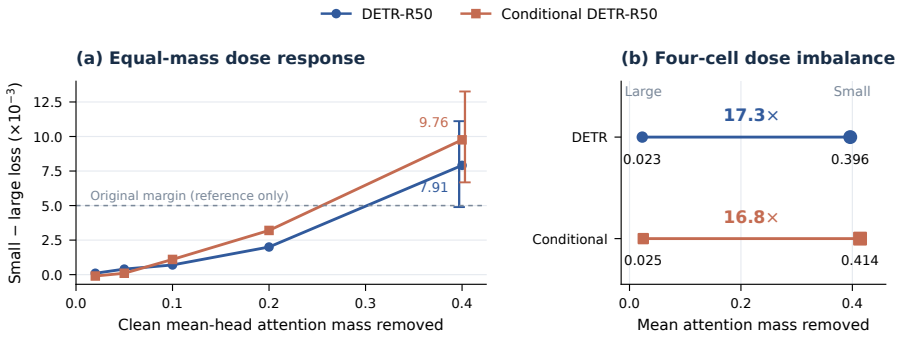

Figure 2: Dose response and intervention imbalance. (a) Exploratory dose sweep on the common $m = .40$-eligible population (DETR: 697 small/3,185 large; Conditional: 735 small/3,192 large). Values below $m = .40$ use the supplied rounded summary; only $m = .40$ has a recoverable 95% image-bootstrap interval in the delivered data and is shown with error bars. The dashed line is a descriptive reference. (b) Mean mass removed by four cells on the original eligible population; paired endpoints show large and small objects.

a unique causal mediator. The m=.05 contrast is $+0.000356$ $[0.000046, 0.000674]$ in DETR and $+0.000210$ $[-0.000453, 0.000946]$ in Conditional DETR.

Head and high-dose checks delimit the conclusion. Matching $m = .02$ only after averaging heads permits unequal head-specific doses: in 78.5% of eligible DETR small anchors and 90.2% of Conditional small anchors, at least one head loses more than 0.05 clean mass. A separately registered, post-freeze per-head operator, with every head losing 0.02, gives small-minus-large contrasts $+0.000082$ $[-0.000152, 0.000306]$ and -0.000066 $[-0.000535, 0.000378]$ on its narrower eligible populations (655 and 549 small anchors). These are exploratory findings and do not change the frozen mean-head test.

After observing the DETR low-dose result, we also registered $m \in \{0.10, 0.20, 0.40\}$. On the common population eligible at all five displayed doses, the reported point estimates increase with dose in both detectors (Figure 2). At $m = .40$ the DETR contrast is $+0.00791$ $[0.00489, 0.01111]$ and the Conditional contrast is $+0.00976$ $[0.00668, 0.01326]$. The DETR interval still overlaps the old $+0.005$ margin, and no confirmatory high-dose threshold test was registered. The descriptive trend directly rules out the broad interpretation that equal attention mass eliminates size-conditioned effects at every dose.

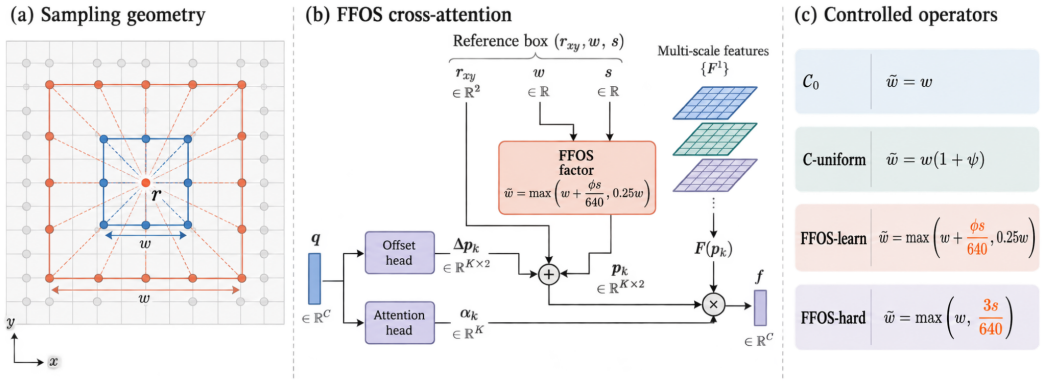

Figure 3: FFOS sampling-footprint schematic supplied by the author. Panels show sampling geometry, the cross-attention pathway, and four controlled operators. In this scalar illustration, w is a reference-box side length and s is the feature stride, not another box coordinate. The implementation applies the factor componentwise and includes the point-scale and offset-scale terms specified in the text.

Exploratory decomposition of the fixed-count contrast. An E3 analysis was specified after the primary outcome and uses only archived local JSONL. On the common $m = .40$-eligible anchors, the fixed-k size gap divided by the $m = .40$ size gap is 3.13 (95% image-bootstrap interval $[2.26, 4.81]$) for DETR and 2.61 ($[2.02, 3.66]$) for Conditional DETR. This comparison holds anchor identity fixed but compares different lesion operators and does not identify a dose-mediated share. Dose-bin analysis reveals a more fundamental limit: in the $[.20, .40)$ and $[.40, 1]$ fixed-k removed-mass bins, DETR has only six and one large anchors, respectively, and Conditional DETR has none. Conversely, the two lowest-mass bins contain only three and four small DETR anchors, and two and seven small Conditional anchors. The sparse overlap prevents a reliable within-dose small-versus-large comparison in these tails. All E3 results are exploratory and are retained in the source run log.

Readout and transfer boundaries. The prediction-selected AP bridge changes AP by -0.051 and -0.023 points for in-box deletion relative to clean inference in DETR and Conditional DETR; the paired bootstrap interval crosses zero for the latter. Exploratory first/final layer, DC5, extra token seeds, and R101 analyses are reported in Appendix B. The R101 comparison changes both checkpoint and decoder layer. All extensions reuse the same holdout and do not constitute independent confirmations. The sparse footprint premise passes a geometry gate, but the sparse grid-lesion implementation fails consistency checks and yields no interpretable causal effect.

