RECASTING THE DESTROY STEP OF LARGE NEIGHBORHOOD SEARCH AS DENSE SEGMENTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Large neighborhood search improves an incumbent by releasing selected variables and repairing the resulting subproblem under a time limit. FOVEA casts variable selection as dense semantic segmentation on a fixed 128×128 canvas. A small U-Net reads twelve semantic channels and predicts cell scores, which are decoded into a variable set subject to a fixed variable budget. Family-specific layouts connect the search state to the canvas, while one set of network weights serves four problem families. The FP32 network input occupies 786 kB across instance sizes. Our analysis bounds constraint coupling for spatially coherent regions, gives a rank certificate for rounds with zero possible improvement, and bounds the coupling advantage attainable on an expander family. On the tested families and hardware, FOVEA trails the strongest graph encoder at 10^4 variables and overtakes it near 1.5×10^4; the cost model provides an approximate crossover estimate. At larger sizes, where the graph-encoder baselines exceed device memory, FOVEA reduces the primal integral by 12 to 16% relative to the best runnable baselines. On the expander family, FOVEA performs comparably to random selection, with a coupling advantage close to one.

1 INTRODUCTION

Mixed integer programming models routing, scheduling, network design and resource allocation. At large scales, decisions often require a feasible solution before branch-and-bound closes the optimality gap. Large neighborhood search (Shaw, 1998; Pisinger & Ropke, 2007) maintains a feasible incumbent, releases a small subset of its variables, solves the resulting subproblem under a time limit, and accepts improvements. The destroy step chooses one of the $\binom{n}{k}$ subsets of k variables; a solver then repairs the selected neighborhood. Classical rules (Danna et al., 2005; Fischetti & Lodi, 2003; Pisinger & Ropke, 2007) combine hand-designed signals from the instance and incumbent. Learned operators adapt the selection rule to the search state.

Learned destroy operators often use representations indexed by the instance (Song et al., 2020; Wu et al., 2021; Sonnerat et al., 2021; Huang et al., 2023b; Li et al., 2021). For the full-graph encoders evaluated here, a message-passing layer costs $O(\mathrm{nnz}(A))$, and the graph and activations occupy device memory. On the tested families and hardware, these encoders reach the memory limit beyond a few times 10^4 variables. This motivates a selector whose neural input and output dimensions remain fixed as the instance grows.

A spatial grid provides such a representation domain. Visual representations already support combinatorial optimisation through height maps for bin packing, rendered constraint matrices and images of tours (Zhao et al., 2021a; Ling et al., 2021; Steever et al., 2022; Samizadeh & Tong, 2023; Chen et al., 2025; Yin et al., 2026). These applications include constructing solutions with a decoder or graph encoder. For destroy selection, the spatial output directly identifies candidate neighborhoods: a mask selects cells, and a decoder maps their entities to variables for repair.

Two close visual approaches highlight the roles of resolution and action geometry. Chen et al. (2025) fuse image and graph features and increase the total pixel count linearly with instance size to support scale transfer. Yin et al. (2026) select axis-aligned bounding boxes on a rendered tour and repair their interiors exactly. FOVEA uses a fixed-resolution canvas and a mask over its cells, keeping the number of neural outputs independent of instance size. An entity-to-variable map provides the interface to each repair problem. This separates the spatial action from the family-specific definition

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

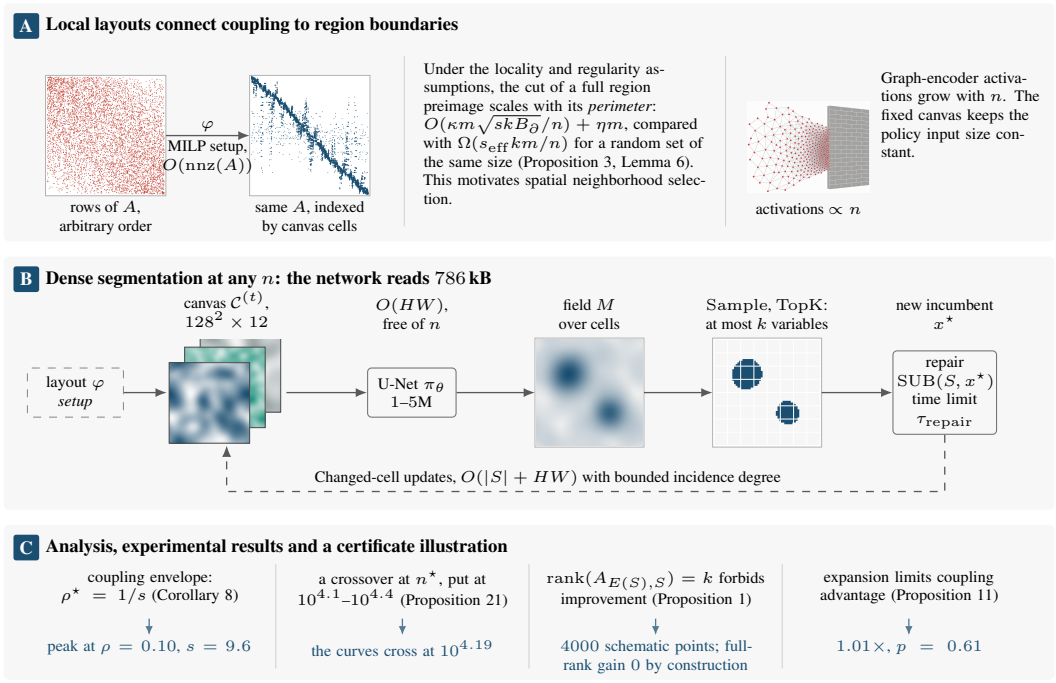

Figure 1: FOVEA overview. **(A)** A precomputed local layout relates the coupling of a full region preimage to its perimeter. **(B)** Each round rasterises the search state into a fixed-size tensor, predicts a field over cells, truncates the variables under the mask to at most k, and solves the resulting repair formulation within the time budget. The one-time layout shown here applies to MILP and routing. Changed cells are updated incrementally (Section 3.5). **(C)** Coupling, crossover, and rank predictions are paired with experimental outcomes and the rank schematic, including the expander control. Panels (A) and (B) use an illustrative instance satisfying the hypotheses of Section 3.6. Panel (C) summarises the text; its rank example uses generated points.

of a released variable. The resolution and mask-shape comparisons (Table 4) measure the associated quality tradeoffs, including scale-dependent resolution and rectangular selections.

FOVEA rasterises the search state onto a fixed canvas and predicts cell scores with a U-Net. Entity mapping and budgeted decoding turn the selected cells into the variables released for repair. Spatial adjacency makes perimeter penalties, morphological post-processing and connected-component analysis available to the selector. Family-specific layouts provide the spatial interface: routing and general mixed integer programs reuse precomputed layouts, while scheduling updates its layout with the incumbent. Incremental rasterisation accounts for variable, constraint and periodic relaxation updates (Section 3.5).

The shared canvas protocol connects one segmentation policy to variable-level repair across routing, mixed integer programming, job shop scheduling and bin packing. Related cross-problem interfaces use heterogeneous graph types, token schemes or attribute vectors (Drakulić et al., 2025; Zhou et al., 2024; Zong et al., 2025); Section D compares these interfaces with the canvas. Our experiments separate the effects of spatial layout and learning: layout accounts for most of the gain on set cover and routing, and learning improves on the same-layout control. The analysis relates region boundaries to constraint coupling, gives a rank certificate for zero-improvement neighborhoods, and identifies coupling limits on expanders. These results motivate the layout, budget and resolution comparisons, including a resolution schedule based on Chen et al. (2025).

2 Preliminaries

We consider mixed integer linear programs $\min\{c^\top x : x \in \mathcal{F}\}$ over $\mathcal{F} = \{x : A_j x = b_j \ (j \in E),\ A_j x \le b_j \ (j \in [m] \setminus E),\ x_i \in \mathbb{Z} \ (i \in \mathcal{I}),\ l \le x \le u\}$, with variable set $[n]$ and constraint set

$[m]$. The equality rows $E \subseteq [m]$ define the rank certificate in Proposition 1; $\theta_E = |E|/m$ is their share. The support of constraint j is $\mathrm{supp}(A_j) = \{i : A_{ji} \neq 0\}$ with size s_j, the mean support is $s = \mathrm{nnz}(A)/m$, and $d = \mathrm{nnz}(A)/n$ is the mean number of constraints per variable. Routing and scheduling enter this framework through their integer formulations and the entity mapping in Section 3.1.

Large neighborhood search maintains a feasible incumbent $x^\star$ and repeats two steps. The *destroy* step chooses $S \subseteq [n]$ with $|S| = k$. The *repair* step solves, under a time limit,

$$\mathrm{SUB}(S, x^\star): \quad \min_x \; c^\top x \quad \text{s.t.} \quad x \in \mathcal{F}, \qquad x_i = x_i^\star \;\; \forall i \notin S, \tag{1}$$

and replaces $x^\star$ if the result improves it. We write $\Delta(S) = c^\top x^\star - \min\{c^\top x : x \in \mathcal{F}(\mathrm{SUB}(S, x^\star))\}$ for the improvement the best possible repair of S would deliver.

Coupling helps characterise the freedom available in a destroy set. It counts the constraints straddling released and fixed variables,

$$\mathrm{cut}(S) = \big|\{\, j \in [m] \;:\; \mathrm{supp}(A_j) \cap S \neq \emptyset \;\wedge\; \mathrm{supp}(A_j) \setminus S \neq \emptyset \,\}\big|. \tag{2}$$

A straddling constraint enters (1) as a constraint on x_S, with the fixed variables contributing a constant. Section 3.6 characterises the resulting freedom through the rank of the equality submatrix: full column rank makes $x^\star$ the unique feasible point and certifies zero possible improvement.

We evaluate solution quality over time using the primal integral. Let $v^\star$ be a reference value and $\gamma(x)$ the primal gap: zero if $c^\top x = v^\star$, one if the two have opposite signs, and $|c^\top x - v^\star| / \max\{|c^\top x|, |v^\star|\}$ otherwise. The case split keeps $\gamma \in [0, 1]$ when incumbent and reference straddle zero, which makes the integral comparable across instances. Following Berthold (2013), the primal gap function $p(t)$ is 1 before the first feasible solution and γ(incumbent at t) afterwards, and the primal integral is $P(T) = \int_0^T p(t)\, dt$. Section C.7 also defines the confined primal integral of Berthold & Csizmadia (2021), an exponentially discounted measure with a finite infinite-horizon limit. Comparing it with P assesses sensitivity to the evaluation horizon. Since $p(0) = 1$, $P(T)$ contains a term equal to the time t_0 of the first feasible solution, which belongs to the shared initialisation rather than to any destroy policy; Figure 4 therefore draws that segment explicitly, and we keep t_0 and $P(T) - t_0$ apart.

We study a fixed absolute variable budget $k \in [10^2, 10^3]$, independent of n. The corresponding destroy fraction is $\rho = k/n$, which decreases as the instance grows. Holding a fraction fixed would instead give $k = \Theta(n)$: for example, $n = 10^6$ and $\rho = 0.1$ release 10^5 variables for repair. Section B.1 describes the computational and coverage properties of the fixed-budget regime. Under the assumptions of Corollary 8, the geometric lower bound on coupling advantage peaks at $\rho^\star = 1/s$. At fixed k, instances with $n > ks$ lie on the lower-ρ side of this peak, where the bound rises with ρ.

3 FOVEA

FOVEA evaluates a segmentation policy on a fixed-size canvas. Band B of Figure 1 shows one round: rasterise the search state, predict cell scores, sample cells, decode variables within the budget, and repair the selected neighborhood. Algorithm 1 gives the search loop. The entity layer defines how each problem family connects its spatial representation to the variables of the repair problem.

3.1 The entity layer

The layout φ maps *variables* to canvas positions. In routing, coordinates belong to customers while the decision variables are edges x_{ij}. An entity layer connects these two representations and defines the variable set used by $\mathrm{cut}(S)$ and Section 3.6. Each family supplies an entity set $\mathcal{E}$, a row-stochastic placement $P \in \mathbb{R}^{|\mathcal{E}| \times HW}$ spreading each entity's unit mass over cells, and an association $\mathrm{ent} : [n] \to 2^{\mathcal{E}}$ recording the entities referenced by each variable. An entity is selected when its mass inside the chosen cells reaches a threshold ϑ; a variable is released when all of its entities are selected:

$$\mathcal{E}_{\mathrm{sel}} = \Big\{ e \in \mathcal{E} : \textstyle\sum_{c \in P_{\mathrm{sel}}} P_{ec} \geq \vartheta \Big\}, \qquad S = \big\{\, i \in [n] : \mathrm{ent}(i) \subseteq \mathcal{E}_{\mathrm{sel}} \,\big\}. \tag{3}$$

Requiring *all* entities selects an edge only when both endpoints are selected. For general programs the entity is the variable and the association is the identity; for routing the entity is the customer and S contains candidate edges among selected customers (Table 5 gives all four instantiations). Routing uses a sparsified candidate edge set to support locality: on a complete graph, degree constraints span the canvas (Remark 5).

3.2 LAYOUT

The layout aims to make constraints spatially local, as required by (A1). Routing uses native coordinates. For general programs, the constraint matrix supplies a two-dimensional sparsity pattern indexed by (constraint, variable). We reorder its rows and columns to concentrate nonzeros near the diagonal. A variable co-occurrence graph adds $\binom{s_j}{2}$ edges for a constraint of support s_j. To retain sparse processing, we run reverse Cuthill-McKee (Cuthill & McKee, 1969) on the bipartite adjacency $\left(\begin{smallmatrix} 0 & A \\ A^\top & 0 \end{smallmatrix}\right)$ at $O(\mathrm{nnz}(A))$. The resulting order induces both permutations and brings variables sharing constraints closer along the column axis. The ablation design includes a one-dimensional canvas, motivated by the dimension-dependent bound in Corollary 13. An index-order control isolates the contribution of spatial layout.

3.3 THE CANVAS

At round t the state is rasterised into $\mathcal{C}^{(t)} \in \mathbb{R}^{H\times W\times K}$ with $H = W = 128$ and $K = 12$. Cell (p, q) owns the variables $V_{pq} = \{i : \varphi(i) \in \mathrm{cell}(p, q)\}$. Each channel aggregates a numerical quantity over V_{pq}, covering instance geometry, current state, dual information or search history (Table 6). The network receives the aggregated channel values as FP32 numbers. Several draw on signals used by classical heuristics: incumbent–relaxation discrepancy in RINS (Danna et al., 2005), reduced cost in variable fixing, and pseudocost in branching. The policy learns how to combine these signals when selecting a region.

The network's tensor is one part of the search state. The rasteriser maintains per-cell aggregates, and a host-side index maps cells back to variables in the original problem. These structures connect the fixed neural representation to variable-level repair: the policy reuses the same input shape across scales, while the index retains the instance-specific information needed for decoding (Section A.8).

For a fixed layout, the twelve channels have three update schedules: four static channels are computed once, four state-dependent channels use incremental updates at $O(|S| + HW)$ for grid and variable state, and four dual channels are refreshed every R_{lp} rounds. Selection time includes the amortised refresh cost $\tau_{\mathrm{lp}}/R_{\mathrm{lp}}$ and the constraint-update cost specified in Section 3.5. Routing and constraint-programming sub-solvers supply no LP relaxation, so their dual channels use the neutral values in Section B and incur no LP refresh cost. Mixed integer programming includes that additional cost.

Across three decades in n, a fixed canvas receives occupancy counts that also span three decades. We use within-instance normalisation and GroupNorm (Wu & He, 2018) throughout to control this input variation. The reported resolution comparisons use normalised inputs and examine the scale dependence identified by Chen et al. (2025). The ablation design also specifies a separate normalisation study.

Count channels enter as $\log(1 + v)$ rescaled by the within-instance 99th percentile of the same transform, which handles both the three orders of magnitude of variation in n and the within-canvas non-uniformity of clustered instances. Mean-valued channels enter as within-instance robust z-scores, using the median and the median absolute deviation so that a few extreme cells do not set the scale. History channels are normalised by the current round index and by the largest improvement seen so far, so that both are bounded and comparable across a run. All statistics are computed per instance.

3.4 POLICY AND TWO-STAGE DECODING

The policy $\pi_\theta : \mathbb{R}^{H\times W\times K} \to [0, 1]^{H\times W}$ is a U-Net (Ronneberger et al., 2015) with four downsampling stages and one to five million parameters. We train it for the twelve semantic channels and target latency below 50 ms on one CPU core to support CPU-only deployment.