4.1 FFOS: A TRAINED FOOTPRINT-OF-SAMPLING OPERATOR

Premise (measured). On the official DEIMv2-S checkpoint, the deformable decoder's sampling footprint is disproportionately small for small objects. Measured forward-only on 1,504 previously inspected VOC2007 development images, the mean sub-cell attention mass inside a GT box for small objects is $F_{sub} = 0.3533$ (95% image-bootstrap interval $[0.3230, 0.3821]$), and the small/large median ratio of effective sampling-cell counts is 0.4065 ($[0.3713, 0.4313]$) against a frozen criterion of ≤ 0.5; the median effective cell count is 26.41 for small and 64.96 for large matched objects. The normalized learned offset is *smaller* for small objects (0.675 vs 0.871), so this checkpoint does not compensate small references by learning larger offsets. A static source check passes a proportionality gate: sampling displacement is exactly $\mathbf{o} \odot n_l^{-1} \odot w_h \odot s_{off}$ with $s_{off} = 0.5$.

Operator. DEIMv2-S computes sampling locations as $\mathbf{p} = \mathbf{r}_{xy} + \mathbf{o} \odot n_l^{-1} \odot w_h \odot s_{off}$, where $\mathbf{r}_{xy} \in \mathbb{R}^2$ is the reference-box center, $\mathbf{o}$ the learned offset, w_h the reference width/height factor, and n_l the per-level point scale. FFOS changes only w_h, per decoder layer (4), per head (8) and per feature level (strides 8/16/32): **C0** leaves w_h unchanged; **C-uniform** uses $w_h(1 + \psi_{\ell hl})$; **FFOS-learn** uses $\max\big(w_h + \phi_{\ell hl}\, s_l/640,\ \ 0.25\, w_h\big)$; **FFOS-hard** uses $\max(w_h, 3s_l/640)$ with no free

Table 2: VisDrone development AP (points; mean ± SD across three seeds). The candidate improvement condition passes but the C0 feasibility gate fails.

Arm	AP	AP50	APs	ΔAPs vs C0
C0	5.823 ± 1.708	7.181 ± 2.097	0.432 ± 0.150	+0.000
C-uniform	5.973 ± 1.090	7.618 ± 1.573	0.565 ± 0.034	+0.133
FFOS-hard	6.500 ± 0.557	8.200 ± 0.737	1.050 ± 0.079	+0.618
FFOS-learn	6.900 ± 0.721	8.600 ± 0.917	1.203 ± 0.068	+0.772

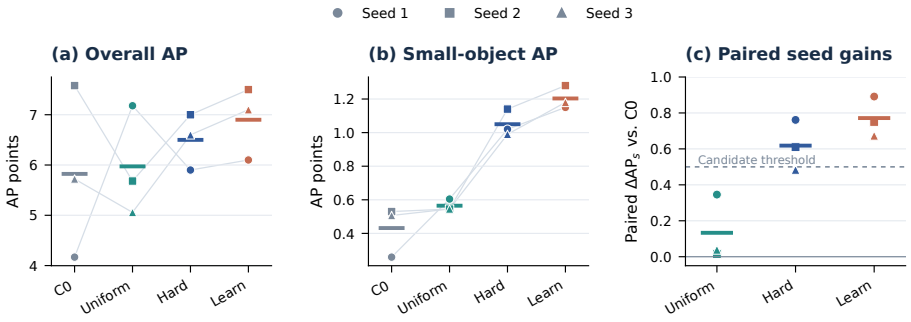

Figure 4: FFOS development results over three paired seeds. Symbols identify seeds; horizontal ticks mark means, not confidence intervals. Panels (a,b) show overall and small-object AP. Panel (c) shows paired APs differences against C0 and the $+0.5$ candidate threshold. The separate baseline-feasibility requirement is C0 APs ≥ 5.0; its observed mean is 0.432.

parameters. ψ and ϕ are zero-initialised, so C-uniform and FFOS-learn equal C0 at step 0. FFOS-hard generally changes the operator immediately wherever its floor is active. The learnable factors are optimised in a separate AdamW group at 10^{-2} with zero weight decay. Only w_h is edited; every learned weight of the network keeps its role.

Protocol. VisDrone2019-DET train (6,471 images) is split deterministically into train-dev (5,500) and dev (971); the official val and test-dev sets stay untouched. Ten-way classification heads are initialised by copying the semantically mapped rows of the official COCO 80-way head (pedestrian, bicycle, car, truck, bus and motor inherit their COCO classifiers; people, van, tricycle and awning-tricycle take the fixed unused rows 76–79, documented as arbitrary). $N = 10{,}500$ steps, batch 2, AdamW (backbone 10^{-5}, detector 10^{-4}, weight decay 10^{-4}), 50-step warm-up, cosine decay to 1%, horizontal flip only, no EMA. Seeds 20271101–20271103 for all four arms; the final-step checkpoint is the only one evaluated. Gates were frozen before any arm result: **G-V2-1** requires the mean C0 dev APs over three seeds to be ≥ 5.0 points, read before any other arm; **G-V2-2** requires the better of FFOS-learn and FFOS-hard to exceed the C0 mean dev APs by ≥ 0.5 point with at least two of three paired per-seed differences positive and to exceed the C-uniform mean, subject to a 0.2-point overall-AP guard.

Development results and feasibility. The measured FFOS summary gives mean dev APs of 0.432 points for C0, 0.565 for C-uniform, 1.050 for FFOS-hard, and 1.203 for FFOS-learn. FFOS-learn improves over C0 by 0.772 APs points with three of three paired seed differences positive, while mean overall AP increases from 5.823 to 6.900. However, the prespecified feasibility gate requires C0 mean dev APs of at least 5.0 points; the measured 0.432 falls far short. The positive candidate comparison therefore cannot be interpreted as a validated detector improvement. These results use the corrected measurement summary supplied by the author. They support a development-set comparison, not a successful prespecified transfer claim.

Untrained footprint probes. A zero-training perturbation of the same functional form, $\max(w_h, \kappa s_l/640)$ with $\kappa = 1, 2$, was applied to the same VOC development set. It raised the small-object median effective cell count from 26.41 to 26.43 ($\kappa = 1$) and 27.82 ($\kappa = 2$) but changed overall AP/APs by $-0.012/+0.031$ and $-0.023/-1.258$ points, respectively. These VOC probes

show no consistent untrained benefit under the tested settings; they do not establish that training is necessary in general or explain the VisDrone comparison. The multiplicative arm C-uniform isolates footprint *scaling* from the additive stride-proportional term used by FFOS-learn, but the failed feasibility gate prevents a confirmatory transfer claim.

5 Related work

DETR introduced set prediction with object queries and Hungarian matching (Carion et al., 2020); Conditional DETR changes the cross-attention query construction to improve training convergence (Meng et al., 2021). Deformable DETR uses sparse multi-scale sampling (Zhu et al., 2021), which falls outside our dense-memory lesion operator. EIVE treats decoder cross-attention as an instance-level explanation and combines attention signals across decoder layers (Xiang et al., 2026). Our intervention operates on encoder-memory links inside a frozen decoder and separates selected-query, all-query, rematched-set and prediction-selected AP readouts; this is a difference in estimand, not a quantitative comparison to EIVE. Local Gains deletes query-to-query attention edges and compares local, fixed-assignment and native set readouts (Zhang & Zhang, 2026). We adopt the need for multiple readouts but perturb spatial cross-attention instead, and explicitly measure detector AP and size-conditioned dose. Parameter-level ablations in detection transformers (Hütten et al., 2025) and query-trajectory counterfactuals for small objects (Shi & Ma, 2026) ask related but different mechanistic questions. Sparse and deformable attention sample a small number of points per query (Zhu et al., 2021); our FFOS operator acts on the reference-box factor of those sampling offsets, which is a degree of freedom of the operator rather than of the data, and we evaluate it by fine-tuning a public detector rather than by lesion. D-RISE uses perturbations to explain object detectors, while the attention-as-explanation debate cautions against identifying weights with causal evidence (Petsiuk et al., 2021; Jain & Wallace, 2019). DropKey also masks attention links, but as a training regularizer for vision-transformer self-attention rather than our inference-time detection cross-attention lesion (Li et al., 2023). SaCo compares salience and perturbation responses across pixel groups in vision transformers (Wu et al., 2024). ODExAI evaluates localization, faithfulness, and cost in object-detector explanations (Nguyen et al., 2025); VX-CODE models collective pixel contributions using Shapley values and interactions (Yamauchi et al., 2026). These works motivate explicit perturbation and evaluation choices, but do not supply this study's fixed-count versus matched-mass comparison. The attention debate also includes tests under which attention can remain informative (Wiegreffe & Pinter, 2019); our lesion results neither certify nor universally reject attention as explanation. Our contribution is the controlled comparison and its population boundary, rather than a new masking primitive.

6 Limitations

Population and intervention. The primary result is conditional on two related frozen R50 detectors, decoder layer three, and GT-matched eligible objects. The four-cell gate excludes 1,277 of 2,116 candidate small anchors (60.3%), including objects poorly resolved by the feature grid. It therefore does not cover the entire COCO small-object population. Matching mean-head mass does not match each head, cell count, feature content, position, or the consequences of renormalization. The high-dose results show that the low-dose conclusion does not extend to all masses.

Inference and generalization. The ± 0.005 margin lacks an external AP or perceptual anchor. Percentile-bootstrap coverage has not been validated by simulation. Image resampling conditions on fixed weights and cell-selection seeds, and later analyses repeatedly inspect the same holdout; their intervals are exploratory. Additional cell seeds do not estimate training variability. R101 changes both checkpoint and layer, while DC5 changes both weights and stride. No isolated architecture, layer, or resolution effect follows from these comparisons.

Readout and transfer. Prediction-selected AP and GT-matched local quality describe different populations. Sparse grid-lesion consistency checks failed, so sparse causal effects are excluded. FFOS has a positive development-set comparison but fails its baseline-feasibility gate; no official validation or test result is established. The corrected FFOS summary is supplied as measured by the

author; the archive's earlier CSV is superseded. Independent run-level verification, a public artifact release, and the outstanding same-layer checkpoint check remain necessary for stronger claims.

7 CONCLUSION

A four-cell spatial-attention lesion yields a size-conditioned local contrast in the frozen R50 comparison, whereas the $m = .02$ equal-mass contrast satisfies the registered equivalence margin. Additional token-selection seeds retain that classification. Exploratory R101 runs satisfy the old interval rule at a different decoder layer, so they do not establish same-layer checkpoint transfer. High-dose and DC5 checks demonstrate dependence on intervention dose and checkpoint setting. Detector-level AP changes are small in the measured R101 bridges, and their paired intervals cross zero. These results support recording dose, eligibility, checkpoint and readout with spatial-attention ablations. FFOS shows a positive development-set comparison but fails its feasibility gate. The supported result is conditional lesion sensitivity, with intervention dose and population selection attached to each claim.

ETHICS STATEMENT

We use public object-detection benchmarks and publicly released checkpoints; no new human-subject data are collected. Detector errors can disproportionately affect small objects, and a local attention effect must not be presented as a validated explanation of an individual decision. Negative and exploratory findings are identified explicitly. No unobserved confirmation result is imputed.

REPRODUCIBILITY STATEMENT

The intervention, anchor eligibility, estimands, seeds, and bootstrap unit are specified in Sections 2–3; supplementary source logs and per-run tables document subsequent analyses. The appendix records the protocol changes and failed checks. The manuscript archive contains the LaTeX source and figure-generation inputs; a public artifact release remains pending.