Decoding first selects cells and then variables. $\mathrm{Sample}(M)$ perturbs the log-heatmap with Gumbel noise at scale ε and visits cells in decreasing perturbed order until they hold βk variables. At

Algorithm 1 FOVEA search with a fixed variable budget

Require: Instance $(\mathcal{F}, c)$, feasible incumbent $x^\star$, trained policy π_θ, budget k, wall-clock deadline
1: Build the family layout, variable index and initial canvas $\mathcal{C}$.
2: **while** time remains **do**
3: If $x^\star$ changed, update any incumbent-dependent layout; refresh affected canvas state.
4: Refresh available relaxation channels when due (Section 3.3).
5: $M \leftarrow \pi_\theta(\mathcal{C})$; $P_{\text{sel}} \leftarrow \text{Sample}(M)$, including morphological processing.
6: $S \leftarrow \text{BudgetDecode}(P_{\text{sel}}, k)$ using the rule in Section 3.4.
7: If time remains, solve $\text{SUB}(S, x^\star)$ for at most $\min\{\tau_{\text{repair}}, \text{time left}\}$.
8: Accept a returned feasible solution if it improves $c^\top x^\star$.
9: Record the round's outcome and update the search history.
10: **end while**
11: **return** $x^\star$

unit noise scale, this samples without replacement proportional to M (Kool et al., 2019); the scale controls sampling concentration. Morphological closing reduces boundary fragmentation, motivated by Corollary 12. If the positively responding cells are exhausted first, the round uses the available cells and a smaller budget.

The variable stage applies the budget to the variables associated with the sampled cells. The default rule consumes whole cells in decreasing cell score and truncates the final cell by per-variable score. When a relaxation is available, the score is the absolute difference between incumbent and relaxation values; routing and scheduling use family-specific quantities. Cell scores have dimension HW for every n, while the variable stage provides finer selection within cells. Section B.5 describes the default rule and the score-only alternative.

The coupling bounds concern the full region preimage S_R. Budgeted decoding returns a subset, and the full-preimage bound does not automatically apply after truncation. Section B.5 describes the proposed comparison of truncation rules. At large n, when a cell can contain thousands of variables, the optional refinement in Section B.6 re-rasterises the selected region and reapplies π_θ.

3.5 Incremental rasterisation and training

Between rounds only variables in $S^{(t)}$ change value, so updates visit the cells meeting $\varphi(S^{(t)})$. Rebuilding the grid costs $O(HW)$, and redistributing placement mass costs $O(k \max_i |\text{ent}(i)|)$. The tightness channel additionally revisits incident constraints at cost $O(\sum_{j \in E(S^{(t)})} s_j)$. With bounded constraint degrees and support sizes, these updates are independent of n; dense rows increase their cost beyond the grid-and-variable term $O(HW + k)$. The first rasterisation costs $O(n)$ and takes seconds at $n = 10^6$. We report this setup separately and measure incremental update time across instance sizes.

Training imitates a local branching oracle (Fischetti & Lodi, 2003; Sonnerat et al., 2021), uses hindsight relabelling from diverse behaviour policies, and optionally applies REINFORCE (Williams, 1992) to improvement per unit time (Section B). Positive labels cover the variables that a sub-solve actually changes, focusing supervision on effective moves. All families share one π_θ without a family embedding, so held-out families use the same trained parameters and input protocol.

3.6 Analysis

The analysis relates region boundaries to constraint coupling, identifies neighborhoods with zero possible improvement, and bounds coupling advantage on expander instances. Improvement in the objective also depends on the repair problem. The geometric results use locality (A1), density (A2) and region regularity (A3); Section A gives their definitions and the assumptions for each result. The tag [CONDITIONAL] identifies results that use additional distributional, geometric or cost-model assumptions. Table 16 distinguishes associated empirical criteria from the illustrative rank panel.

Coupling depends on the boundary of the selected region. Write $S_R = \varphi^{-1}(R)$ for its variable preimage. At scaling $\delta = c_\delta \sqrt{s/n}$ and $|R| = k/n$, under the locality, density, regularity and budget conditions of Proposition 3, the coupling bound is $O(\kappa m \sqrt{skB_\partial}/n) + \eta m$. Its boundary term grows with the square root of the released mass; ηm accounts for constraints that are nonlocal under the

chosen layout. Combining this bound with the random-selection lower bound in Lemma 6 gives the geometric factor $\kappa^{-1}\min\{\sqrt{sk},\, n/\sqrt{sk}\}$. Under the additional support and defect assumptions of Corollary 8, this lower bound on coupling advantage has peak $\Omega(\sqrt{n}/\kappa)$ at $\rho^\star = 1/s$. For the set cover family with $s \approx 10$, this places the geometric peak near $\rho^\star \approx 0.1$. The random baseline uses a truncated support mean, and the bound applies to S_R before decoder truncation (Section A).

Proposition 1 (Degree-of-freedom certificate)**.** *Let E be the equality rows and $E(S)$ those meeting S. If $\mathrm{rank}(A_{E(S),S}) = |S| = k$, then $x^\star$ is the unique feasible point of $\mathrm{SUB}(S, x^\star)$ and the round cannot improve. Rank deficiency, $\mathrm{rank}(A_{E(S),S}) < k$, is therefore necessary for improvement. Since $\mathrm{rank}(A_{E(S),S}) \le \mathrm{cut}_E(S) + \mathrm{in}_E(S)$, fewer than k incident equality rows guarantee rank deficiency: $\mathrm{cut}_E(S) + \mathrm{in}_E(S) < k$ implies $\mathrm{rank}(A_{E(S),S}) < k$.*

A row count of at least k can coexist with deficient rank and possible improvement; the certificate itself uses $\mathrm{rank}(A_{E(S),S}) = k$. At budgets k in the hundreds, the rank test takes milliseconds. The sufficient condition also motivates an online budget rule: enlarge k until the row count falls below it. In the parameter setting of Section B.8, this occurs above a budget threshold for spatial regions, while random sets fail the condition over the analysed range. The certificate applies to equality rows, including routing degree constraints, packing assignment rows and many MIPLIB instances. The four synthetic families contain only inequalities; there the budget rule uses the tight-inequality heuristic of Remark 10.

Coupling limits on expanders. For a left d-regular lossless-expander incidence graph satisfying the uniform-support and neighborhood-size conditions of Proposition 11, every selection cuts at least $(1-2\epsilon)$ times the expected random coupling. For $\epsilon < 1/2$, the attainable coupling advantage is at most $(1-2\epsilon)^{-1}$, uniformly over admissible S on each instance. Even with $s = O(1)$ and $d = O(1)$, spatial layouts must satisfy both this lower bound and the geometric upper bound wherever their hypotheses overlap. These bounds constrain the achievable locality and density parameters; their precise compatibility is discussed in Section A. Explicit expander constructions (Capalbo et al., 2002) provide the test family. We measure its coupling advantage and primal integral separately. Five further results in Section A guide the design and cost accounting: boundary fragmentation motivates the perimeter loss and sampler post-processing (Corollary 12); dimension-dependent isoperimetry motivates the one-dimensional canvas comparison (Corollary 13); the resolution bound predicts saturation under regularity assumptions, with a fitted constant (Corollary 18); and the critical-size and memory analyses specify the scaling tests (Proposition 21 and Sections A.8 and 4).

4 Experiments

We measure search quality by the primal integral, charging both selection and repair time. Routing and mixed integer programming provide training data, while scheduling and bin packing test zero-shot transfer of the shared weights. A fifth family, set cover with a lossless-expander incidence graph, tests the coupling limit in Proposition 11. The main experiments use a 200 s single-threaded horizon and five seeds. A region-growing control shares the canvas, layout, truncation rule, budget schedule, sub-solver and seeds, isolating the contribution of the network. Section C gives the splits, seventeen baselines, criteria fixed before the runs, statistical tests and separate MIPLIB protocol. It also describes a thirty-five-study ablation design, with results for five studies in Table 4.

Scaling and computational cost. At 10^4 variables, FOVEA has a 2.8% higher primal integral than CL-LNS on set cover and a 5.6% higher integral on auctions (Table 1). At this scale, the 128^2-cell representation trades direct variable scoring for aggregated state information. At 10^5 and 10^6 variables, both graph encoders exceed device memory. FOVEA lowers the primal integral relative to the best runnable baselines by 13.3% and 15.7% on set cover, and 12.4% and 13.5% on auctions. Each large-scale comparison has $p < 0.05$ and concerns methods that fit the stated hardware budget.

Selection time in Figure 2(a) varies by at most 23% across three decades, rising from 31 to 38 ms as incremental updates touch more constraints per released variable. The graph encoder grows approximately in proportion to instance size until its memory limit. Region growing has lower selection cost throughout, by a factor of ten at $n = 10^3$; the learned policy provides the quality improvement examined below. Below $n \approx 5 \times 10^3$, FOVEA's selection step also costs more than the graph encoder's. Overall performance reflects both this cost and neighborhood quality, including the

Table 1: Synthetic mixed integer programs: primal integral in seconds over a 200 s compute-charged horizon. Lower is better; best values are bold and FOVEA is shaded. The last row gives FOVEA's percentage improvement over the best runnable baseline. OOM denotes an encoder exceeding device memory; a dash denotes a sub-solve that returned no result within the horizon. Table 12 separately lists the published results of Huang et al. (2023b), obtained with a 60 min protocol at smaller sizes. The other two families appear in Table 9.

Method	SC			CA		
	10^4	10^5	10^6	10^4	10^5	10^6
SCIP, default	58.4	128.6	171.2	74.9	149.3	184.1
SCIP, aggressive	46.2	112.4	158.9	63.1	134.7	176.3
Random LNS	34.7	71.3	96.4	68.2	103.8	138.9
RINS	31.9	68.9	94.8	59.4	98.2	132.4
Local branching	62.8	—	—	88.3	—	—
LB-RELAX	28.6	64.2	89.1	52.7	92.4	127.8
IL-LNS	25.3	OOM	OOM	47.6	OOM	OOM
CL-LNS	**21.4**	OOM	OOM	**41.2**	OOM	OOM
Region growing	24.1	55.7	78.5	49.8	84.6	115.2
FOVEA	22.0	**48.3**	**66.2**	43.5	**74.1**	**99.6**
gain over best runnable	-2.8	**+13.3**	**+15.7**	-5.6	**+12.4**	**+13.5**

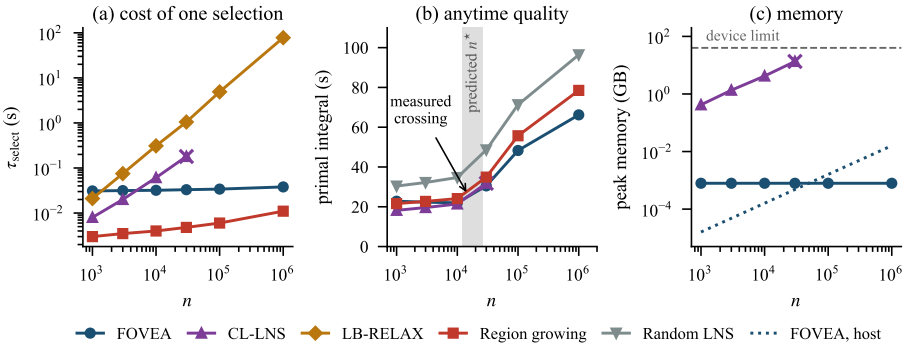

Figure 2: Set cover across three decades of instance size, with budget $k = 500$ fixed independently of n. Half-decade samples below 10^5 bracket the crossing in (b). A cross marks the last size completed within the 40 GB device budget. The band is the $n^\star$ interval from Proposition 21 using the constants in Table 11. The dotted line in (c) gives FOVEA's $O(n)$ host-side state.

CL-LNS advantage at 10^4 in Table 1. The constants in Table 11 place $n^\star$ between $10^{4.1}$ and $10^{4.4}$, using the seed spread of quality ratio $r = 1.03$ and a 15% tolerance on the fitted encoder slope. The curves cross at $10^{4.19}$. The estimate of r comes from compute-matched runs, independently of the primal-integral comparison. Because $n^\star$ equates bounds with different tightness, we interpret the estimate at factor-of-two precision (Remark 25).

Routing: search trajectories and final quality. FOVEA has the lowest reported primal integral among the neighborhood search methods on CVRPLIB X and XL, lowering it relative to region growing by 9.2% and 15.1%. Specialised solvers achieve better final objectives: HGS reaches a 0.19% gap on CVRPLIB X compared with FOVEA's 0.88%, and FOVEA's objective averages 6.1% above FILO2's across the large sets at the same deadline (Table 2). With weights trained at $N \leq 2000$, FOVEA also has a lower aggregate objective ratio to FILO2 than region growing on the large sets extending to 10^6 customers. Only endpoint results are reported at these sizes, and the million-customer tier contains a single instance.

Where the advantage comes from. Spatial layout accounts for most of the gain in Table 3: relative to index order, it reduces the fixed rule's primal integral by 28.7% on set cover and 37.2% on routing. Learning on that layout lowers the integral by a further 8.7% and 9.2% at $p < 0.01$. Under index

Table 2: Routing. CVRPLIB X and XL report primal integral and gap to the best known value. Large sets, from AGS through Lazio, use the objective ratio to FILO2 at equal wall clock, consistently across the column because FILO2 supplies the reference solutions at the largest sizes. HGS, LKH-3 and FILO2 are re-run under the same 200 s limit. This truncates the first two relative to their published protocols in Table 14; FILO2's short configuration completes its iteration budget and reproduces its published value. For solvers whose incumbent traces were not recorded, we report only endpoint gaps or objective ratios. A dash in the ratio column denotes no returned result on the million-customer instance within the deadline. Lower is better.

Method	X, 10^2–10^3		XL, 10^3–10^4		AGS through Lazio, to 10^6
	P	gap	P	gap	ratio
Random LNS	18.7	2.41	31.4	4.82	1.212
ALNS	13.2	1.63	24.6	3.71	1.157
L2D, k=5	10.4	1.18	23.8	3.16	—
L2D, k=10	9.6	1.04	21.3	2.88	—
LKH-3	—	0.51	—	1.94	—
HGS	—	**0.19**	—	0.71	—
FILO2, short	—	0.34	—	**0.62**	1.000
Region growing	9.8	1.12	17.9	2.35	1.094
FOVEA	**8.9**	0.88	**15.2**	2.02	**1.061**

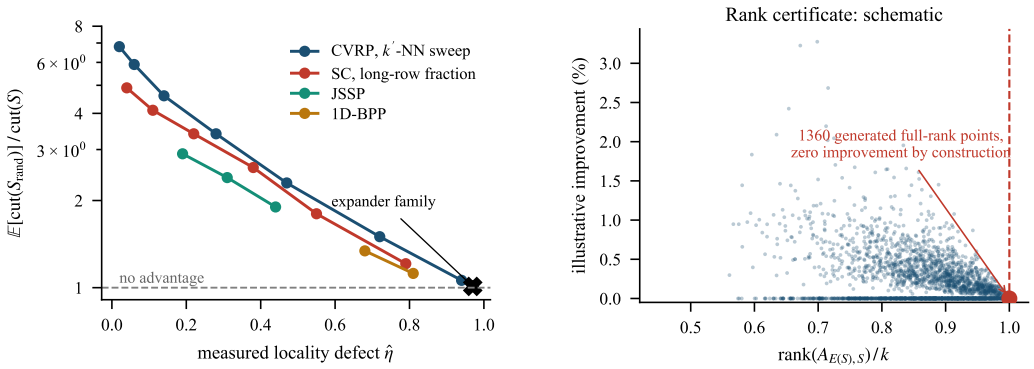

Figure 3: Left: locality and coupling from the within-family sweeps associated with Proposition 3. Routing varies the candidate-edge cutoff k'; set cover varies the fraction of long rows. These sweeps hold the problem family fixed. The cross marks the expander result, whose coupling advantage is limited by Proposition 11. Right: a schematic of Proposition 1 containing 4000 generated points. Of these, 1360 satisfy $\mathrm{rank}(A_{E(S),S}) = k$ and are assigned zero improvement by construction. The panel illustrates the certificate and provides no independent empirical test.

order, the corresponding reductions are 2.7% and 3.2%, both statistically insignificant. This contrast indicates that useful spatial adjacency supports the learned policy.

On scheduling, the shared weights trained on routing and mixed integer programming lower the primal integral relative to random selection by 7.5% at $p = 0.012$. End-to-end CP-SAT retains better performance. On bin packing, FOVEA's primal integral is 1.8% higher than random selection, with $p = 0.44$. Its coupling advantage is 1.1 to 1.3, showing that the relation between coupling and trajectory quality depends on the repair problem. On expanders, the coupling advantage is 1.01, compared with 3.4 on ordinary set cover at the same density. Proposition 11 bounds this advantage uniformly over admissible selections on each instance.

Resolution. The resolution ablation shows little change in primal integral from $g = 128$ to $g = 256$ and an increase at 512 (Table 4). The larger output dimension under a fixed supervision budget is one possible explanation for this pattern. On set cover at $n = 10^5$, the $\sqrt{n}$ resolution schedule of Chen et al. (2025) with normalised inputs has a 3.7% higher primal integral than fixed $g = 128$.