AI ASSISTANCE DISCLOSURE

ChatGPT/Codex was used for language editing, reference searches, consistency review, LaTeX editing, and generation of plotting code. ImageGen was used to explore figure concepts; those concept images are not included in this version. The authors remain responsible for verifying every result, bibliographic record, figure, and interpretation before submission. AI-assisted figure concepts were redrawn as data-based vector graphics with their plotted values checked against the designated summary.

REFERENCES

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision*, pp. 213–229, 2020. doi: 10.1007/978-3-030-58452-8_13. URL https://arxiv.org/abs/2005.12872.

Nils Hütten, Florian Hölken, Hasan Tercan, and Tobias Meisen. Detection transformers under the knife: A neuroscience-inspired approach to ablations. In *British Machine Vision Conference*, 2025. URL https://arxiv.org/abs/2507.21723.

Sarthak Jain and Byron C. Wallace. Attention is not explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 3543–3556, 2019. doi: 10.18653/v1/N19-1357. URL https://aclanthology.org/N19-1357/.

Bonan Li, Yinhan Hu, Xuecheng Nie, Congying Han, Xiangjian Jiang, Tiande Guo, and Luoqi Liu. Dropkey for vision transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22700–22709, 2023. doi: 10.1109/CVPR52729.2023.02174. URL https://openaccess.thecvf.com/content/CVPR2023/html/Li_DropKey_for_Vision_Transformer_CVPR_2023_paper.html.

Depu Meng, Xiaokang Chen, Zejia Fan, Gang Zeng, Houqiang Li, Yuhui Yuan, Lei Sun, and Jingdong Wang. Conditional detr for fast training convergence. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3651–3660, 2021. URL https://openaccess.thecvf.com/content/ICCV2021/html/Meng_Conditional_DETR_for_Fast_Training_Convergence_ICCV_2021_paper.html.

Loc Phuc Truong Nguyen, Hung Truong Thanh Nguyen, and Hung Cao. Odexai: A comprehensive object detection explainable ai evaluation. *arXiv preprint arXiv:2504.19249*, 2025. URL https://arxiv.org/abs/2504.19249.

Vitali Petsiuk, Rajiv Jain, Varun Manjunatha, Vlad I. Morariu, Ashutosh Mehra, Vicente Ordonez, and Kate Saenko. Black-box explanation of object detectors via saliency maps. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11443–11452, 2021. URL https://openaccess.thecvf.com/content/CVPR2021/html/Petsiuk_Black-Box_Explanation_of_Object_Detectors_via_Saliency_Maps_CVPR_2021_paper.html.

Zhaoning Shi and Bo Ma. Reading decoder trajectories: Training-free counterfactual query-trajectory reliability for small-object detection. *arXiv preprint arXiv:2609.06581*, 2026. URL https://arxiv.org/abs/2609.06581.

Sarah Wiegreffe and Yuval Pinter. Attention is not not explanation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pp. 11–20, 2019. doi: 10.18653/v1/D19-1002. URL https://aclanthology.org/D19-1002/.

Junyi Wu, Weitai Kang, Hao Tang, Yuan Hong, and Yan Yan. On the faithfulness of vision transformer explanations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10936–10945, 2024. URL https://arxiv.org/abs/2404.01415.

Jianlin Xiang, Yanshan Li, and Linhui Dai. Eive: End-to-end instance-specific visual explanations for detection transformers. *arXiv preprint arXiv:2606.01601*, 2026. URL https://arxiv.org/abs/2606.01601.

Toshinori Yamauchi, Hiroshi Kera, and Kazuhiko Kawamoto. Explaining object detectors via collective contribution of pixels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026. URL https://arxiv.org/abs/2412.00666.

Zhen Yang, Zhipeng Wang, Wenshan Xu, Xiuying He, Zhichao Wang, and Zhijian Yin. Region-aware random erasing. In *2019 IEEE 19th International Conference on Communication Technology*, 2019. doi: 10.1109/ICCT46805.2019.8947189. URL https://doi.org/10.1109/ICCT46805.2019.8947189.

Ze Zhang and Yang Zhang. Local gains and fixed-assignment set losses in shared set decoders. *arXiv preprint arXiv:2608.14717*, 2026. URL https://arxiv.org/abs/2608.14717.

Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. *arXiv preprint arXiv:1708.04896*, 2017. URL https://arxiv.org/abs/1708.04896.

Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. In *International Conference on Learning Representations*, 2021. URL https://arxiv.org/abs/2010.04159.

A Reproducibility and registered decisions

The immutable Stage 6 preregistration has SHA256 `c6b33651a590f85abccb3609a6dbaaba2a5c25a3ab1d639b59070c177dbcebe8`. The COCO val annotation has SHA256 `e8c7f7908f1d7278341fae127d0da654f102f11bd7b21d8aeefa635b8c810b6f`. The project manifest records the public source commits, checkpoint SHA256 hashes, completed run IDs, byte-preserved stdout, per-image JSONL hashes and scripts that regenerate tables and figures. Failed MVE hypotheses, cancelled and repaired runs, the one-seed protocol change, late compute-budget checkpoints, the DC5 cost amendment, and post-freeze additions are retained in append-only decision/change logs. The technical repeat on the second four-GPU host produced byte-identical raw JSONL files, as checked against the original raw hashes. No COCO test-dev data were used. The supplied summaries reproduce the reported aggregate tables. The primary per-image records, full bootstrap arrays, and frozen protocol text were not included in this delivery, so this revision cannot independently regenerate the underlying experiments or bootstrap intervals. The figure scripts regenerate all displayed data graphics from bundled summaries.

Table 3: Analysis provenance. All COCO extensions reuse the primary holdout.