Table 3: Left: a factorial comparison of destroy rule and layout, varying one factor between adjacent cells. Right: zero-shot transfer of the shared weights and the expander family used to test Proposition 11. Values are primal integrals, except for the final column, which gives coupling advantage over random selection. The three rules have similar coupling on expanders ($p = 0.61$).

Method	SC, 10^4		CVRP X	
	spatial	index	spatial	index
Region grow.	24.1	33.8	9.8	15.6
FOVEA	**22.0**	**32.9**	**8.9**	**15.1**

Method	Zero-shot P		Expander SC	
	JSSP	BPP	P	gain
Random LNS	42.6	**27.3**	51.2	1.00
Region grow.	40.1	27.9	50.8	1.02
FOVEA	**39.4**	27.8	**50.6**	1.01

Table 4: Five ablation studies on set cover (SC, $n = 10^5$) and CVRP XL. Values are primal integrals (lower is better). Deployed settings are shaded; best values within each study are bold.

Study	Arm	SC, 10^5	CVRP XL
		P	P
A2, Canvas resolution	$g = 32$	61.4	19.8
	$g = 64$	52.7	16.4
	$g = 128$	48.3	15.2
	$g = 256$	**47.9**	**15.0**
	$g = 512$	48.8	15.4
A7, Fixed against scaled resolution	fixed $g = 128$	**48.3**	**15.2**
	$g \propto \sqrt{n}$	50.1	15.6
B3, Action space	free-form mask	**47.8**	**15.1**
	connected	48.3	15.2
	one rectangle	56.2	17.9
B7, Budget schedule	fixed k	48.3	15.2
	adaptive growth	46.9	14.8
	rank rule	**45.2**	**14.6**
D5, Shared against per-family weights	one shared set	48.3	15.2
	per-family	**46.4**	**14.3**

Action space, budget and shared weights. Free-form masks have slightly lower primal integrals than connected masks on both datasets. On set cover at $n = 10^5$, a single rectangle has a 16.4% higher integral than a connected mask. In this comparison, allowing irregular boundaries makes a larger difference than relaxing connectivity. Shared weights have a 4.1% higher integral than per-family weights, quantifying the quality cost of using one parameter set across families. Those shared parameters also serve the held-out scheduling and packing families, connecting this quality comparison to the transfer results in Table 3. The main scaling comparisons use a fixed budget; in the separate set-cover ablation, the online rule motivated by Proposition 1, implemented with the tight-inequality heuristic for pure covering programs, lowers the integral by 6.4% relative to fixed k. CVRP XL shows the same directions of change.

The rank certificate is established analytically. Figure 3 illustrates it with generated points whose full-rank cases have zero improvement by construction; no empirical evaluation using per-round rank and improvement logs is reported.

5 Conclusion

FOVEA represents destroy selection as dense segmentation on a fixed canvas, connecting a shared policy to variable-level repair across problem families. On set cover and routing, spatial layout supplies most of the observed gain, and learning improves on the same-layout control. Comparisons that include selection and repair time show a scale-dependent crossover with the evaluated graph encoders. The expander analysis and experiments clarify how constraint locality affects this design. A fixed spatial interface therefore offers a practical way to organise neighborhood selection, with benefits that depend on the layout and repair problem.

REFERENCES

Luca Accorsi and Daniele Vigo. Routing one million customers in a handful of minutes. *Computers & Operations Research*, 164:106562, 2024.

Florian Arnold, Michel Gendreau, and Kenneth Sörensen. Efficiently solving very large-scale routing problems. *Computers & operations research*, 107:32–42, 2019.

Timo Berthold. Measuring the impact of primal heuristics. *Operations research letters*, 41(6): 611–614, 2013.

Timo Berthold and Zsolt Csizmadia. The confined primal integral: a measure to benchmark heuristic minlp solvers against global minlp solvers. *Mathematical Programming*, 188(2):523–537, 2021.

Federico Berto, Chuanbo Hua, Nayeli Gast Zepeda, André Hottung, Niels Wouda, Leon Lan, Junyoung Park, Kevin Tierney, and Jinkyoo Park. Routefinder: Towards foundation models for vehicle routing problems. *arXiv preprint arXiv:2406.15007*, 2024.

Ksenia Bestuzheva, Mathieu Besançon, Wei-Kun Chen, Antonia Chmiela, Tim Donkiewicz, Jasper Van Doornmalen, Leon Eifler, Oliver Gaul, Gerald Gamrath, Ambros Gleixner, et al. Enabling research through the scip optimization suite 8.0. *ACM Transactions on Mathematical Software*, 49 (2):1–21, 2023.

Matthew X Burns, Qingyuan Hou, and Michael C Huang. Provable accuracy bounds for hybrid dynamical optimization and sampling. *arXiv preprint arXiv:2410.06397*, 2024.

Michael Capalbo, Omer Reingold, Salil Vadhan, and Avi Wigderson. Randomness conductors and constant-degree lossless expanders. In *Proceedings of the thiry-fourth annual ACM symposium on Theory of computing*, pp. 659–668, 2002.

Jinbiao Chen, Jiahai Wang, Zhiguang Cao, and Yaoxin Wu. Neural multi-objective combinatorial optimization via graph-image multimodal fusion. In *International Conference on Learning Representations*, volume 2025, pp. 33964–33983, 2025.

Elizabeth Cuthill and James McKee. Reducing the bandwidth of sparse symmetric matrices. In *Proceedings of the 1969 24th national conference*, pp. 157–172, 1969.

Giacomo Da Col and Erich C Teppan. Industrial-size job shop scheduling with constraint programming. *Operations Research Perspectives*, 9:100249, 2022.

Emilie Danna, Edward Rothberg, and Claude Le Pape. Exploring relaxation induced neighborhoods to improve mip solutions. *Mathematical Programming*, 102(1):71–90, 2005.

Maxence Delorme, Manuel Iori, and Silvano Martello. Bpplib: a library for bin packing and cutting stock problems. *Optimization Letters*, 12(2):235–250, 2018.

Darko Drakulić, Sofia Michel, and Jean-Marc Andreoli. Goal: A generalist combinatorial optimization agent learner. In *International Conference on Learning Representations*, volume 2025, pp. 52465–52488, 2025.

Matteo Fischetti and Andrea Lodi. Local branching. *Mathematical programming*, 98(1):23–47, 2003.

Maxime Gasse, Didier Chételat, Nicola Ferroni, Laurent Charlin, and Andrea Lodi. Exact combinatorial optimization with graph convolutional neural networks. *Advances in neural information processing systems*, 32, 2019.

Ambros Gleixner, Gregor Hendel, Gerald Gamrath, Tobias Achterberg, Michael Bastubbe, Timo Berthold, Philipp Christophel, Kati Jarck, Thorsten Koch, Jeff Linderoth, et al. Miplib 2017: data-driven compilation of the 6th mixed-integer programming library. *Mathematical Programming Computation*, 13(3):443–490, 2021.

Nir Goren, Shai Yehezkel, Omer Dahary, Andrey Voynov, Or Patashnik, and Daniel Cohen-Or. Visual diffusion models are geometric solvers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 43187–43196, 2026.

Keld Helsgaun. An extension of the lin-kernighan-helsgaun tsp solver for constrained traveling salesman and vehicle routing problems. *Roskilde: Roskilde University*, 12(966-980):1, 2017.

Taoan Huang, Aaron Ferber, Yuandong Tian, Bistra Dilkina, and Benoit Steiner. Local branching relaxation heuristics for integer linear programs. In *International Conference on Integration of Constraint Programming, Artificial Intelligence, and Operations Research*, pp. 96–113. Springer, 2023a.

Taoan Huang, Aaron M Ferber, Yuandong Tian, Bistra Dilkina, and Benoit Steiner. Searching large neighborhoods for integer linear programs with contrastive learning. In *International conference on machine learning*, pp. 13869–13890. PMLR, 2023b.

Richard M Karp. Probabilistic analysis of partitioning algorithms for the traveling-salesman problem in the plane. *Mathematics of operations research*, 2(3):209–224, 1977.

Wouter Kool, Herke Van Hoof, and Max Welling. Attention, learn to solve routing problems! *arXiv preprint arXiv:1803.08475*, 2018.

Wouter Kool, Herke Van Hoof, and Max Welling. Stochastic beams and where to find them: The gumbel-top-k trick for sampling sequences without replacement. In *International conference on machine learning*, pp. 3499–3508. PMLR, 2019.

Yeong-Dae Kwon, Jinho Choo, Byoungjip Kim, Iljoo Yoon, Youngjune Gwon, and Seungjai Min. Pomo: Policy optimization with multiple optima for reinforcement learning. *Advances in neural information processing systems*, 33:21188–21198, 2020.

Kevin Leyton-Brown, Mark Pearson, and Yoav Shoham. Towards a universal test suite for combinatorial auction algorithms. In *Proceedings of the 2nd ACM conference on Electronic commerce*, pp. 66–76, 2000.

Sirui Li, Zhongxia Yan, and Cathy Wu. Learning to delegate for large-scale vehicle routing. *Advances in neural information processing systems*, 34:26198–26211, 2021.

Zhengxuan Ling, Xinyu Tao, Yu Zhang, and Xi Chen. Solving optimization problems through fully convolutional networks: An application to the traveling salesman problem. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 51(12):7475–7485, 2020.

Zhengxuan Ling, Rui Liu, Yu Zhang, and Xi Chen. Can deep learning solve parametric mathematical programming? an application to 0–1 linear programming through image representation. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 52(9):5656–5667, 2021.

Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3431–3440, 2015.

Fu Luo, Xi Lin, Fei Liu, Qingfu Zhang, and Zhenkun Wang. Neural combinatorial optimization with heavy decoder: Toward large scale generalization. *Advances in Neural Information Processing Systems*, 36:8845–8864, 2023.

Wenbin Ouyang, Sirui Li, Yining Ma, and Cathy Wu. Learning to segment for vehicle routing problems. In *International Conference on Learning Representations*, volume 2026, pp. 149846–149882, 2026.

Laurent Perron, Frédéric Didier, and Steven Gay. The cp-sat-lp solver (invited talk). In *29th International Conference on Principles and Practice of Constraint Programming (CP 2023)*, pp. 3–1. Schloss Dagstuhl–Leibniz-Zentrum für Informatik, 2023.

David Pisinger and Stefan Ropke. A general heuristic for vehicle routing problems. *Computers & operations research*, 34(8):2403–2435, 2007.

Eduardo Queiroga, Rafael Martinelli, Anand Subramanian, Eduardo Uchoa, and Thibaut Vidal. The xl instances and the cvrplib best known solution challenge. *arXiv preprint arXiv:2601.11467*, 2026.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241. Springer, 2015.

Mina Samizadeh and Guangmo Tong. Vn-solver: Vision-based neural solver for combinatorial optimization over graphs. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pp. 4269–4273, 2023.

Paul Shaw. Using constraint programming and local search methods to solve vehicle routing problems. In *International conference on principles and practice of constraint programming*, pp. 417–431. Springer, 1998.

Jialin Song, Yisong Yue, Bistra Dilkina, et al. A general large neighborhood search framework for solving integer linear programs. *Advances in Neural Information Processing Systems*, 33: 20012–20023, 2020.

Nicolas Sonnerat, Pengming Wang, Ira Ktena, Sergey Bartunov, and Vinod Nair. Learning a large neighborhood search algorithm for mixed integer programs. *arXiv preprint arXiv:2107.10201*, 2021.

Zachary Steever, Chase Murray, Junsong Yuan, Mark Karwan, and Marco Lübbecke. An image-based approach to detecting structural similarity among mixed integer programs. *INFORMS Journal on Computing*, 34(4):1849–1870, 2022.

Eric Taillard. Benchmarks for basic scheduling problems. *european journal of operational research*, 64(2):278–285, 1993.

Eduardo Uchoa, Diego Pecin, Artur Pessoa, Marcus Poggi, Thibaut Vidal, and Anand Subramanian. New benchmark instances for the capacitated vehicle routing problem. *European Journal of Operational Research*, 257(3):845–858, 2017.

Thibaut Vidal. Hybrid genetic search for the cvrp: Open-source implementation and swap* neighborhood. *Computers & Operations Research*, 140:105643, 2022.

Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3):229–256, 1992.

Yaoxin Wu, Wen Song, Zhiguang Cao, and Jie Zhang. Learning large neighborhood search policy for integer programming. *Advances in Neural Information Processing Systems*, 34:30075–30087, 2021.

Yuxin Wu and Kaiming He. Group normalization. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 3–19, 2018.

Yudong W Xu, Wenhao Li, Scott Sanner, and Elias B Khalil. Large neighborhood search meets iterative neural constraint heuristics. In *International Conference on the Integration of Constraint Programming, Artificial Intelligence, and Operations Research*, pp. 631–648. Springer, 2026.

Zhuoli Yin, Yi Ding, Reem Khir, and Hua Cai. Vitsp: A vision language models guided framework for solving large-scale traveling salesman problems. In *International Conference on Learning Representations*, volume 2026, pp. 132572–132599, 2026.

Hao Yuan, Changwen Zhang, Yong Sun, Liming Gong, Junchi Yan, et al. Btbs-lns: Binarized-tightening, branch and search on learning lns policies for mip. In *International Conference on Learning Representations*, volume 2025, pp. 50682–50706, 2025.

Hang Zhao, Qijin She, Chenyang Zhu, Yin Yang, and Kai Xu. Online 3d bin packing with constrained deep reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 741–749, 2021a.

Hang Zhao, Yang Yu, and Kai Xu. Learning efficient online 3d bin packing on packing configuration trees. In *International conference on learning representations*, 2021b.

Jianan Zhou, Zhiguang Cao, Yaoxin Wu, Wen Song, Yining Ma, Jie Zhang, and Chi Xu. Mvmoe: Multi-task vehicle routing solver with mixture-of-experts. *arXiv preprint arXiv:2405.01029*, 2024.

Zefang Zong, Xiaochen Wei, Guozhen Zhang, Chen Gao, Huandong Wang, and Yong Li. Unico: Towards a unified model for combinatorial optimization problems. *arXiv preprint arXiv:2505.06290*, 2025.

A PROOFS

This appendix gives the full statements and proofs of the results in Section 3.6. Throughout, $\mathrm{per}(R) = \mathcal{H}^1(\partial R)$, $S_R = \varphi^{-1}(R)$, and B_∂ counts boundary components: a connected region with one hole has two boundary components and one connected component.

The analysis uses three standing hypotheses, each measurable on an instance before a solver is run. (A1) (δ, η)-*locality*: at most ηm constraints satisfy $\mathrm{diam}(\varphi(\mathrm{supp}\, A_j)) > \delta$. (A2) κ-*bounded density*: every ball of radius $r \geq \delta$ contains at most $\kappa m r^2$ constraint barycentres. (A3) *admissible regions*: $\mathcal{R}_{C,B}$ consists of the unions of cells R whose boundary has $B_\partial(R) \leq B$ components and satisfies $\mathrm{per}(R) \leq C\sqrt{B_\partial(R)\,|R|}$, with the perimeter condition imposed on each region.

The distributional form of (A1) accommodates the small number of global constraints present in every family studied here. The geometric upper bounds concern S_R, the full preimage of a region. The deployed decoder returns $S \subseteq S_R$ after truncation. Since $\mathrm{cut}(\cdot)$ is not monotone under inclusion, these bounds do not transfer to the truncated set; its coupling is evaluated empirically.

A.1 THE ISOPERIMETRIC UPPER BOUND

The crossing lemma holds for arbitrary regions.

Lemma 2 (Crossing lemma)**.** *Let $R \subseteq \mathbb{R}^2$ be arbitrary, $p \in R$ and $q \notin R$. Then $[p, q] \cap \partial R \neq \emptyset$.*

Proof. If $p \in \partial R$ or $q \in \partial R$ the conclusion follows immediately, so assume $p \in \mathrm{int}\, R$ and $q \in \mathrm{int}(R^c)$. The plane decomposes as the disjoint union $\mathrm{int}\, R \sqcup \partial R \sqcup \mathrm{int}(R^c)$. If $[p, q] \cap \partial R = \emptyset$ then the connected set $[p, q]$ is contained in the union of the two disjoint open sets $\mathrm{int}\, R$ and $\mathrm{int}(R^c)$ and meets each of them, contradicting connectedness. □

Proposition 3 (Isoperimetric upper bound)**.** *Under* (A1)*,* (A2) *and* (A3)*, for every $R \in \mathcal{R}_{C,B}$ with $|R| = a$,*

$$\mathrm{cut}(S_R) \;\leq\; 9\kappa m\big(\delta\,\mathrm{per}(R) + B_\partial \delta^2\big) + \eta m \;\leq\; 9\kappa m\big(C\delta\sqrt{B_\partial a} + B_\partial \delta^2\big) + \eta m. \tag{4}$$

If moreover $\delta = c_\delta\sqrt{s/n}$, $a = k/n$ and $k \gtrsim B_\partial s$, then $\mathrm{cut}(S_R) = O(\kappa m\sqrt{skB_\partial}/n) + \eta m$.

Proof. Write $J_{\mathrm{near}} = \{j : \mathrm{diam}\,\varphi(\mathrm{supp}\, A_j) \leq \delta\}$. By (A1), $|J_{\mathrm{near}}^c| \leq \eta m$, and counting all of these constraints as cut gives their contribution to the upper bound.

Let $j \in J_{\mathrm{near}}$ be cut by S_R, so there are $i, i' \in \mathrm{supp}\, A_j$ with $\varphi(i) \in R$ and $\varphi(i') \notin R$. By Lemma 2 the segment $[\varphi(i), \varphi(i')]$ contains a point $z \in \partial R$. Since $j \in J_{\mathrm{near}}$ we have $|\varphi(i) - \varphi(i')| \leq \delta$, and z lies on that segment, so $|z - \varphi(i)| \leq \delta$. Every point of $\varphi(\mathrm{supp}\, A_j)$ lies within δ of $\varphi(i)$, hence within 2δ of z, so $\varphi(\mathrm{supp}\, A_j) \subseteq B(z, 2\delta)$ and in particular the barycentre satisfies $z_j \in (\partial R)^{2\delta}$.

It remains to count barycentres in the tube. Because R is a union of cells, ∂R is a disjoint union of B_∂ closed rectilinear polylines; let the c-th have length ℓ_c, so $\sum_c \ell_c = \mathrm{per}(R)$. Marking points along the c-th polyline at arc-length spacing δ produces at most $\ell_c/\delta + 1$ centres whose closed δ-balls cover it, and enlarging each radius to 3δ covers $(\partial R)^{2\delta}$. Hence $(\partial R)^{2\delta}$ is covered by at most $\mathrm{per}(R)/\delta + B_\partial$ balls of radius 3δ. Applying (A2) with $r = 3\delta \geq \delta$, each such ball contains at most $\kappa m(3\delta)^2 = 9\kappa m\delta^2$ barycentres, so

$$\#\{j \in J_{\mathrm{near}} : j \text{ is cut}\} \;\leq\; 9\kappa m\delta^2\Big(\frac{\mathrm{per}(R)}{\delta} + B_\partial\Big) = 9\kappa m\big(\delta\,\mathrm{per}(R) + B_\partial\delta^2\big).$$

Adding ηm gives the first inequality, and substituting $\mathrm{per}(R) \leq C\sqrt{B_\partial a}$ from (A3) gives the second.

For the specialisation, put $\delta = c_\delta\sqrt{s/n}$ and $a = k/n$. Then $\delta\,\mathrm{per}(R) \le Cc_\delta\sqrt{skB_\partial}/n$ and $B_\partial\delta^2 = B_\partial c_\delta^2 s/n$, and the ratio of the first to the second is $(C/c_\delta)\sqrt{k/(sB_\partial)}$, which is at least one exactly when $k \ge (c_\delta/C)^2 B_\partial s$. Here $k \gtrsim B_\partial s$ abbreviates this condition, with the same constant c_δ used in Corollary 12. In that regime the first term dominates and the stated bound follows. □

Remark 4 (The role of the isoperimetric inequality). The continuum inequality $\inf_{|R|=a}\mathrm{per}(R) = 2\sqrt{\pi a}$ establishes that the constraint $\mathrm{per} \le C\sqrt{B|R|}$ in (A3) is satisfiable and that the exponent of a is optimal. The proof above uses the upper bound on perimeter supplied by (A3). On a grid of resolution g a union of cells of area a has $\mathrm{per} \ge 4\lceil g\sqrt{a}\rceil/g \ge 4\sqrt{a}$, so the attainable constant is $C = 4$, realised by the axis-aligned square. The continuum value $2\sqrt{\pi} \approx 3.54$ belongs to the larger class of measurable regions. Squares are the grid isoperimetric optimum.

Remark 5 (Dependence of locality on the formulation). Reformulation changes the parameter η. On a complete-graph routing formulation, the degree constraint at customer i has as its support every edge incident to i, whose image spans the canvas, so $\eta \approx 1$ and the global-constraint term dominates Proposition 3. On a k-nearest-neighbour candidate edge set the same constraint has local support and η falls close to zero. Candidate sparsification is therefore a precondition in Section 3.1. The routing sweep in Figure 3 varies the candidate cutoff, taking $\hat{\eta}$ from near zero to near one within one family.

A.2 The random baseline

Lemma 6 (Random baseline). *Let S_{rand} be uniform among k-subsets of $[n]$ and $\rho = k/n$. For every j,*

$$\Pr[\mathrm{supp}\,A_j \subseteq S_{\mathrm{rand}}] = \prod_{i=0}^{s_j-1}\tfrac{k-i}{n-i} \le \rho^{s_j},$$
$$\Pr[\mathrm{supp}\,A_j \cap S_{\mathrm{rand}} = \emptyset] = \prod_{i=0}^{s_j-1}\tfrac{n-k-i}{n-i} \le (1-\rho)^{s_j},$$

so $\Pr[j \text{ cut}] \ge 1 - \rho^{s_j} - (1-\rho)^{s_j}$. Let $J_2 = \{j : s_j \ge 2\}$ contain the constraints that can be cut, and let $J_{\mathrm{sml}} = \{j \in J_2 : s_j \le 1/\rho\}$ contain those for which the linear estimate applies. Define the truncated mean support size by $s_{\mathrm{eff}} = \frac{1}{m}\sum_{j\in J_{\mathrm{sml}}} s_j$. For $\rho \le 1/4$,

$$\frac{m\,s_{\mathrm{eff}}\,k}{4n} + 0.56\,\big|J_2 \setminus J_{\mathrm{sml}}\big| \;\le\; \mathbb{E}[\mathrm{cut}(S_{\mathrm{rand}})] \;\le\; \min\Big\{m,\ \frac{msk}{n}\Big\}. \tag{5}$$

Proof. Counting k-subsets containing a fixed s_j-set gives $\Pr[\mathrm{supp}\,A_j \subseteq S_{\mathrm{rand}}] = \binom{n-s_j}{k-s_j}/\binom{n}{k} = \prod_{i=0}^{s_j-1}\frac{k-i}{n-i}$, and each factor is at most $k/n = \rho$ because $k \le n$. Similarly $\Pr[\mathrm{supp}\,A_j \cap S_{\mathrm{rand}} = \emptyset] = \binom{n-s_j}{k}/\binom{n}{k} = \prod_{i=0}^{s_j-1}\frac{n-k-i}{n-i}$ with each factor at most $1-\rho$. A constraint is cut precisely when neither event occurs, giving the stated bound. Sampling without replacement makes each event *less* likely than under independent sampling, which gives the stated lower bound.

For the upper bound in (5), a constraint can only be cut if it is met, so by a union bound $\Pr[j \text{ cut}] \le s_j\rho$, and summing gives $\mathbb{E}[\mathrm{cut}(S_{\mathrm{rand}})] \le \rho\sum_j s_j = \rho\,\mathrm{nnz}(A) = msk/n$; the bound by m is trivial.

For the lower bound, fix $j \in J_{\mathrm{sml}}$ and set $u = s_j\rho \in (0,1]$. Since $s_j \ge 2$ and $\rho \le 1/4$ we have $\rho^{s_j} \le \rho^2 \le \rho/4$, and $(1-\rho)^{s_j} \le e^{-s_j\rho} = e^{-u}$. Hence $\Pr[j \text{ cut}] \ge 1 - e^{-u} - \rho/4$. For $u \le 1$, $1 - e^{-u} \ge u - u^2/2 \ge u/2$. Also $u = s_j\rho \ge 2\rho$, so $\rho/4 \le u/8$, giving $\Pr[j \text{ cut}] \ge u/2 - u/8 = 3u/8 \ge s_j\rho/4$. Summing over J_{sml} yields $\frac{\rho}{4}\sum_{j\in J_{\mathrm{sml}}} s_j = \frac{m s_{\mathrm{eff}} k}{4n}$.

For $j \in J_2 \setminus J_{\mathrm{sml}}$ we have $s_j > 1/\rho \ge 4$, so $\rho^{s_j} \le \rho^2 \le 1/16$ and $(1-\rho)^{s_j} \le e^{-s_j\rho} < e^{-1}$, whence $\Pr[j \text{ cut}] > 1 - 1/16 - e^{-1} > 0.56$. Adding the two groups gives (5). □

Remark 7 (Truncated support sizes in the lower bound). The map $s \mapsto 1-(1-\rho)^s$ is concave, so replacing each s_j by the mean s yields an upper bound on $\mathbb{E}[\mathrm{cut}(S_{\mathrm{rand}})]$. On families whose support sizes are heavy tailed, the long supports are almost surely cut and contribute $\Theta(1)$ each. The lower bound therefore uses s_{eff}, truncated at $1/\rho$, together with a separate contribution from long supports. The approximation $\mathbb{E}[\mathrm{cut}(S_{\mathrm{rand}})] \approx msk/n$ can overstate the random baseline on these families and, consequently, the coupling advantage.

Corollary 8 (Peak of the coupling-advantage envelope, [CONDITIONAL]). *Let η be negligible, $B_\partial = 1$ and $\rho \le 1/4$. Dividing* (5) *by* (4) *gives a lower envelope on the coupling advantage with geometric factor $\kappa^{-1}\min\{\sqrt{sk},\, n/\sqrt{sk}\}$, unimodal in ρ with peak $\Omega(\sqrt{n}/\kappa)$ at $\rho^\star = 1/s$. The two branches rest on different hypotheses. Below the peak the envelope rises as $\sqrt{s\rho n}$ provided $s_{\text{eff}} \asymp s$. Above it the envelope falls as $\sqrt{n/(s\rho)}$ provided a constant fraction of constraints have $s_j \ge 1/\rho$, a condition separate from $s_{\text{eff}} \asymp s$.*

Every s_j entering s_{eff} is at most $1/\rho$, so $s_{\text{eff}} \asymp s$ implies $s\rho = O(1)$ and applies to the lower-ρ branch. For uniform support size, the transition occurs at $s\rho = 1$.

Proof of Corollary 8. With $B_\partial = 1$ and η negligible, Proposition 3 gives $\mathrm{cut}(S_R) = O(\kappa m\sqrt{sk}/n) = O(\kappa m\sqrt{s\rho/n})$. For the numerator we take the two branches separately. When $s\rho \le 1$ and $s_{\text{eff}} \asymp s$, the first term of (5) gives $\mathbb{E}[\mathrm{cut}(S_{\text{rand}})] = \Omega(m s_{\text{eff}}\rho) = \Omega(ms\rho)$. When $s\rho \ge 1$, the bound $s_{\text{eff}} \le 1/\rho$ limits the lower estimate available from the first term to $\Omega(m)$ at best, and $s_{\text{eff}} \asymp s$ imposes the corresponding restriction on support size. The second term gives $\mathbb{E}[\mathrm{cut}(S_{\text{rand}})] \ge 0.56|J_2 \setminus J_{\text{sml}}| = \Omega(m)$ under the stated hypothesis that a constant fraction of constraints have $s_j \ge 1/\rho$. Combining, $\mathbb{E}[\mathrm{cut}(S_{\text{rand}})] = \Omega(m\min\{1, s\rho\})$ on the respective branches, and their ratio is

$$\Omega\left(\frac{\min\{1, s\rho\}}{\kappa\sqrt{s\rho/n}}\right) = \frac{1}{\kappa}\cdot\begin{cases}\sqrt{s\rho n} = \sqrt{sk}, & s\rho \le 1,\\ \sqrt{n/(s\rho)} = n/\sqrt{sk}, & s\rho \ge 1,\end{cases}$$

which is $\kappa^{-1}\min\{\sqrt{sk}, n/\sqrt{sk}\}$. The two branches are increasing and decreasing in ρ respectively and meet where $sk = n$, that is at $\rho = 1/s$, where the common value is $\sqrt{n}/\kappa$. □

The ratio of a lower bound to an upper bound gives a lower bound on the true coupling advantage. The constants, the ratio s_{eff}/s and κ vary with ρ, so the peak height is specified by Ω rather than Θ. Both branches inherit the restriction $\rho \le 1/4$ from Lemma 6, and the peak $\rho^\star = 1/s$ lies in this range when $s \ge 4$. The set cover family satisfies this condition with $\hat{s} = 9.6$. In vertex cover, every row has support two, placing $\rho^\star = 1/2$ outside the proved range of Lemma 6. The ρ sweep on vertex cover provides a descriptive comparison of curve shapes, rather than a test of this corollary.

A.3 THE DEGREE-OF-FREEDOM CERTIFICATE

Proposition (Proposition 1 restated). *If $\mathrm{rank}(A_{E(S),S}) = |S| = k$ then $x^\star$ is the unique feasible point of $\mathrm{SUB}(S, x^\star)$. Consequently $\mathrm{rank}(A_{E(S),S}) < k$ is necessary for improvement, and $\mathrm{cut}_E(S) + \mathrm{in}_E(S) < k$ is sufficient for that necessary condition.*

Proof. In $\mathrm{SUB}(S, x^\star)$ every variable outside S is fixed to $x^\star$. An equality row $j \in E$ whose support lies entirely in $\bar{S}$ is satisfied identically by $x^\star$ and constrains nothing. Each remaining equality row meets S, and collectively these rows impose the affine system

$$A_{E(S),S}\, x_S = b_{E(S)} - A_{E(S),\bar{S}}\, x^\star_{\bar{S}},$$

of which $x^\star_S$ is a solution. If the coefficient matrix has full column rank k then this solution is unique, so every point of $\mathcal{F}(\mathrm{SUB}(S, x^\star))$, integral or otherwise, agrees with $x^\star$ on S and hence equals $x^\star$; the objective cannot decrease. The matrix $A_{E(S),S}$ has exactly $|E(S)| = \mathrm{cut}_E(S) + \mathrm{in}_E(S)$ rows, since an equality row meets S iff its support either straddles S or lies inside it, and rank never exceeds the number of rows. □

Remark 9 (Direction of the implication). The diagnostic chain is $\mathrm{cut}_E(S) + \mathrm{in}_E(S) < k \Rightarrow \mathrm{rank} < k \Rightarrow$ improvement is not excluded. A row count of k or more remains compatible with rank deficiency and improvement. The rank itself determines the certificate, while the row count provides a sufficient check for rank deficiency. The corresponding figure therefore uses $\mathrm{rank}\,/k$ on its horizontal axis.

Remark 10 (Tight inequalities). Inequality rows that are tight at $x^\star$ constrain local motion. Counting them together with equality rows gives a practical heuristic: an inequality tight at $x^\star$ may become slack at a better point. We distinguish this active-constraint diagnostic from the equality-rank certificate wherever it is used.

A.4 THE EXPANDER LOWER BOUND

Proposition 11 (Expander lower bound). *Let the incidence graph be left d-regular with $s_j \equiv s \geq 2$ and a $(\gamma n, (1-\epsilon)d)$ lossless expander, with $\epsilon \leq 1/2$. Then for every S with $|S| = k \leq \gamma n$,*

$$\operatorname{cut}(S) \geq dk\Big(1 - \epsilon - \frac{\epsilon}{s-1}\Big) \geq (1-2\epsilon)dk \geq (1-2\epsilon)\mathbb{E}[\operatorname{cut}(S_{\mathrm{rand}})].$$

Proof. Let $n_r = \#\{j : |\operatorname{supp} A_j \cap S| = r\}$ for $r \geq 1$. Every constraint meeting S is counted exactly once, so $\sum_{r\geq 1} n_r = |N(S)|$, and counting incidences from the variable side, which is d-regular, gives $\sum_{r\geq 1} r\, n_r = dk$. Subtracting,

$$dk - |N(S)| = \sum_{r\geq 2}(r-1)n_r \;\geq\; (s-1)n_s,$$

since all terms are non-negative and the term at $r = s$ contributes $(s-1)n_s$. Lossless expansion gives $|N(S)| \geq (1-\epsilon)dk$, so $n_s \leq \epsilon dk/(s-1)$. A constraint meeting S fails to be cut exactly when its whole support lies in S, which for uniform support size s means $|\operatorname{supp} A_j \cap S| = s$; therefore $\operatorname{cut}(S) = |N(S)| - n_s \geq (1-\epsilon)dk - \epsilon dk/(s-1)$. Since $s \geq 2$ we have $\epsilon/(s-1) \leq \epsilon$, giving the second inequality. The third follows from Lemma 6, whose upper bound reads $\mathbb{E}[\operatorname{cut}(S_{\mathrm{rand}})] \leq \rho \operatorname{nnz}(A) = \rho n d = dk$, multiplied by the non-negative factor $1-2\epsilon$, using $\epsilon \leq 1/2$. For $\epsilon > 1/2$, the expansion hypothesis permits small $\operatorname{cut}(S)$, and the comparison gives no positive lower bound. □

The construction is realisable: random d-regular bipartite graphs are lossless expanders with high probability, and explicit constructions are known (Capalbo et al., 2002). Taking $\sum_{i\in\operatorname{supp} A_j} x_i \geq 1$ over such a graph gives a set cover instance in the family. Comparing this lower bound with Proposition 3 constrains the jointly attainable locality, density and region parameters wherever their hypotheses overlap. Limits such as $\eta \to 1$ or $\kappa \to \infty$ describe possible failures of favourable layout conditions; establishing either limit requires assumptions on the instance sequence and region scale.