Analysis	Timing	Inferential status
Primary layer-3 R50	Protocol fixed before primary outcomes	Primary; one selection seed after amendment
Extra selection seeds	Added after primary outcomes	Cell-sampling sensitivity
Per-head / high dose	Post-primary; high dose outcome-motivated	Exploratory; different eligible populations
DC5 / common anchors	Post-primary; pairing followed aggregate results	Exploratory; weights and stride change
R101	Post-primary; layer 4 rather than 3	Joint checkpoint/layer extension
AP bootstrap	Added after AP point estimates	Conditional image-sampling intervals
FFOS	Separate development-set experiment	Positive comparison; feasibility gate fails

B Exploratory extensions and readout checks

Exploratory first/final-layer sensitivity. Before their outcomes, we fixed the first and final decoder layers as two post-freeze checks on the same holdout, models, seed and intervention arms. Table 4 reports every model–layer outcome. The fixed-four-cell size interaction is positive at both layers but varies substantially: at layer one it is $+0.11506$ for DETR and $+0.04444$ for Conditional DETR, compared with $+0.01399$ and $+0.01349$ at layer six. The $m = .02$ interactions are close to zero at both layers, and all four image-bootstrap intervals lie within the old ± 0.005 margin as a *descriptive reference*. This does not constitute four new preregistered equivalence tests. Fixed-four-cell removed attention mass also varies by layer and size; consequently the layer pattern does not isolate a unique causal source.

Exploratory DC5 grid and common-anchor checks. After observing the original layer-three result, we registered official DC5 DETR-R50 and Conditional DETR-R50 variants, which use a denser encoder feature grid and separately trained official weights. Their full-val clean AP values, 43.234 and 43.730, satisfy the pre-extension two-AP-point reproduction gate. Table 5 reports all four model–grid combinations at the same intervention layer and seed. The fixed-four-cell small-minus-large contrasts are $+0.00670$ and $+0.00649$ in the DC5 variants, compared with $+0.02136$ and $+0.02325$ in the original stride-32 variants. The $m = .02$ contrasts remain near zero, but the original equivalence criterion was not registered as a new DC5 test. DC5 admits 1,796 of 2,116 candidate small anchors (84.9%) under the fixed-four-cell gate, compared with 839 (39.7%) originally. This population change makes an aggregate cross-checkpoint comparison especially difficult to interpret.

After seeing those aggregate DC5 results, we registered a separate common-anchor diagnostic. We paired original and DC5 rows by image ID, GT annotation ID and intervention arm, averaged multiple targets within each image, and reused the same 5,000 whole-image bootstrap draws for both

Table 4: Post-freeze exploratory small-minus-large local-quality contrasts at first/final decoder layers. Each entry is an estimate and 95% whole-image bootstrap interval from 5,000 resamples, conditional on one selection seed. The frozen layer-3 margin is only a reference here.

Layer	Checkpoint	Fixed $k = 4$ [95% CI]	Equal mass $m = .02$ [95% CI]
1	DETR-R50	+0.115060 [+0.098546,+0.132119]	-0.000825 [-0.003462,+0.002367]
1	Conditional DETR-R50	+0.044435 [+0.035670,+0.053798]	-0.000756 [-0.002975,+0.001799]
6	DETR-R50	+0.013993 [+0.011920,+0.016102]	+0.000200 [+0.000121,+0.000281]
6	Conditional DETR-R50	+0.013490 [+0.011945,+0.015104]	+0.000089 [+0.000028,+0.000152]

Table 5: Post-freeze exploratory size interactions for original and official DC5 checkpoints. Each entry is an estimate and 95% whole-image bootstrap interval; the number of eligible small anchors exposes the changed population. DC5 changes both feature stride and pretrained weights.

Checkpoint	Grid	Fixed $k = 4$ [95% CI]	Equal mass $m = .02$ [95% CI]	Eligible small $k = 4$
DETR-R50	stride 32	+0.021364 [+0.017395,+0.025528]	+0.000147 [+0.000009,+0.000286]	839
DETR-R50	DC5 stride 16	+0.006698 [+0.005489,+0.007950]	+0.000153 [+0.000017,+0.000299]	1796
Conditional DETR-R50	stride 32	+0.023249 [+0.019712,+0.027121]	-0.000024 [-0.000294,+0.000238]	839
Conditional DETR-R50	DC5 stride 16	+0.006487 [+0.005467,+0.007657]	+0.000091 [-0.000228,+0.000437]	1796

checkpoints. Table 6 includes every model and both registered measures on their respective eligibility intersections. On the shared fixed-four-cell population, the small-minus-large contrast falls by 0.01877 for DETR and 0.02049 for Conditional DETR. On the shared $m = .02$ population, the interaction differences are near zero and their intervals cross zero. The common-anchor result shows that eligibility changes alone do not numerically account for the observed fixed-count difference; it cannot distinguish feature stride, checkpoint weights, selected-query identity or the changed attention distribution.

Exploratory DC5 token-selection seeds. After the first DC5 result we registered two further cell-selection seeds for the same DC5 checkpoints and audited them independently from the frozen raw records (Table 7). Across the two models and three seeds the fixed-four-cell small-minus-large contrast stays in $+0.00649$ to $+0.00744$, and all six $m = .02$ rows are classified equivalent at fixed mass with intervals inside the original $\pm.005$ margin. This is a cell-sampling stability check on fixed weights and images; it is not a training-seed replication and it does not promote the DC5 extension to confirmatory status.

Local, set, and AP readouts. The raw local experiment records query-specific and all-query effects, fixed-Hungarian utility, rematched utility and assignment switches for every eligible anchor; these are distinct estimands. As an AP bridge, Table 8 reports native COCO AP on the same 4,000 holdout images under the *prediction-selected* policy. Random in-box four-cell deletion changes AP by -0.051 and -0.023 AP points relative to clean inference, while the corresponding out-of-box arm is nearly unchanged at the table's precision. There are 201/219 paired whole-grid fallback images in DETR/Conditional. An outcome-motivated, post-freeze paired whole-image bootstrap reconstructed each original AP exactly and then drew 500 image samples (Table 9). The in-box minus clean 95% interval is $[-0.0915, -0.0110]$ AP points for DETR but $[-0.0641, +0.0281]$ for Conditional DETR. Thus the detector-level AP decrement is small and supported by image-sampling uncertainty in only one of the two checkpoints; it is not a GT-local to AP mediation estimate. Attention-ranked deletion remains a dose-unmatched exploratory reference and is not evidence for attention-ranking faithfulness.

B.1 Post-freeze sensitivity and transfer checks

Token-selection sensitivity in the original checkpoints. After the initial R50 analysis, we fixed two further cell-selection seeds, 20270931 and 20270932, without changing checkpoints, images, layer or dose rule. The independent raw-record audit applies the original 5,000-draw image bootstrap to each seed separately. All four new model–seed $m = .02$ intervals lie within the original ± 0.005 equivalence margin. The three-selection-seed range of the fixed-four-cell small-minus-large interaction is $[0.02041, 0.02136]$ for DETR-R50 and $[0.02207, 0.02368]$ for Conditional DETR-R50.

Table 6: Outcome-motivated, post-freeze common-anchor diagnostic. Original and DC5 columns are small-minus-large local-quality contrasts on annotations eligible in both variants. The difference and 95% interval use paired whole-image bootstraps. Paired counts are annotations, not independent bootstrap units.

Checkpoint	Measure	Original common	DC5 common	DC5−original [95% CI]	Paired small/large
DETR-R50	D_4	+0.02136	+0.00260	-0.01877 [-0.02290,-0.01487]	839/3194
DETR-R50	$L_{.02}$	+0.00015	+0.00008	-0.00007 [-0.00032,+0.00020]	821/3192
Conditional DETR-R50	D_4	+0.02325	+0.00276	-0.02049 [-0.02432,-0.01694]	839/3194
Conditional DETR-R50	$L_{.02}$	-0.00002	+0.00017	+0.00019 [-0.00034,+0.00089]	836/3194

Table 7: Token-selection stability of the DC5 size interactions. The seed-20270930 rows are the originally registered DC5 comparison; the two later seeds were registered after viewing it and are exploratory. Entries are small-minus-large local-quality contrasts with 95% whole-image bootstrap intervals, conditional on one selection seed.

Token seed	Model	Fixed $k = 4$ [95% CI]	Equal mass $m = .02$ [95% CI]
20270930	DETR-R50 DC5	$+0.006698$ [$+0.005489$,$+0.007950$]	$+0.000153$ [$+0.000017$,$+0.000299$]
20270930	Conditional DC5	$+0.006487$ [$+0.005467$,$+0.007657$]	$+0.000091$ [-0.000228,$+0.000437$]
20270931	DETR-R50 DC5	$+0.007442$ [$+0.006204$,$+0.008767$]	$+0.000167$ [$+0.000021$,$+0.000309$]
20270931	Conditional DC5	$+0.006725$ [$+0.005706$,$+0.007826$]	$+0.000191$ [-0.000066,$+0.000541$]
20270932	DETR-R50 DC5	$+0.006573$ [$+0.005367$,$+0.007839$]	$+0.000089$ [-0.000052,$+0.000223$]
20270932	Conditional DC5	$+0.007377$ [$+0.006096$,$+0.008818$]	$+0.000112$ [-0.000101,$+0.000295$]

This measures sensitivity to sampled cells conditional on the same model weights and images; it does not estimate training-seed variation. The audit and raw file hashes are recorded in the experiment log for run `f5cf029f-1574-400a-83a5-e5e6ea0cc2cc`.

Official R101 checkpoints. The full-COCO-val clean AP baselines are 43.5141 and 42.8042 for official DETR-R101 and Conditional DETR-R101, respectively; both pass the predeclared ± 2 AP-point checkpoint gate. We used the same intervention arms, holdout and selection seed but targeted the fourth decoder layer (zero-based index 3), whereas the frozen R50 comparison targeted the third (zero-based index 2). This index mismatch was identified during the final source/log audit and is retained as a protocol-description error; the numerical R101 results remain exploratory. The independent raw audit gives fixed-four-cell small-minus-large contrasts 0.01706 (95% image-bootstrap interval $[0.01403, 0.02025]$) and 0.01498 ($[0.01308, 0.01690]$). At $m = .02$, the contrasts are 0.0000721 ($[-0.0000371, 0.0001785]$) and 0.0000263 ($[-0.0000864, 0.0001524]$); both satisfy the original interval criterion, with at least 821 eligible small-anchor images per model. These R101 tests were added after observing R50 and use the same images, so they are exploratory checkpoint extensions rather than independent confirmations. The two R101 first-layer raw runs are complete; their contrasts are withheld pending an independent audit.

Grid concentration and detector AP. For the R101 matched queries, median mean-head effective grid count rises from 6.38 small to 46.20 large in DETR and from 8.40 to 66.22 in Conditional DETR. This is descriptive concentration, not evidence that cell count causes the lesion effect. In prediction-selected all-query evaluation, random in-box deletion changes AP relative to each R101 clean checkpoint by -0.00649 and -0.03134 AP percentage points, respectively. The independent audit verifies all 4,000 images, 100 detections per arm and checkpoint, and the raw prediction hashes. A separately registered, post-outcome paired-image bootstrap gives inside-minus-clean AP intervals $[-0.03910, +0.02533]$ and $[-0.07429, +0.00690]$ AP percentage points, respectively. Both cross zero; neither supports an AP improvement or statistically resolved loss. GT-selected local and prediction-selected AP populations remain distinct.