A.5 FRAGMENTATION AND THE DIMENSION OF THE CANVAS

Corollary 12 (Cost of a fragmented mask, [CONDITIONAL]). *Under the hypotheses of Proposition 3, let R consist of B_∂ disjoint near-circular blocks of total area $a = k/n$, with $\delta = c_\delta\sqrt{s/n}$. Then*

$$\operatorname{cut}(S_R) = O\Big(\kappa\frac{m}{n}\big(\sqrt{skB_\partial} + B_\partial s\big)\Big) + \eta m.$$

If in addition $s_{\mathrm{eff}} \asymp s$ and $s\rho \leq 1$, so that Lemma 6 bounds the random baseline at $\Theta(msk/n)$, then

$$\mathrm{Gain}^{-1} = O\left(\kappa\sqrt{\frac{B_\partial}{sk}} + \kappa\frac{B_\partial}{k}\right) + \frac{\eta n}{sk}.$$

There are two thresholds. At $B^\dagger \asymp k/(c_\delta^2 s)$, the block radius reaches the locality scale δ, and the perimeter and boundary-component terms in the upper bound exchange dominance. At $B^\ddagger \asymp k$, each block holds $O(1)$ variables, and this bound no longer guarantees an advantage over random selection.

Proof. For B_∂ disjoint discs of total area a, each has radius $r = \sqrt{a/(\pi B_\partial)}$ and the total perimeter is $B_\partial \cdot 2\pi r = 2\sqrt{\pi a B_\partial}$. A union of cells of the same area has a larger perimeter, by at most the factor $4/(2\sqrt{\pi})$ of Remark 4, so the disc calculation is correct up to that constant. Substituting into Proposition 3 with $a = k/n$ gives $\delta\operatorname{per}(R) = 2\sqrt{\pi}c_\delta\sqrt{skB_\partial}/n$ and $B_\partial\delta^2 = B_\partial c_\delta^2 s/n$, which is the first display. Under the added hypotheses Lemma 6 gives $\mathbb{E}[\operatorname{cut}(S_{\mathrm{rand}})] \asymp msk/n$, and dividing the first display by it gives the second, the last term being ηm over that same baseline. Solving $r = \delta$, that is $\sqrt{(k/n)/(\pi B_\partial)} = c_\delta\sqrt{s/n}$, gives $B_\partial = k/(\pi c_\delta^2 s)$. The second threshold is where the leading term of Gain^{-1} reaches $\Theta(1)$, namely $B_\partial \asymp k$. □

The condition $s_{\mathrm{eff}} \asymp s$ specialises Lemma 6 to the support-size regime used by this corollary. On a heavy-tailed family, applying Lemma 6 with the truncated mean introduces the factor s/s_{eff} in

the inverse-advantage bound. Retaining c_δ in $B^\dagger$ makes the locality scale explicit: a family whose measured $\hat{\delta}/\sqrt{s/n}$ is 5 has a threshold twenty-five times smaller than the expression omitting that constant. Similarly, $B^\dagger$ and the crossover $B_\partial \asymp (C/c_\delta)^2 k/s$ at which the terms of Proposition 3 exchange dominance agree up to a constant factor, which may be a few tens. The scale $B_\partial \approx sk$ exceeds the feasible range of the whole-cell release variant: at most k cells are selected, and the numbers of outer boundaries and enclosed holes give $B_\partial = O(k)$. Counting boundary components includes the contribution of holes. The $B_\partial \delta^2$ term captures this fine-fragmentation regime.

Corollary 13 (Dimension and locality, [CONDITIONAL])**.** *Let $s_{\text{eff}} \asymp s$ and $s\rho \le 1$. On a D-dimensional canvas with ideal locality scale $\delta_D = \Theta((s/n)^{1/D})$, a region of volume a whose surface measure is $O(a^{(D-1)/D})$ satisfies $\mathrm{cut}() = O(\kappa m\, a^{(D-1)/D}\delta_D) + \eta_D m$, so its geometric advantage obeys*

$$\text{Gain} \;\gtrsim\; \frac{1}{\kappa}\cdot\frac{s\rho}{a^{(D-1)/D}(s/n)^{1/D}} \;=\; \frac{s}{\kappa}\Big(\frac{k}{s}\Big)^{1/D},$$

which for $k > s$ is strictly decreasing in D. Carrying the locality defect as well,

$$\mathrm{Gain}(D) = \Omega\Big(\min\Big\{\tfrac{s}{\kappa}(k/s)^{1/D},\ \tfrac{s\rho}{\eta_D}\Big\}\Big).$$

Proof. Ideal locality in D dimensions requires a set of diameter δ_D to hold s variables at density n, so $n\delta_D^D \asymp s$. Substituting $a = \rho = k/n$ into the displayed ratio gives $s\rho^{1/D}(n/s)^{1/D} = s(\rho n/s)^{1/D} = s(k/s)^{1/D}$. At $D = 2$ this is $\sqrt{sk}$ and at $D = 1$ it is k, so the exponent is consistent with Proposition 3. Monotonicity in D is immediate from $k/s > 1$. For the last display, $\mathrm{Gain} \ge \mathbb{E}[\mathrm{cut}(S_{\mathrm{rand}})]/(\mathrm{Geo} + \eta_D m) \ge \frac{1}{2}\min\{\mathbb{E}[\mathrm{cut}(S_{\mathrm{rand}})]/\mathrm{Geo},\ \mathbb{E}[\mathrm{cut}(S_{\mathrm{rand}})]/(\eta_D m)\}$, and under the stated hypotheses $\mathbb{E}[\mathrm{cut}(S_{\mathrm{rand}})] \asymp ms\rho$, so the second argument of the minimum is $s\rho/\eta_D$. □

The locality contribution retains the factor $s\rho/\eta_D$. Replacing it by $1/\eta_D$ changes the bound by $1/(s\rho) = n/(sk)$, a factor of one hundred at $s = 10$, $k = 10^3$ and $n = 10^6$. The geometric result is a lower bound on the advantage, obtained from the upper bound on the cut in Proposition 3; a matching lower bound on the cut would be needed to establish the order of the advantage.

The geometric term favours a one-dimensional canvas, while the locality defect η_D governs the choice of $D = 2$. For intrinsically two-dimensional structure, a $\sqrt{n} \times \sqrt{n}$ grid graph has bandwidth $\Theta(\sqrt{n})$, so a one-dimensional layout loses locality and $\eta_1 \to 1$. For intrinsically one-dimensional structure, such as staircase or banded programs arising from multi-period planning and time-expanded network flow, $\eta_1 \approx \eta_2 \approx 0$ and the bound favours a one-dimensional canvas. The ablation design includes a test of this prediction with a one-dimensional variant on a banded subfamily. This variant uses its own weights, while the shared-weight experiments concern the two-dimensional network.

A.6 Resolution and granularity

We define the perimeter of a destroy set through a reference grid finer than any grid used by the policy. A definition over all measurable regions would admit $R = \varphi(S)$, a finite point set with empty interior and $\partial R = R$, giving $\mathcal{H}^1(\partial R) = 0$ for every destroy set. The reference grid supplies the regularity needed for a geometric measure of the set.

Definition 14. Fix a reference resolution $g_0 \gg g$ and call $S \subseteq [n]$ *resolvable* at g_0 if some union R of g_0-cells has $\varphi^{-1}(R) = S$, which holds as soon as the g_0-grid separates S from $[n] \setminus S$. For such S let $\mathrm{per}(S) = \min\{\mathcal{H}^1(\partial R) : R \text{ a union of } g_0\text{-cells}, \varphi^{-1}(R) = S\}$, the minimum being over a finite family and therefore attained, and let $B^\star$ be the number of boundary components of a minimiser. Every such $\mathrm{per}(S)$ is a positive multiple of $1/g_0$.

Theorem 15 (Discretisation error, [CONDITIONAL])**.** *Let $S^\star$ be resolvable at g_0. Assume* (A2′)*, that every measurable U with $|U| \ge 2/g^2$ contains at most $\rho_{\max}|U|$ variables, and* (A4)*, that there exists L with $\Delta(S^\star) - \Delta(\hat{S}) \le L|S^\star \setminus \hat{S}|$ for every union of cells $\hat{S} \subseteq S^\star$. Let $\hat{S}$ be the best destroy set representable as a union of cells at resolution g. Then, deterministically,*

$$\Delta(S^\star) - \Delta(\hat{S}) \;\le\; L\rho_{\max}\Big(\frac{C_1\,\mathrm{per}(S^\star)}{g} + \frac{C_2 B^\star}{g^2}\Big), \qquad C_1 = 2\sqrt{2},\ C_2 = 2\pi.$$

Proof. Let $R^\star$ attain the infimum in Definition 14. Let $\mathcal{Q}_{\text{in}}$ be the cells contained in $\operatorname{int} R^\star$ and put $\tilde{S} = \varphi^{-1}(\bigcup \mathcal{Q}_{\text{in}})$. Then $\tilde{S}$ is a union of cells with $\tilde{S} \subseteq S^\star$, so $|\tilde{S}| \le k$ and $\tilde{S}$ is feasible; by optimality of $\hat{S}$ among unions of cells, $\Delta(\hat{S}) \ge \Delta(\tilde{S})$.

If $i \in S^\star \setminus \tilde{S}$ then $\varphi(i) \in R^\star$ but the cell containing $\varphi(i)$ is not contained in $\operatorname{int} R^\star$, so that cell meets $\partial R^\star$ and therefore $\operatorname{dist}(\varphi(i), \partial R^\star) \le \sqrt{2}/g$, the cell diameter. Hence $S^\star \setminus \tilde{S} \subseteq \varphi^{-1}\big((\partial R^\star)^{\sqrt{2}/g}\big)$. The boundary consists of $B^\star$ closed curves of total length $\operatorname{per}(S^\star)$, and a tube of radius t about them has area at most $2t \operatorname{per}(S^\star) + \pi B^\star t^2$; with $t = \sqrt{2}/g$ this is $2\sqrt{2}\operatorname{per}(S^\star)/g + 2\pi B^\star/g^2$. The tube also contains a disc of radius t, so its area is at least $2\pi/g^2$ and the cutoff in (A2′) is met; applying (A2′) bounds $|S^\star \setminus \tilde{S}|$ by $\rho_{\max}$ times the area, and (A4) converts the count into a value difference. □

Remark 16 (Density cutoff in (A2′)). The cutoff makes the density hypothesis meaningful for a finite point set. A disc of radius ε about any $\varphi(i)$ contains one variable in area $\pi\varepsilon^2$, which would force $\rho_{\max} \ge 1/(\pi\varepsilon^2)$ and hence $\rho_{\max} = \infty$ if arbitrarily small discs were included. Likewise, (A2) applies to balls of radius $r \ge \delta$. The proof uses a tube of radius $\sqrt{2}/g$, whose area meets the cutoff.

Remark 17 (Sub-cell components and random grid offsets). The second term, $O(L\rho_{\max}B^\star/g^2)$, measures loss from sub-cell components. A term $O(Lk/g^2)$ would instead scale with the variable budget k. The boundary-component term dominates when $\operatorname{per}(R^\star) \le B^\star/g$, corresponding to components smaller than one cell. Under (A2′) the bound is deterministic and applies to every grid offset. Random offsets offer two additional properties. A component of diameter ε/g that straddles a grid line contributes no complete cell, whereas under a uniform offset it lies in a single cell with probability at least $(1-\varepsilon)^2$. Averaging over offsets also replaces $\rho_{\max}$ by the mean density along the boundary, $\mathbb{E}_u|S^\star \setminus \tilde{S}_u| \le \frac{C}{g}\int_{\partial R^\star} \rho_{1/g}\, d\mathcal{H}^1$. This form can sharpen the bound on clustered instances when the mask boundary passes through sparse regions.

The next corollary considers a compact optimum whose reference-grid minimiser consists of $B^\star$ near-circular blocks. Its derivation also uses the boundary-density substitution in Remark 17, associated with averaging over random grid offsets.

Corollary 18 (Saturation resolution, [CONDITIONAL]). *Let $\zeta = \Delta(S^\star)/(Lk) \in (0,1]$ be the budget efficiency. Requiring the first term of Theorem 15 to contribute at most a relative error ϵ gives*

$$g^\star = \frac{2C_1\sqrt{\pi}}{\epsilon\zeta}\sqrt{\frac{B^\star\Lambda_\partial}{n\rho}} \asymp c_g\sqrt{\frac{B^\star\,\Lambda_\partial/n}{\rho}}, \qquad \rho = k/n,$$

where Λ_∂ is the mean density along $\partial R^\star$. At $g = g^\star$ the second term is $O(\epsilon^2\zeta)$ and is therefore lower order, so the balance is self-consistent.

Proof. Dividing Theorem 15 by $\Delta(S^\star) = \zeta Lk$ gives relative error $\rho_{\max}(C_1 \operatorname{per}/g + C_2B^\star/g^2)/(\zeta k)$. For a compact optimum whose minimiser has $B^\star$ near-circular blocks holding k variables at density $\rho_{\max}$, the total area is $a = k/\rho_{\max}$ and hence $\operatorname{per} = 2\sqrt{\pi a B^\star} = 2\sqrt{\pi k B^\star/\rho_{\max}}$. The first term is then $2C_1\sqrt{\pi}\sqrt{\rho_{\max}B^\star/k}/(\zeta g)$, and setting it equal to ϵ and substituting $\rho_{\max} \to \Lambda_\partial$ from Remark 17 and $k = \rho n$ gives the display. Substituting $g = g^\star$ into the second term gives $C_2\epsilon^2\zeta/(4C_1^2\pi)$. □

On clustered instances, using the global maximum density in place of Λ_∂ increases the predicted $g^\star$ by a factor $\sqrt{\Lambda_{\max}/\Lambda_\partial}$, so boundary density and maximum density lead to different resolution estimates under (A2′). Here $B^\star$ counts boundary components of the minimiser in Definition 14, while B_∂ in (A3) describes the regions available to the policy. They can coincide when the optimum is itself admissible. The theory specifies the scaling law $\sqrt{B^\star/\rho}$. Its constant c_g absorbs ϵ, ζ and C_1, so fitting c_g on one family is required before predicting on another. The multi-scale refinement in Section B.6 remains relevant below the single-canvas resolution condition $g \gtrsim c_g\sqrt{B^\star/\rho}$. At $g = 128$ with $B^\star = 4$, the implied threshold on ρ varies over three orders of magnitude across plausible values of c_g.

Remark 19 (Representation and learning error). Theorem 15 bounds the representational loss of restricting to unions of cells, with $\hat{S}$ defined as the best representable set. A learned policy also incurs estimation and optimisation error. In a sweep over g, the output dimension grows as g^2 at

a fixed sample count, so estimation error can cause empirical saturation to occur *earlier* than the approximation bound predicts. The comparison between predicted and measured saturation reflects these combined errors.

Remark 20 (Relation to partition-based approximation). Classical partitioning results for the Euclidean travelling salesman problem (Karp, 1977) solve each group independently and stitch the pieces together. Their error arises from discarding structure across group boundaries and decreases with larger groups. FOVEA solves the released subproblem within the fixed incumbent, preserving cross-boundary coupling through the fixed variables. Its analysis therefore concerns the loss from representing the destroy set on a grid; the stitching error in partition-based approximation belongs to a different construction.

A.7 From per-round cost to the primal integral

We write λ for the per-round contraction rate of the primal gap, reserving ρ for the destroy fraction.