Sparse-model boundary. On official DEIMv2-S, the separate forward-only grid-footprint diagnostic passes its frozen small-object gate: mean sub-grid attention mass is 0.4705 and the small/large median cell-effective-number ratio is 0.3207 on the COCO holdout. This establishes a geometric premise only. The subsequent sparse grid-lesion implementation failed repeated clean-versus-lesion consistency and native-output reconstruction checks. We exclude its partial lesion rows from every

Table 8: COCO AP points (AP×100) on the 4,000-image holdout under prediction-selected all-query interventions. One selection seed; all 4,000 image IDs and 100 native postprocessed detections per image enter every arm. The separate paired image-bootstrap intervals are in Table 9.

Checkpoint	Clean	Inside	Outside	Ranked	Whole grid	Δ inside
DETR-R50	42.174	42.123	42.178	42.058	42.175	-0.051
Conditional DETR-R50	40.945	40.922	40.944	40.908	40.930	-0.023

Table 9: Post-freeze, outcome-motivated paired whole-image bootstrap of original prediction-selected AP contrasts. Intervals use 500 draws of the same 4,000 image IDs with replacement; they condition on a single checkpoint and token-selection seed. Original AP was reconstructed to within 10^{-8} before resampling.

Checkpoint	Contrast	Original AP-point Δ	Paired image-bootstrap 95% CI
DETR-R50	Inside−clean	-0.0512	[-0.0915,-0.0110]
DETR-R50	Inside−outside	-0.0550	[-0.0959,-0.0117]
Conditional DETR-R50	Inside−clean	-0.0232	[-0.0641,+0.0281]
Conditional DETR-R50	Inside−outside	-0.0224	[-0.0610,+0.0331]

effect table. D-FINE-S also failed strict checkpoint loading because the available model construction omitted six official encoder projection state keys. A separate FFOS line is reported in Section 4.1.

C A PREREGISTERED SPARSE-ATTENTION TRANSFER GATE THAT DID NOT PASS

After the dense-attention lesion study in the main paper, we considered a separate training-time method, M1, for the public DEIMv2-S sparse-attention detector. M1 would remove exactly $m = 0.20$ of the realized sampling-weight mass per query and head during training, with a matched DropKey-style control. Before any M1 training or detector AP evaluation, we froze a mechanism gate: in the clean checkpoint, last-layer attention concentration $S_2 = \sum_{i=1}^{12} a_i^2$ had to be higher for small than large ground-truth objects by at least 0.020, with a positive lower image-bootstrap bound. This was an operational precondition for the proposed size-aware transfer argument, not a test of whether M1 could ever improve AP under another rationale.

We evaluated all 1,504 images in a fixed, public VOC2007 trainval development split. The original detector outputs were matched to non-difficult ground truth using the original pre-change focal Hungarian cost (class 2, box L1 5, GIoU 2). We measured all eight heads at the final of four decoder layers, used original-image area thresholds of 32^2 and 96^2 pixels, and resampled whole images for 1,000 preregistered bootstrap replicates. The official checkpoint loaded with an exact-key check; its safetensors file omits only two frozen constructor constants, which were restored from the recorded model config before strict loading. The raw run, converted VOC split hashes, frozen protocol, and independent recount script are retained in the repository.

Table 11 shows a positive small-minus-large mean difference, and its bootstrap lower bound is above zero. However, the measured difference is below the frozen 0.020 minimum. The preregistered gate therefore *failed*. No M1 arm was trained, no M1-versus-DropKey AP result exists, and this appendix makes no claim that M1 improves or harms detector performance. The medium-object mean is below both the small and large means; the pattern does not support a simple monotone size-concentration account. The small stratum has only 146 objects and larger dispersion than the other strata. All statements are conditional on one public checkpoint, the fixed VOC split, the original Hungarian matching rule, and the final decoder layer.

The failure blocks the proposed M1 training study under the frozen protocol. It does not invalidate the main paper's dense-attention lesion measurements: they involve different architectures, interventions and estimands. A future study could test another sparse detector or an alternative mechanism only under a new preregistration; lowering this gate after observing the result would not be confirmatory evidence.

Table 10: Token-selection and joint checkpoint/layer extensions on the same COCO holdout. Layer is one-based: R50 uses 3 and R101 uses 4. Each local-quality contrast is the small-minus-large point estimate with its 95% whole-image bootstrap interval. The original R50 seed 20270930 is the only frozen primary comparison; both extra token seeds and the R101 rows are exploratory. The $m = .02$ interval criterion is applied unchanged to each row, without treating these rows as independent confirmations. Values are generated from byte-verified independent audits.

Checkpoint	Layer	Token seed	Fixed $k = 4$	Equal mass $m = .02$	Small n
DETR-R50	3	20270930	+0.021364 [+0.017395,+0.025528]	+0.000147 [+0.000009,+0.000286]	826
DETR-R50	3	20270931	+0.020408 [+0.016594,+0.024566]	+0.000128 [-0.000002,+0.000257]	826
DETR-R50	3	20270932	+0.021041 [+0.017131,+0.025337]	+0.000096 [-0.000048,+0.000240]	826
Conditional-R50	3	20270930	+0.023249 [+0.019712,+0.027121]	-0.000024 [-0.000294,+0.000238]	836
Conditional-R50	3	20270931	+0.023684 [+0.020192,+0.027593]	-0.000095 [-0.000480,+0.000263]	836
Conditional-R50	3	20270932	+0.022072 [+0.018755,+0.025684]	+0.000119 [-0.000135,+0.000375]	836
DETR-R101	4	20270930	+0.017060 [+0.014035,+0.020245]	+0.000072 [-0.000037,+0.000178]	821
Conditional-R101	4	20270930	+0.014978 [+0.013075,+0.016903]	+0.000026 [-0.000086,+0.000152]	838

Table 11: Frozen DEIMv2-S last-layer sampling-weight concentration on GT-matched VOC2007 development objects. Each object contributes eight head values. The table is generated from byte-preserved run stdout. The lower bound is the 2.5th percentile of 1,000 image-cluster bootstrap differences.