Proposition 21 (Critical instance size, [conditional]). *Suppose the first feasible solution appears at t_0 with gap γ_0, each round costs $\tau = \tau_{\text{select}} + \tau_{\text{repair}}$, and* (B1) *holds: $\mathbb{E}[\gamma_{t+1} \mid \gamma_t] \le (1-\lambda)\gamma_t$. Then with $N = \lfloor (T - t_0)/\tau \rfloor$,*

$$\mathbb{E}[P(T)] \;\le\; t_0 + \frac{\tau\gamma_0}{\lambda}\big(1 - (1-\lambda)^{N+1}\big) \;\le\; t_0 + \frac{\tau\gamma_0}{\lambda}.$$

Compare a selector whose per-round cost is $c_F = c_0 + c_1 n/N$ against one of cost $c_G d n$, with quality ratio $r = \lambda_G/\lambda_F \ge 1$. Here $c_0 = \tau_{\text{net}} + \tau_{\text{lp}}/R_{\text{lp}}$ is independent of n and $c_1 = c_{\text{layout}} + c_{\text{raster}}$ is the one-off setup amortised over the N rounds of the run. The two bounds coincide at

$$n^\star = \frac{r\,c_0 + (r-1)\tau_{\text{repair}}}{c_G d - r\,c_1/N}, \qquad \textit{defined only when} \quad c_G d > \frac{r\,c_1}{N}.$$

Proof. Before t_0 the integrand is 1, contributing t_0. Afterwards the gap is constant within each round, so $\int_{t_0}^{T} p \le \tau \sum_{t=0}^{N-1} \gamma_t + \tau\gamma_N$. Taking expectations and iterating (B1) gives $\mathbb{E}\gamma_t \le q^t \gamma_0$ with $q = 1 - \lambda$, so the sum is at most

$$\tau\gamma_0\Big(\frac{1-q^N}{\lambda} + q^N\Big) = \frac{\tau\gamma_0}{\lambda}\big(1 - q^N(1-\lambda)\big) = \frac{\tau\gamma_0}{\lambda}\big(1 - q^{N+1}\big).$$

For the second claim, equate $(c_F + \tau_{\text{repair}})/\lambda_F$ with $(c_G d n + \tau_{\text{repair}})/\lambda_G$, which gives $r c_F + (r-1)\tau_{\text{repair}} = c_G d n$. Substituting $c_F = c_0 + c_1 n/N$ puts n on both sides, and collecting it yields $n(c_G d - r c_1/N) = r c_0 + (r-1)\tau_{\text{repair}}$ and hence the display. □

Remark 22 (Amortised setup cost). FOVEA's per-round cost includes the amortised share $c_1 n/N$ of a layout and a first rasterisation that are both $\Theta(n)$. This makes the cost affine in n. A positive $n^\star$ exists when the graph selector's slope $c_G d$ exceeds FOVEA's amortised slope $r c_1/N$. Short runs at very large n can be dominated by setup and have no positive crossing. Reporting c_1/N alongside $n^\star$ identifies this regime.

Remark 23 (Finite-horizon correction). The geometric sum contributes $-\frac{\tau\gamma_0}{\lambda}(1-\lambda)^{N+1}$, which reduces the infinite-horizon bound by the unspent tail of the series. This finite-horizon correction is exact within the geometric bound.

Remark 24 (The plateau bounds the applicability). Real runs converge to some $\gamma_\infty > 0$. Writing $\tilde{\gamma}_t = \gamma_t - \gamma_\infty$ and assuming (B1′), the same argument gives $\mathbb{E}[P(T)] \le t_0 + \gamma_\infty(T - t_0) + \tau\tilde{\gamma}_0/\lambda$. The middle term is linear in T and independent of both τ and λ. At long horizons, the plateau height therefore dominates the comparison and reflects neighbourhood quality. The transient dominates while $T - t_0 \ll \tau\tilde{\gamma}_0/(\lambda\gamma_\infty)$, defining the regime in which the selection-cost comparison of Proposition 21 is informative.

Remark 25 (How to read $n^\star$). $n^\star$ equates two upper bounds of different tightness and provides an order-of-magnitude estimate of the crossing. It depends on the quality deficit and encoder slope. The numerator contains two terms. $(r-1)\tau_{\text{repair}}$ overtakes $r c_0$ once r exceeds $1 + c_0/\tau_{\text{repair}}$, which on the constants of Table 11 is under two and a half percent. At the measured $r = 1.03$, the terms are 0.044 and 0.054 s and both contribute materially. Their sensitivities differ: halving the encoder slope doubles $n^\star$, whereas doubling the quality deficit $r - 1$ raises it by about half because $r c_0$ remains. We report an interval that propagates measurement error in both quantities. Empirical consistency

includes a measured crossing inside that interval, or a predicted crossing beyond the largest available instance accompanied by no observed crossing; the unknown tightness of the bounds limits the precision of either comparison.

A.8 Memory accounting

Memory accounting separates the network input, resident raster state and host-side index. The network reads a $128 \times 128 \times 12$ single-precision tensor occupying 786 kB. The rasteriser's resident state is four times that size: it holds the grid in double precision together with the per-cell sums and counts used by incremental updates. Both reside on the device and are constant in n. A graph encoder stores the incidence structure and its activations on the device, requiring $O(\mathrm{nnz}(A))$ memory. Figure 2(c) reports device memory for these selectors; on this family the graph baselines exceed the 40 GB budget between $n = 3 \times 10^4$ and $n = 10^5$.

FOVEA also stores three $O(n)$ arrays on the host: the layout, with two 32-bit coordinates per variable; each variable's cell identifier; and an inverted index from cells to variables, with one 32-bit identifier per variable and a table of HW offsets. These arrays occupy sixteen bytes per variable, or 16 MB at $n = 10^6$, shown by the dotted line in Figure 2(c). Per-variable channel quantities are streamed from the solver at each refresh, keeping resident host storage at sixteen bytes per variable. The representation thus places the instance-sized arrays in host memory while keeping the network input independent of n.

B Method details

B.1 The budget regime

Holding k fixed in absolute terms defines the scaling regime. Under the locality conditions of Section 3.6, the geometric coupling-advantage bound scales as $\sqrt{sk}/\kappa$ and remains constant in n on families with constant density parameters; Proposition 11 characterises a contrasting regime. When $n \gg ks$, the destroy fraction $\rho = k/n$ lies below the predicted peak location $\rho^\star = 1/s$ of the coupling-advantage envelope. These instances occupy the branch that rises with ρ, where the geometric factor is $\sqrt{sk}/\kappa$. We evaluate the empirical trend motivated by Corollary 8 at $n \le 10^4$. By Corollary 18, a small ρ also raises the saturation resolution $g^\star$, motivating the multi-scale extension in Section B.6. Every figure with n on the horizontal axis states that k is fixed, so the scaling curves describe a common absolute repair budget.

B.2 Entity layers

Table 5 instantiates the layer of Section 3.1 for the four families. Two placements distribute mass across cells, so (3) is written with a mass threshold: a variable of a general mixed integer program occupies one column and as many rows as it has constraint blocks, and a scheduling operation occupies a horizontal run proportional to its duration.

Table 5: The entity layer per family. *Arity* is $|\mathrm{ent}(i)|$, the number of entities a variable refers to; the layer is vacuous exactly when the arity is one and the association is the identity.

Family	Entity	Placement	Arity	Released set S
MILP	variable	column = reordered column rank, rows = its constraint blocks	1	variables whose mass in the region reaches ϑ
CVRP	customer	native coordinates, one-hot	2	candidate edges with both endpoints selected
JSSP	operation	(interval midpoint, machine by load), spread over the duration	2	start times of the operations under the mask, plus the ordering variables they induce
1D-BPP	(item, bin) pair	(item rank by size, bin rank by residual capacity), one-hot	1	assignment variables under the mask

Scheduling uses the same two-level structure as routing. A disjunctive formulation carries an ordering variable $z_{o,o'}$ for each pair of operations competing for a machine, so a variable refers to two entities and the arity is two. Releasing the ordering variables induced by the selected operations allows the sub-solver to change their machine sequence along with their start times. Operations correspond to customers in routing, and ordering variables to edges. The scheduling layout is rebuilt whenever the incumbent changes because an operation's time coordinate depends on the current solution; it is the one layout computed more than once. Packing uses artificial axes, with bins ordered by residual capacity so that adjacent positions have comparable slack. Variables without spatial meaning, such as bin-used indicators, are marked non-spatial and handled by the sub-solver.

B.3 CHANNELS

Table 6: The twelve input channels. *Class* determines refresh cost: for a fixed layout, static channels are computed once; state channels are updated incrementally, and dual channels require a linear programming solve and are refreshed every R_{lp} rounds. HGS and CP-SAT provide signals for eight channels on routing and scheduling; an availability mask identifies the remaining relaxation-dependent channels.

#	Channel	Class	Aggregation	Semantics
1	occupancy	static	count	placement mass of the cell
2	solution	state	mean	incumbent variable values
3	lp_relaxation	dual	mean	relaxation values
4	integrality_gap	dual	mean	$\lvert x_i^\star - \tilde{x}_i \rvert$; the RINS signal and the default score
5	reduced_cost	dual	mean	variable reduced costs
6	objective_coeff	static	mean	coefficients of the objective
7	pseudocost	dual	mean	the solver's own difficulty estimate
8	tightness	state	mean	mass of tight constraints attributed to i
9	improvement_history	state	mean	gain when i was last released
10	recency	state	mean	rounds since i was last released
11	type_mask	static	mean	binary 1, integer $1/2$, continuous 0
12	boundary	static	mean	share of i's constraints reaching far across the canvas

The input tensor has twelve channels for every family. Channels 3, 4, 5 and 7 carry relaxation-dependent signals only when the sub-solver is a mixed integer programming solver. General programs and packing therefore provide signals for all twelve channels, while routing and scheduling provide eight; the remaining four use neutral values and are marked unavailable. The channel-removal ablation design uses the five groups geometry $\{1, 12\}$, solution $\{2, 8\}$, relaxation $\{3, 4, 5, 7\}$, history $\{9, 10\}$ and entity $\{6, 11\}$.

Normalisation. The channel transforms and within-instance statistics are specified in Section 3.3.

B.4 SAMPLING

Sample takes the heatmap $M \in [0,1]^{H \times W}$, the budget k, an oversample factor $\beta \geq 1$ and a Gumbel scale ε, and returns a cell set. It perturbs each cell to $\log M_{pq} + \varepsilon g_{pq}$ with independent $g_{pq} \sim \mathrm{Gumbel}(0, 1)$, then walks the cells in decreasing perturbed order, admitting them until the admitted cells hold βk variables or the cells with $M_{pq} > 0$ run out. A morphological closing follows, and then the singleton components are discarded.

At unit Gumbel scale, taking cells in that order samples without replacement with probability proportional to M (Kool et al., 2019); varying ε adjusts the concentration of the sampling distribution. When the positively responding cells run out before βk is reached, the round uses the available cells and a smaller budget. The morphological steps control different features: closing joins components separated by a single-cell gap, and singleton pruning removes isolated cells. Both contribute to controlling the boundary-component count B_∂ in Corollary 12.

B.5 TRUNCATING A REGION TO THE BUDGET

The default truncation rule preserves cell structure within the budget. It selects whole cells in decreasing cell score, then uses the per-variable score to fill the remaining budget from the next cell. The score-only alternative takes the k highest-scoring variables across the sampled region and can distribute the selection across its cells. The latency accounting assumes this alternative. The two rules tend to agree for sharply peaked heatmaps and differ for flatter ones, where the perimeter penalty has greater influence. The ablation design compares both rules with a cell-level rule that omits per-variable scoring to assess the resulting selection sets. The bound in Proposition 3 applies to full region preimages, as specified in Section 3.4.

B.6 MULTI-SCALE REFINEMENT

When a single cell holds thousands of variables, per-variable scoring supplies much of the selection detail. The optional refinement stage re-rasterises the bounding box of P_{sel} at the same $H \times W$ resolution and applies π_θ to that crop, increasing spatial detail within the selected region. The second heatmap replaces the first inside the box, and truncation proceeds as before. Refinement adds a forward pass at fixed resolution and a rasterisation whose cost depends on the variables in the crop. Corollary 18 describes the resolution threshold up to the constant c_g, which is fitted on one family for the refinement ablation.

B.7 TRAINING

The primary training signal imitates a local branching oracle. At each large neighborhood search step, the oracle solves an auxiliary program to find the best assignment within a Hamming ball around the incumbent. Positive labels cover only variables changed by the sub-solve, as described in Section 3.5. Because oracle calls are expensive, we supplement them with hindsight relabelling of traces collected under diverse behaviour policies. Each round's realised improvement is attributed to the released cells, and the network is trained with weighted binary cross-entropy against the resulting targets.

The optional reinforcement stage maximises improvement per unit time, accounting for the repair cost of larger neighborhoods. It uses REINFORCE with a leave-one-out baseline formed from G masks sampled at the same state. This unbiased baseline uses the sampled returns directly and supports the shared policy across instance scales without an additional value network.

B.8 BUDGET AND LOCALITY THRESHOLDS FOR THE DEGREE-OF-FREEDOM CERTIFICATE

Substituting representative parameters into Proposition 3 gives the budget and locality thresholds at which the certificate's sufficient condition holds. Consider a formulation in which half the rows are equalities, as in the degree constraints of a routing model and the item-assignment rows of a packing model, with $d = 5$ constraints per variable, mean support $s = 10$, equality fraction $\theta_E = 0.5$, density parameter $\kappa = 1$ and locality constant $c_\delta = 1$. Pure covering or packing programs have $\theta_E = 0$; the tight-inequality variant in Remark 10 provides a heuristic for these cases.

A random destroy set of size k meets about dk constraints, of which roughly $\theta_E dk = 2.5k$ are equalities. This estimate exceeds k throughout the budget range, leaving the rank to be measured directly at these parameters.

For a single compact block, the equality rows meeting S are a θ_E share of all rows meeting it, so the count is θ_E times the sum of two terms: the cut bound of Proposition 3, and the rows lying wholly inside. With one boundary component and area k/n the perimeter is $2\sqrt{\pi k/n}$, so the geometric term is $9\kappa m\delta\,\mathrm{per} = 18\sqrt{\pi}\,c_\delta\kappa\,d\sqrt{k/s} \approx 31.9\,d\sqrt{k/s}$, the second term is $9\kappa c_\delta^2 d = 45$, and the rows lying wholly inside number about dk/s. Multiplying by θ_E,

$$\text{rows} \;\approx\; 0.5\big(31.9\,d\sqrt{k/s} + 9d + dk/s\big) \;=\; 79.8\sqrt{k/10} + 22.5 + 0.25k.$$

At $k = 100$, the estimated count is about 300, above the certificate threshold. The condition rows $< k$ first holds at $k \approx 1.2 \times 10^3$; at $k = 100$, it also holds once $c_\delta \lesssim 0.3$. The reported runs use $k \le 10^3$, placing them in the latter regime: the layout's measured c_δ falls into that range on the families whose formulations contain equality rows. Budgets in the low thousands offer another route to the threshold and extend beyond the range evaluated here.

The certificate therefore identifies a threshold governed by budget and locality. In this example, the random row count is linear in k with a coefficient above one. The compact count combines a $\sqrt{k}$ boundary term with a smaller linear interior term and eventually falls below the budget. A row count below k certifies rank deficiency for the compact set. For random sets, a count above the budget leaves rank undetermined, so assessing their degrees of freedom requires measuring the rank itself.

C Experimental protocol in full

C.1 Instances

Table 7: Evaluation splits. *Size* denotes customers for routing, variables for mixed integer programs and operations for scheduling. *Seen* records whether the family appears in training; the two held-out families test zero-shot transfer of shared weights. Huang et al. (2023b) publish four synthetic families at 1,000 to 6,000 variables. We generate all four at 10^4 to align the main table with a point in the scaling sweep; this size difference is part of the protocol distinction from Table 12. Routing training reaches 2000 customers. XL, AGS and the million-scale sets test size extrapolation, while CVRPLIB X evaluates the training size range. The combinatorial auction family uses CATS (Leyton-Brown et al., 2000) with the parameters of Huang et al. (2023b).

Family	Split	Size	Count	Seen
CVRP	CVRPLIB X (Uchoa et al., 2017)	100–1000	100	
CVRP	CVRPLIB XL (Queiroga et al., 2026)	10^3–10^4	100	
CVRP	Arnold et al. (2019)	3×10^3–3×10^4	10	
CVRP	Accorsi & Vigo (2024)	2×10^4–10^6	20	
CVRP	clustered, distribution shift	5000	32	
CVRP	generated training set	$N \leq 2000$	256	yes
MILP	SC, MIS, CA, MVC after Huang et al. (2023b)	10^4	400	yes
MILP	the same at twice the size	2×10^4	400	yes
MILP	scaling tiers, SC and CA	10^5, 10^6	80	
MILP	MIPLIB 2017 (Gleixner et al., 2021)	mixed	240	
MILP	banded (staircase) programs	banded	20	
MILP	expander set cover (falsification)	10^3, 10^4	40	
JSSP	Taillard (Taillard, 1993)	≤ 2000	80	no
JSSP	Large-TA (Da Col & Teppan, 2022)	$\leq 10^5$	—	no
1D-BPP	BPPLIB (Delorme et al., 2018)	mixed	220	no

The four mixed integer families use the generators of Huang et al. (2023b): Barabási–Albert graphs for vertex cover, after Song et al. (2020); Erdős–Rényi graphs for independent set; CATS arbitrary relations for auctions (Leyton-Brown et al., 2000); and the set covering construction of Wu et al. (2021). Our implementations assemble the constraint matrix directly in sparse form for the 10^6 variable tier. Generator parameters follow the source papers, with the variable count adjusted to each size tier.

The scheduling transfer set reaches 10^5 operations under the selected CP-SAT sub-solver. Da Col & Teppan (2022) construct instances with up to 10^6 operations and report that CP-SAT returns no feasible solution on that group within six hours. An incumbent is required to initialise our LNS pipeline, so this group falls outside the evaluation with CP-SAT. The same study reports that CP Optimizer finds solutions for six of the ten million-operation instances on one core and all ten on four cores, indicating a route to extending the evaluation with another sub-solver. Taillard's set reaches 2000 operations; Large-TA supplies the larger instances.

C.2 Baselines

Learning to Delegate is evaluated at both subproblem sizes, $k = 5$ and $k = 10$, which define different action spaces. FILO2 is listed at both published effort levels: the long configuration uses a larger iteration budget and achieves a smaller gap (Table 14). Table 2 includes the short configuration; the long configuration's published iteration budget exceeds the 200 s deadline at every tier. FOVEA

Table 8: The seventeen baselines grouped by the aspect of performance they assess: solver configuration, destroy policy, encoder representation, domain performance and the contribution of learning to a spatial prior.

Layer	Methods	Comparison target
Solver defaults	SCIP default, SCIP aggressive (Bestuzheva et al., 2023)	the contribution of the LNS loop
Handcrafted destroy	random, RINS, local branching, LB-RELAX, adaptive portfolio	learned and fixed destroy rules
Learned destroy	CL-LNS, IL-LNS, L2D (k=5, k=10)	canvas and graph encoders
Domain solvers	HGS (Vidal, 2022), LKH-3 (Helsgaun, 2017), FILO2 (Accorsi & Vigo, 2024), CP-SAT (Perron et al., 2023)	performance relative to domain solvers
Spatial control	region-growing on the same canvas	**learning's contribution to the spatial prior**

and the spatial control use identical cached layouts, accessed through the same keys and checked by matching digests in the evaluation harness.

C.3 Additional results

Table 9: All four synthetic families at 10^4 variables and at twice that, primal integral over a 200 s compute-charged horizon. Best in each column in bold, ours shaded. The two families the main text reports are repeated here so that the four can be read together. The pattern on vertex cover and independent set matches the one on set cover and auctions: CL-LNS is ahead at the smaller tier and the margin is small.

Method	MVC		MIS		CA		SC	
	1×	2×	1×	2×	1×	2×	1×	2×
SCIP, default	39.7	57.2	71.4	93.8	74.9	96.4	58.4	79.3
SCIP, aggressive	32.1	48.6	60.3	82.5	63.1	84.7	46.2	66.8
Random LNS	18.4	26.9	15.2	22.7	68.2	82.6	34.7	45.1
RINS	16.8	24.3	14.6	21.1	59.4	73.5	31.9	42.3
Local branching	44.5	63.1	52.8	74.6	88.3	104.2	62.8	81.5
LB-RELAX	14.9	21.7	13.1	19.4	52.7	66.9	28.6	37.4
IL-LNS	12.3	18.9	11.7	17.2	47.6	60.8	25.3	33.8
CL-LNS	**9.8**	15.4	**9.4**	14.8	**41.2**	54.2	**21.4**	28.9
Region growing	11.6	16.2	10.9	15.3	49.8	61.3	24.1	30.2
FOVEA	10.7	**14.8**	10.2	**13.9**	43.5	**51.7**	22.0	**27.1**

C.4 Published reference points

The following tables provide published reference points under their source protocols. Their captions specify the horizons, hardware and reference values needed to interpret them alongside the matched comparisons in the main text.

Published routing results span two size ranges and two evaluation protocols, shown separately below with their original settings. Our 200 s runs appear in Table 2.

Routing gaps are reported with the date of the best-known-value snapshot; the snapshots in the two tables are four years apart. Since set I uses solver-generated reference values, Section 4 compares objective ratios against FILO2 at an equal deadline. The largest case is one million-customer instance, so the result at 10^6 describes that individual instance.

Table 10: MIPLIB 2017 under the present protocol: 300 s, 5 s per re-optimisation, SCIP 8.0.1, single threaded, three seeds. *Gap* $< 1\%$ counts instances reaching that gap within the horizon. Among the methods rerun under this protocol, FOVEA has the lowest mean gap and primal integral and the largest number of instances with gap below 1%. For context, Yuan et al. (2025) report a smaller gap of 1.75% for BTBS-LNS on the same instance set under the solver version and thread count listed in Table 13. BTBS-LNS was not rerun here, so the two protocols provide separate reference points.

Method	240 instances		
	mean gap %	gap $< 1\%$	P
SCIP, default	14.2	71	194.6
Random LNS	15.1	68	201.3
RINS	13.8	74	188.4
LB-RELAX	11.4	82	171.9
CL-LNS	8.9	96	152.7
Region growing	9.7	91	158.2
FOVEA	**8.4**	**99**	**148.3**

Table 11: Constants used in Proposition 21 for the set-cover estimate in Figure 2(b). The encoder slope $c_G d$ is fitted at the four sizes where CL-LNS runs and is also plotted in that panel. The quality ratio r is obtained from compute-matched runs at $n = 10^4$ with equal round counts. FOVEA's amortised slope is nearly five hundred times smaller than the encoder slope; substituting the constants satisfies the side condition of Proposition 21 and yields a predicted crossing.

Symbol	Meaning	Value	Source
τ_{net}	one forward pass, one CPU core	31 ms	measured
$\tau_{\text{lp}}/R_{\text{lp}}$	amortised relaxation refresh	12 ms	measured at $R_{\text{lp}} = 20$
c_{layout}	reordering, per variable	$0.9\,\mu$s	measured
c_{raster}	first rasterisation, per variable	$0.4\,\mu$s	measured
τ_{repair}	one sub-solve	1.8 s	measured
$c_G d$	graph encoder, per variable	5.9×10^{-6} s	fitted, Figure 2(a)
N	rounds in a run	106	$\lfloor (T - t_0)/\tau \rfloor$
c_1/N	FOVEA's own amortised slope	1.2×10^{-8} s	derived, Remark 22
r	quality ratio λ_G/λ_F	1.03	compute-matched at $n = 10^4$

C.5 Empirical checks motivated by the analysis

C.6 Ablations

The block-cap sweep follows Corollary 12. With $k = 100$ and $s \approx 10$ on set cover, the threshold $B^\dagger \asymp k/(c_\delta^2 s)$ places the useful range at a handful of blocks. The selected arms cover that range. The stronger condition $k \gtrsim B_\partial^2 s$ would instead place the estimated range below one block, illustrating the effect of the threshold's dependence on the block count.

C.7 Metrics and compute accounting

Alongside the primal integral in Section 2, the protocol defines the confined primal integral $P_\alpha(T) = \int_0^T p(t) e^{t/\alpha}\, dt$ with $\alpha < 0$ (Berthold & Csizmadia, 2021), which exponentially discounts late progress. Extending a run from T to $2T$ can change the ranking induced by P. The limit of P_α as $T \to \infty$ instead provides a finite, discounted measure of the full trajectory, although finite-horizon rankings can still change before that limit. The tables use P; comparison with P_α would assess sensitivity to the deadline.

A recorded trace supports three compute conventions. *Compute-charged* includes selection and repair time and is used for our tables and figures. *Compute-free* counts repair time alone, isolating behavior when selection overhead is excluded. *Compute-matched* fixes the round count and compares neighborhood quality separately from its cost; the quality ratio r in Table 11 uses this convention.

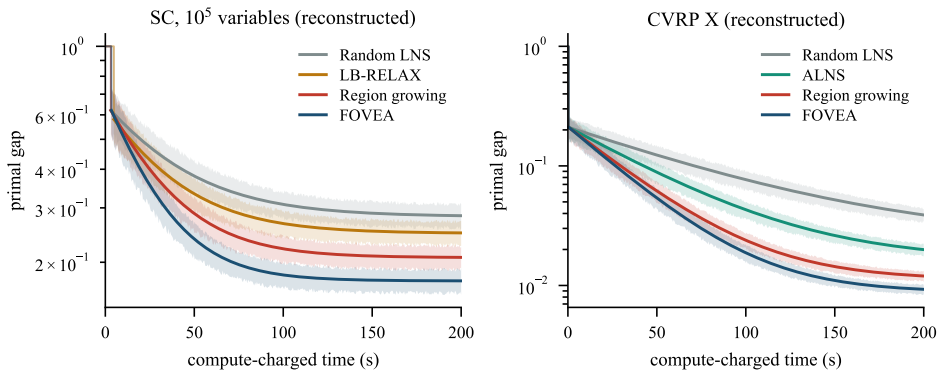

Figure 4: Illustrative primal-gap curves against compute-charged time. Exponential curves are fitted so that their areas match the aggregate primal integrals in Table 1 and Table 2. The shaded bands are generated illustrations, not empirical interquartile ranges over seeds. Each reconstructed curve includes an initial $p(t) = 1$ segment of duration t_0, which is longer in the set-cover panel and shorter than one second in the routing panel. This segment contributes t_0 to the integral, as defined in Section 2.

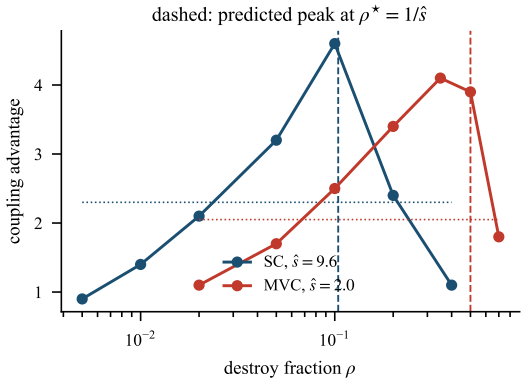

Figure 5: A destroy-fraction sweep motivated by the coupling envelope in Corollary 8. The empirical criterion is a peak exceeding both tails twofold, marked by the dotted lines; both plotted families meet it. On set cover, $\hat{s} = 9.6$ places the predicted peak $\rho^\star \approx 0.1$ within the proved range $\rho \leq 1/4$. Vertex cover provides a descriptive comparison outside that range: each row has support two, so its prediction $\rho^\star = 0.5$ falls outside Lemma 6. Its observed peak is one grid point away, at $\rho = 0.35$, within the reported seed spread. The sweep uses $n \leq 10^4$, where $\rho^\star$ is reachable; at larger scales, the fixed absolute budget places runs on the ascending branch (Section B.1).

Comparing the conventions separates the effect of selection latency from the improvement achieved per round.

C.8 Horizon, threading and statistics

The evaluation criteria were fixed before the runs: a primary target of a 10% primal-integral reduction over the best runnable baseline at large scales; a secondary target of staying within 10% of CL-LNS at the smaller scale; and a 5% reduction over region growing to assess the contribution of learning. The transfer comparison uses random selection as its reference. Limited benefit was expected on bin packing because its layout has limited metric structure.

The main experiments use a $200\,\mathrm{s}$ horizon to study the transient regime in which selection cost affects the primal integral. The MIPLIB evaluation uses the separate protocol stated in Table 10. Remark 24 distinguishes this transient behavior from the longer-horizon regime dominated by the plateau; hour-scale performance requires evaluation at that horizon. All methods in our evaluation run single threaded, keeping repair parallelism fixed when comparing selection costs.

Table 12: Synthetic mixed integer programs as published by Huang et al. (2023b): 60 min horizon, 2 min sub-solve, SCIP 8.0.1, 100 instances per family, mean primal integral ± standard deviation, lower is better. Table 9 gives our separate 200 s protocol at 10^4 variables. The methods shown are those reported by the source, which provides no RINS or Gurobi results for these families. GRAPH denotes its reimplementation of the learned partitioning operator of Song et al. (2020). †Local branching is quoted from the source appendix and uses a 10 min sub-solve, compared with 2 min for the other rows.

Method	MVC-S	MIS-S	CA-S	SC-S
As published (Huang et al., 2023b), 60 min, SCIP 8.0.1				
Branch and bound (SCIP)	66.1 ± 13.1	222.8 ± 25.9	137.4 ± 25.9	86.7 ± 37.9
Random LNS	38.0 ± 44.8	22.1 ± 5.0	235.6 ± 34.9	124.3 ± 45.4
GRAPH	42.9 ± 44.0	31.8 ± 5.0	277.7 ± 36.5	337.8 ± 76.4
LB-RELAX	57.0 ± 51.2	46.9 ± 6.5	140.5 ± 18.3	63.2 ± 31.6
IL-LNS	19.2 ± 10.2	19.4 ± 5.8	90.0 ± 20.8	63.2 ± 34.3
RL-LNS	29.6 ± 11.5	17.2 ± 5.2	249.2 ± 35.9	77.8 ± 28.9
CL-LNS	8.7 ± 6.7	12.8 ± 5.4	50.7 ± 22.7	26.2 ± 12.8
Local branching†	102.2 ± 35.9	130.9 ± 13.6	272.1 ± 26.9	113.7 ± 35.2

Table 13: MIPLIB 2017, 240 instances, mean primal gap in percent, as published by Yuan et al. (2025): 300 s horizon, 5 s per re-optimisation, SCIP 7.0.3 with Gurobi 9.5.0 as the commercial reference, four cores, three seeds. The second and third columns give SCIP at longer horizons. BTBS denotes the full method; BTBS-F uses the earlier bound-tightening scheme in place of binarised tightening. Gaps use the best solution found by any compared method as the reference. The absolute gaps shown imply a relative improvement of 11.6% between the two corresponding cells, compared with ten percent in the source abstract. Our evaluation of this instance set follows the separate protocol in Table 10.

	SCIP	600 s	900 s	U-LNS	R-LNS	FT-LNS	BTBS	BTBS-F	Gurobi
As published	15.15	11.08	8.79	16.26	15.94	13.07	1.75	3.11	1.98

Configurations in the main tables use five policy seeds. The ablation design specifies the seed counts in Table 17: five for studies supporting main claims, three for the remaining studies, and one for the study requiring an exact oracle solve in every arm. For stochastic sub-solvers, each policy seed is paired with a solver seed shared across methods. Aggregates are means and standard deviations over instances. Significance is assessed by a Wilcoxon signed-rank test on paired per-instance differences, with Holm correction across the methods in a table.

Table 14: CVRPLIB X, 100 instances with 100 to 1000 customers, mean percentage gap to the best known solution (lower is better). The upper block quotes Vidal (2022): a $2.4N$ second time limit, single threaded on an Intel Gold 6148 at 2.4 GHz, averages over ten runs and best known values from November 2020. That paper provides the original references for the competing heuristics; †SISR denotes its reimplementation. The lower block quotes Accorsi & Vigo (2024) on the same instances with an iteration budget, using an AMD Ryzen 5 PRO 4650GE at 3.3 GHz. FILO's reported gaps differ by 0.17 points across these protocols; comparisons should be made within each block.

Method	Protocol	Gap (%)	Time (s)
As published (Vidal, 2022), $2.4N$ s, ten runs			
HGS-CVRP	time limit, $2.4N$ s	0.11	
SISR†	time limit, $2.4N$ s	0.19	
FILO	time limit, $2.4N$ s	0.20	
HGS-2012	time limit, $2.4N$ s	0.21	
KGLS	time limit, $2.4N$ s	0.53	
HILS	time limit, $2.4N$ s	0.66	
LKH-3	time limit, $2.4N$ s	1.00	
OR-Tools	time limit, $2.4N$ s	4.01	
As published (Accorsi & Vigo, 2024), iteration budget, ten runs			
FILO	10^5 iterations	0.37	75
FILO2	10^5 iterations	0.34	76
FILO (long)	10^6 iterations	0.22	786
FILO2 (long)	10^6 iterations	0.20	801

Table 15: Large routing instances, mean percentage gap and mean wall clock, as published by Accorsi & Vigo (2024). Set B contains the ten instances of Arnold et al. (2019), with 3×10^3 to 3×10^4 customers. Set I contains the twenty instances introduced by Accorsi & Vigo (2024), with 2×10^4 to 10^6 customers. Set B gaps use CVRPLIB best known values. Set I gaps use the authors' best solutions from their experiments and therefore quantify distance to those solver-generated reference values.

Set	Method	Gap (%)	Time (s)
B, 10 instances	FILO2	1.08	121
B, 10 instances	FILO2 (long)	0.37	1371
I, 20 instances	FILO2	0.70	370
I, 20 instances	FILO2 (long)	0.30	3474
Lazio, 10^6 customers	FILO2	0.40	532
Lazio, 10^6 customers	FILO2 (long)	0.14	3938

Table 16: Analytical statements, associated empirical criteria and corresponding result locations. Trend and peak criteria test empirical hypotheses motivated by the bounds. The design corollaries Corollary 12, Corollary 13 and Corollary 18 determine ablation ranges in Section 3.6. The rank panel is illustrative; empirical evaluation of its criterion requires per-round rank and improvement logs.