GT size	Objects	Mean S_2	SD	Median	P10	P90
Small	146	0.18724	0.10431	0.15687	0.11137	0.28004
Medium	1235	0.13702	0.04148	0.12828	0.10774	0.16899
Large	2414	0.17292	0.04231	0.16572	0.12634	0.22951
Small − large		0.01432				
Bootstrap lower 95%		0.00364				

D A GRID-AWARE ERASING METHOD THAT FAILED ITS PREREGISTERED COMPARISON

After the M1 sparse-attention transfer gate failed, we evaluated a separately approved fallback, M6, on the public DEIMv2-S checkpoint. This training-time augmentation limits the fraction of stride-16 feature-grid centers covered by each erasing rectangle inside every ground-truth box to at most 0.25. Its motivation was the dense-attention study's sensitivity to spatial coverage; that study did not establish that such a training augmentation would improve detection. M6 was frozen before seeing any M6 detector result.

The five prespecified arms were no erasing (C0), ordinary random erasing (C1), erasing with at most 0.25 pixel-area overlap per ground-truth box (C2), a realized-area-matched random-location control (C3), and the grid cap (M6). Every arm started from the same official COCO-pretrained DEIMv2-S weights and trained for 1,000 optimizer steps on the same fixed 512-image VOC2007 training subset. We used one training seed, 20270930, and evaluated only the final checkpoint on the same 1,504-image VOC development split. These are within- recipe comparisons; we did not evaluate VOC test or claim a comparison to a modern detector's published COCO numbers.

C3 and M6 had exactly the same erase event and rectangle area in all 2,000 paired sample/step records. The C3 record's paired M6 rectangle was identical to the rectangle logged in the M6 run, so the area-matched control was implemented as specified. M6 reduced small-object grid exposure relative to C3, but its reduction was below the frozen 0.10 minimum. In fact, C3's realized small-object exposure was itself below 0.10. Because M6 exposure cannot be negative, this mechanism threshold was unattainable for the realized control. That is a preregistration design error, disclosed rather than corrected after seeing results. M6 also did not exceed every control by the required 0.5 APs point: the pixel-overlap cap C2 was stronger, and its gain over C3 was below the threshold. Overall AP stayed within the allowed 0.2-point drop from C0, which does not override either failed primary condition. Thus the preregistered M6 decision is *failure*. A 1,000-replicate paired whole-image bootstrap of M6–C3 APs with fixed PCG64 seed 20271006 gives a 95% percentile interval of $[-1.833, +3.421]$ APs points. Its lower bound is not positive, so the predefined bootstrap condition

Table 12: The frozen five-arm M6 minimum validation experiment. APs and AP are COCO-style points on VOC2007 development. Small-grid exposure is the mean fraction of occupied stride-16 centers erased per small labeled object during training. Erase area is the mean fraction of the training image erased. One training seed per arm; GPU hours cover training only. Values are generated from byte-preserved run logs and augmentation records.

Arm	APs	AP	Small-grid exposure	Erase area	GPU h
C0	34.52	74.35	0.0000	0.0000	0.083
C1	32.41	74.17	0.0609	0.0434	0.090
C2	37.08	74.23	0.0014	0.0385	0.097
C3	34.88	74.23	0.0606	0.0385	0.104
M6	35.25	74.15	0.0014	0.0385	0.104

Table 13: Post-outcome C2 cross-host technical repeat on the same VOC development split. Both runs use the same training seed and augmentation geometry. Values are generated from the two byte-preserved run logs.

Host	APs	Overall AP
First	37.081	74.227
Second	35.958	74.085
Second−first	-1.124	-0.143

also fails. This interval reflects image sampling on one trained checkpoint per arm, not training variation; it cannot reverse the independently failed APs and mechanism criteria.

C2's positive development-set result is scientifically useful because it shows that a simpler overlap constraint can outperform both ordinary erasing and the proposed grid cap in this small recipe. Its strength was observed as part of the control matrix, after the M6 preregistration. We therefore treat it as an exploratory observation, not evidence for a novel proposed method, out-of-sample gain, seed stability, or superiority to current SOTA. The small development stratum includes only 146 non-difficult objects; one seed and a 512-image fine-tuning subset limit any transfer claim. The M1 and M6 failures are distinct: M1 never reached training because its clean-checkpoint mechanism gate failed, whereas M6 was fully trained and evaluated but failed its effect and mechanism gates.

Box-conditioned erasure already appears in the object-aware detection variants of Random Erasing (Zhong et al., 2017). Region-aware erasing also motivates preserving labeled objects during detector training (Yang et al., 2019). The exact C2 overlap threshold was a control choice here; its positive development outcome does not establish a distinct new augmentation method.

Cross-host technical sensitivity. We preregistered a same-seed C2 repeat on the second four-RTX4090 host after seeing the five-arm result, with absolute agreement tolerances of 1.0 APs and 0.5 overall AP point. The public data and weight hashes, 1,000-step config and all 2,000 augmentation records matched; the augmentation JSONL files are byte identical. Table 13 reports both final-checkpoint outcomes. The APs difference exceeds the frozen technical tolerance, while the overall AP difference does not. Final checkpoint and prediction hashes differ. This is an unresolved training reproducibility concern under nominally identical seeds and data, not an independent-seed estimate. The second host has no C0 control run, so its C2 APs cannot be interpreted as a within-host improvement. We retain both runs and do not use the higher value to claim a robust gain. An additional preregistered repeat on the second host with the same GPU, seed, data and geometry gave APs 36.528 versus 35.958 and overall AP 73.869 versus 74.085. Its final checkpoint and prediction hashes also differ. The within-host changes of $+0.571$ APs and -0.215 overall AP exceed the predefined 0.10-point diagnostic tolerance. Thus even same-host training under this recipe is numerically unstable at the precision relevant to the small-object comparison; this diagnostic cannot isolate the between-host cause or replace independent-seed uncertainty.