Statement	Measurement	Empirical criterion	Result
Proposition 3	locality defect $\hat{\eta}$ against coupling advantage, within a family	explicit cut bound holds for admissible full preimages under the stated hypotheses	Figure 3
Corollary 8	advantage as a function of ρ at $n \leq 10^4$, on a family with $\hat{s} \geq 4$	empirical peak near $\rho \approx 1/\hat{s}$ exceeds both tails twofold	Figure 5
Proposition 1	per-round $\mathrm{rank}(A_{E(S),S})/k$ against realised improvement, on the families with equality rows	all full-rank rounds have zero improvement	Figure 3
Proposition 11	advantage over random on the expander family	coupling ratio at most $(1-2\epsilon)^{-1}$ under the hypotheses, $\epsilon < 1/2$	Table 3
Proposition 21	contraction rate and selection latency, giving a predicted critical size with an interval	agreement with the predicted interval, as an approximate cost-model check	Figure 2

Table 17: Ablation design: thirty-five studies and 117 arms in total. *Arms* counts configurations within a study, including the default. The specified seed counts are five for studies supporting main claims and three or one for the others. Numerical results supplied with the manuscript cover the five studies in Table 4; the remaining rows specify the broader evaluation design.

Study	Arms	Seeds	Sweep
A. Layout and canvas			
A1	4	5	reordering: Cuthill–McKee, spectral, random-index control, original-index control
A2	5	5	resolution $g \in \{32, 64, 128, 256, 512\}$
A3	6	5	drop each of the five channel groups in turn, against the full stack
A4	4	3	cell aggregation: mean, sum, max, softmax
A5	2	3	twelve channels against a single score channel
A6	2	3	the $1 \times HW$ strip with a 1-D U-Net, per Corollary 13
A7	2	5	fixed $g = 128$ against $g \propto \sqrt{n}$ after Chen et al. (2025), over four size tiers
A8	3	5	normalisation: per-instance, one global scale, none
B. Decoding and search			
B1	3	5	two-stage decoding against cell-level only and score only
B2	4	5	perimeter penalty $\lambda_{\mathrm{tv}} \in \{0, 0.01, 0.1, 1\}$
B3	3	5	mask shape: free-form, connected (default), a single rectangle
B4	4	3	Gumbel scale $\varepsilon \in \{0, 0.5, 1, 2\}$
B5	5	3	block cap $B_\partial \in \{\text{none}, 1, 2, 4, 8\}$
B6	3	3	truncation score: integrality gap, uniform, relaxation value
B7	3	5	budget schedule: fixed, adaptive growth, the online rule of Proposition 1
B8	7	5	$\rho \in \{0.01, 0.02, 0.05, 0.1, 0.2, 0.5\}$ at $n \leq 10^4$, and fixed k against fixed ρ
B9	3	1	the same three action spaces under an exact oracle, isolating their cost
B10	4	3	dual refresh $R_{\mathrm{lp}} \in \{1, 5, 20, \infty\}$
C. Architecture			
C1	2	3	U-Net against an FCN at equal parameter count
C2	4	3	capacity in $\{0.5, 1, 5, 20\}$M parameters
C3	4	3	depth in $\{2, 3, 4, 5\}$ at 1M parameters
C4	2	3	two CoordConv planes appended to the twelve channels
C5	3	5	GroupNorm, BatchNorm, LayerNorm
C6	4	3	refinement off, and on with trigger $\alpha \in \{2, 4, 8\}$
D. Training			
D1	3	5	imitation only, imitation then reinforcement, reinforcement only
D2	5	5	training set size in $\{64, 128, 256, 512, 896\}$
D3	3	5	oracle: exact local branching, LB-RELAX, a greedy selector
D4	4	3	hindsight relabelling window $w \in \{0, 1, 2, 4\}$
D5	2	5	single-family against joint multi-family training
D6	2	3	curriculum on instance size
E. Transfer			
E1	2	5	trained at $N \leq 2000$ customers, tested to 10^6
E2	3	5	zero-shot to the two held-out families
E3	2	5	uniform-trained against mixed-trained, on a clustered test set
E4	2	3	one shared checkpoint against a per-family fine-tune of it
E5	3	3	the expander family, testing the coupling-advantage limit in Proposition 11

D Related work

FOVEA draws on learned destroy operators, visual representations of combinatorial state, and shared models for multiple problem families.

D.1 Summary

The comparison follows three axes: the objects over which a destroy action is defined, the cost of encoding the search state, and the interface used to introduce a new problem family. FOVEA defines a cell mask on a fixed grid and decodes it into variables, with classical destroy signals supplied as input channels. The subsections below compare this interface with variable masks, visual operators, and shared models.

Two close visual approaches clarify the design choices: Chen et al. (2025) address scale transfer through resolution scaling, and Yin et al. (2026) select bounding boxes on rendered tours for exact repair. We use input normalisation on a fixed canvas and compare cell masks with rectangular regions. Section D.6 distinguishes our single-neighborhood analysis from the convergence result for the search in Burns et al. (2024).

D.2 Learned destroy operators

The variable-constraint incidence matrix supports several representations of a destroy decision. Song et al. (2020) arrange one row per integer variable with the incumbent value appended as a column, reduce it by PCA, and pass the result to a fully connected network that partitions the integer variables into k disjoint subsets. They train by multiclass classification against the best of several sampled decompositions and, in a second variant, by REINFORCE. Both this approach and Section 3.2 use the constraint matrix as input: row-wise PCA and a multilayer perceptron process it as feature vectors, while FOVEA reorders the matrix to encode adjacency and applies a convolutional network.

Subsequent methods commonly use a bipartite graph encoder in the manner of Gasse et al. (2019) and score each variable. Within this representation, local branching provides an imitation target (Sonnerat et al., 2021); it is also the expert used by FOVEA (Section B). Its computational cost motivates the relaxation-based surrogate of Huang et al. (2023a) and our use of hindsight relabelling. Other approaches use reinforcement learning from improvement rewards (Wu et al., 2021) or a contrastive objective over positive and negative neighborhoods (Huang et al., 2023b). The contrastive method reports the strongest results on the synthetic families of Table 12 under its evaluation protocol.

An extension of this representation uses a tripartite graph and a branching-guided component (Yuan et al., 2025). That work reports results on synthetic integer programs, as well as the MIPLIB 2017 results quoted in Table 13. The synthetic comparisons use different instance distributions and evaluation protocols, so the reported values do not establish a direct ranking across the two studies.

Routing also supports a coarser action space. Li et al. (2021) use a transformer to score $O(N)$ candidate route neighborhoods and select one subproblem, rather than score individual variables.

A per-variable scoring head has n outputs, while the routing selector above scores $O(N)$ candidates. The graph encoders discussed here incur $O(\mathrm{nnz}(A))$ cost per message-passing layer and store the graph and its activations on the device. These computation and memory costs can grow with instance size even when the destroy budget is fixed.

Xu et al. (2026) formalise the destroy set as a binary mask $m \in \{0,1\}^n$ with $m_i = 1$ when variable i is released. Their coordinatewise sampling treats variables independently; the masks come from classical rules or from the logits and gradients of a repair model. Their setting uses learned repair for constraint satisfaction. FOVEA trains a separate destroy policy and uses an exact repair formulation for mixed integer programming. It represents the mask as a field over a grid, where adjacency supports connectivity, perimeter, morphological post-processing and the isoperimetric analysis of Section 3.6.

D.3 Rasterised and visual combinatorial optimisation

Visual combinatorial optimisation uses rendered state for construction, diagnosis, and iterative improvement.

Zhao et al. (2021a) use a height map with a convolutional policy for online three-dimensional bin packing, using the natural grid geometry of the problem. The follow-up work of Zhao et al. (2021b) uses a packing configuration tree to represent placements beyond those reachable on a fixed-resolution grid. The resolution sweep in Table 17 examines the corresponding representation tradeoff for FOVEA.

Ling et al. (2020) render travelling salesman instances as images and predict a tour with a convolutional network, and Ling et al. (2021) do the same for 0-1 programs, solving them end to end. Both use the rendered instance to construct a solution. Steever et al. (2022) encode constraint matrices as images for structural analysis and instance classification. Samizadeh & Tong (2023) render a graph under a chosen layout and classify the image with a standard vision backbone to decide whether the graph is Hamiltonian, outperforming a Graphormer baseline. VN abbreviates vision-based neural. Their comparison of circular, spiral, and random layouts identifies layout as an important design choice. Our layout ablation examines its effect on neighborhood selection.

Chen et al. (2025) study resolution scaling in a model that fuses an instance image with graph features for multi-objective routing and knapsack. They increase the total pixel count linearly with instance size, so the image side length scales as $\sqrt{n}$. Their ablation reports a slight in-distribution improvement from this scaling and a much larger one out of distribution: without it, both their full model and a simpler single-stream fusion variant fall below CNH, their graph-based baseline. The out-of-distribution result identifies a scale-transfer problem relevant to a fixed canvas. Their diagnosis attributes the decline to a shift in occupied-pixel density, including density within a patch, relative to training. FOVEA uses input normalisation to address this shift (Section B), with resolution scaling included as a comparison in the resolution ablation.

Yin et al. (2026) prompt a vision-language model to propose axis-aligned bounding boxes on a rendered travelling salesman tour and repair the interior of each box exactly. This visual improvement operator is closely related to FOVEA. Their action space is a small number of rectangles, whereas a segmentation mask is an arbitrary field over cells. They use a general-purpose model through an API without task-specific training, with latency measured in seconds; FOVEA uses a small trained network with millisecond CPU latency. Their setting is one problem with native coordinates, while the canvas protocol is shared across four families. Remark 4 shows that axis-aligned squares attain the grid isoperimetric optimum at a given area. Free-form masks additionally represent destroy sets with multiple components, whose coupling cost is quantified by Corollary 12.

Goren et al. (2026) observe that rasterisation decouples inference cost from instance size. FOVEA formalises this observation, accounts for setup and update costs in Section 3.3, and evaluates it at 10^6 variables.

D.4 Constructive neural solvers

Constructive neural solvers generate an initial solution, typically by autoregressive decoding over an attention encoder (Kool et al., 2018), with rollout baselines exploiting problem symmetry (Kwon et al., 2020) or a heavy decoder for large-scale generalisation (Luo et al., 2023). FOVEA takes an incumbent as input. In the evaluated families, a cheap construction supplies this incumbent in a small fraction of the time budget, leaving most of the horizon for improvement. The operator comparisons therefore start from an existing solution. Strong domain solvers, including HGS and FILO2 for routing, appear in Table 8 and run end to end on the same clock.

D.5 Cross-problem models

Shared models for multiple combinatorial problems use several interface designs. Heterogeneous bipartite graphs, in the lineage of Gasse et al. (2019), describe a new family by adding node and edge types. Adapters (Drakulić et al., 2025) keep one backbone of mixed-attention blocks and give each family its own lightweight input and output modules over a shared codebook, so a new family is a new adapter. Token sequences (Zong et al., 2025) recast solving as a Markov decision process and

tokenise its trajectories, so a new family is a new tokenisation. Attribute vectors (Zhou et al., 2024; Berto et al., 2024) describe variants of one problem by switching features on and off, providing an interface for variants within routing. These approaches adapt the encoder interface to the problem family.

The canvas interface uses a fixed spatial grid whose adjacency encodes interaction. A new family supplies an entity layout and channel semantics while sharing the network parameters. The zero-shot transfer protocol in Section 4 evaluates these shared parameters on held-out families.

Two terminology distinctions clarify the comparison. Ouyang et al. (2026) learn to segment for vehicle routing by partitioning a tour into stable and unstable edge sequences. This is one-dimensional sequence labelling, while FOVEA predicts a field over a grid. The UniCO cited above is the work of Zong et al. (2025). A separate ICLR 2025 paper with the same name reduces problems to a matrix-encoded general TSP.

D.6 Machinery and theoretical foundations

The policy is a U-Net (Ronneberger et al., 2015) in the fully convolutional lineage of Long et al. (2015), with GroupNorm (Wu & He, 2018) to normalise samples independently within batches that mix instance scales. The sampler is the Gumbel-top-k construction of Kool et al. (2019), with exploration controlled by the Gumbel noise scale. Section B specifies the sampling distribution and the role of the noise scale.

Classical destroy heuristics supply input signals, training targets, and comparison operators. Relaxation-induced neighborhood search (Danna et al., 2005) releases the variables on which the incumbent and the relaxation disagree, providing channel 4 of Table 6. Local branching (Fischetti & Lodi, 2003) searches a Hamming ball and serves as our oracle. Adaptive operator portfolios (Pisinger & Ropke, 2007) choose among fixed rules by measured performance and serve as a baseline. These roles connect the canvas policy to established neighborhood-search methods.

On the theory side, Burns et al. (2024) give the first non-asymptotic convergence result for a hybrid large neighborhood local search, by reduction to a block Langevin diffusion. Their analysis uses a fixed partition of a continuous relaxation and covers randomised or cyclic block selection. Its stated scope excludes discrete decision variables and strongly non-convex potentials, including mixed integer programming. The analysis of Section 3.6 characterises the coupling and improvement conditions of a single neighborhood choice. Convergence of the resulting search trajectory remains a separate question.

E NOTATION

Symbol	Meaning	Symbol	Meaning
n, m	variables, constraints	φ	variable layout
s_j, s	support size, its mean	$\mathcal{E}, \mathrm{ent}$	entity set, association
s_{eff}	truncated mean support size	H, W, g	canvas height, width, resolution
d	constraints per variable	V_{pq}	variables in cell (p, q)
k, ρ	destroy budget, fraction	$\mathcal{C}^{(t)}$	the canvas at round t
$\rho^\star$	coupling-advantage envelope peak location	M	predicted heatmap
R, S_R	admissible region, its preimage	P_{sel}	selected cells
S	destroy set, $S \subseteq S_R$	$\mathcal{R}_{C,B}$	admissible region family
S_{rand}	uniform random k-subset	π_θ	the policy
$x^\star$	incumbent	R_{lp}	dual refresh period
$\mathrm{SUB}(S, x^\star)$	restricted subproblem	ε	Gumbel scale
$\Delta(S)$	best improvement from S	ϑ	entity mass threshold
$\mathrm{cut}(S)$	straddling constraints	$\tau_{\mathrm{select}}, \tau_{\mathrm{repair}}$	selection, repair time
$\mathrm{cut}_E(S), \mathrm{in}_E(S)$	equality rows cut, contained	τ	round time $\tau_{\mathrm{select}} + \tau_{\mathrm{repair}}$
θ_E	equality fraction	t_0	time to first feasible
δ, η	locality scale and defect (A1)	$P(T)$	primal integral
κ	density constant (A2)	$P_\alpha(T)$	confined primal integral
C, B	isoperimetric constant and boundary-component cap, (A3)	γ_t, γ_∞	primal gap, plateau
		L	Lipschitz constant (A4)
c_δ	locality constant, $\delta = c_\delta\sqrt{s/n}$	λ	gap contraction rate
c_g	resolution constant of $g^\star$	$n^\star$	critical instance size
B_∂	boundary components	$g^\star$	saturation resolution
B_{cc}	connected components	$\rho_{\max}$	maximum variable density
$B^\dagger, B^\ddagger$	block thresholds	Λ_∂	mean density on the boundary
$\mathrm{per}(R)$	$\mathcal{H}^1(\partial R)$	$\Lambda_{\max}$	global maximum density
ϵ	expansion defect, Proposition 11	ζ	budget efficiency
$B^\star$	boundary components of the minimiser in Definition 14	r	quality ratio λ_G/λ_F
		c_0, c_1	fixed and amortised per-round cost
g_0	reference resolution, $g_0 \gg g$	N	rounds in a run
β	oversample factor	α	confinement rate of P_α

The locality constant c_δ and the resolution constant c_g use distinct subscripts on c. The boundary component count B_∂, constrained by (A3), counts each boundary curve, while the connected component count B_{cc} counts connected regions: a region with one hole has one component and two boundaries. The mean density along a mask boundary Λ_∂ and the global maximum $\Lambda_{\max}$ also serve different roles. Substituting the latter into Corollary 18 overestimates the saturation resolution by $\sqrt{\Lambda_{\max}/\Lambda_\partial}$ on clustered instances where the uniformity assumption fails. The contraction rate in Proposition 21 is λ, the destroy fraction is ρ, the dual refresh period is R_{lp}, and the region is R. The predicted destroy fraction at the peak of the coupling-advantage envelope is $\rho^\star$. The symbol ϵ denotes the expansion defect in Proposition 11 and the relative error tolerance in Corollary 18. The entity mass threshold is ϑ, distinct from the policy parameter θ in π_θ.