SEPARATING FREE TOKEN GEOMETRY FROM RAW FEATURE-ENERGY ROUTING: AN OAPC DIAGNOSTIC STUDY

Anonymous authors
Paper under double-blind review

ABSTRACT

Fixed-grid tokenization couples uniform allocation with a factorized spatial geometry, so an apparent gain from adaptive placement may instead come from simply removing the grid constraint. We separate three questions: whether free-position geometry helps, whether content conditioning adds value beyond content-independent free geometry, and whether raw feature-energy routing is semantically useful. OAPC (Optimization-driven Adaptive Patch Centroids) is a deliberately non-amortized probe that optimizes N continuous centroids per input using raw-L_2-energy-weighted spatial quantization. Randomized synthetic controls show that geometric and content-conditioned gains are budget-dependent rather than universal. On a stratified 300-image COCO diagnostic subset, raw energy improves the same projected optimizer over constant-density routing in 23 of 24 backbone–budget settings, yet projected OAPC beats a Lloyd-optimized uniform centroidal Voronoi control in only one setting by mean per-image relative MSE reduction. In a 1,500-image, two-backbone mask-prediction stress test, raw-energy routing is statistically unresolved against uniform-CVT on ResNet-50 and underperforms it on ConvNeXt-T; channel standardization removes the resolved deficit without establishing an advantage. Diagnostic GT-mask routing improves micro-IoU through architecture-matched decoders trained under the same protocol on both backbones, showing that aligned guidance can exploit the continuous placement mechanism. Free positions are flexible, but raw feature magnitude is not a stable semantic routing signal under this spatial-quantization objective across the representations tested.

1 INTRODUCTION

Vision Transformers (Dosovitskiy et al., 2021) and their descendants (Liu et al., 2021; Carion et al., 2020) partition input images or intermediate feature maps into a fixed, uniform grid of non-overlapping patches. This design is simple and hardware-friendly, but it also hard-codes a spatial prior: every region receives the same initial token allocation regardless of local feature response. In practice, feature energy is rarely uniform. Objects, boundaries, textures, and clutter concentrate activation in only part of the feature map, while low-response regions may still consume the same number of grid-defined tokens. Fixed grids also impose a factorization constraint on the token budget: depending on N, the baseline may be forced into rectangular patches whose aspect ratio is poorly matched to the image geometry, while transposed factorizations can perform differently despite having the same aspect ratio.

A growing body of work addresses this inefficiency through *post-hoc token pruning*: tokens are first generated by a fixed grid and then selectively dropped (Rao et al., 2021; Meng et al., 2022; Liang et al., 2022; Kong et al., 2022) or merged (Bolya et al., 2023) based on learned attention scores. While effective, these methods retain the initial grid as a hard prior and discard tokens after the fact, rather than placing them where they are needed from the start.

We take a diagnostic approach organized around three falsifiable hypotheses. **H1 (geometry):** free-position tokens can represent non-grid-aligned structure better than a factorized rectangular grid. **H2 (content conditioning):** holding placement dynamics fixed, routing free positions by the

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.

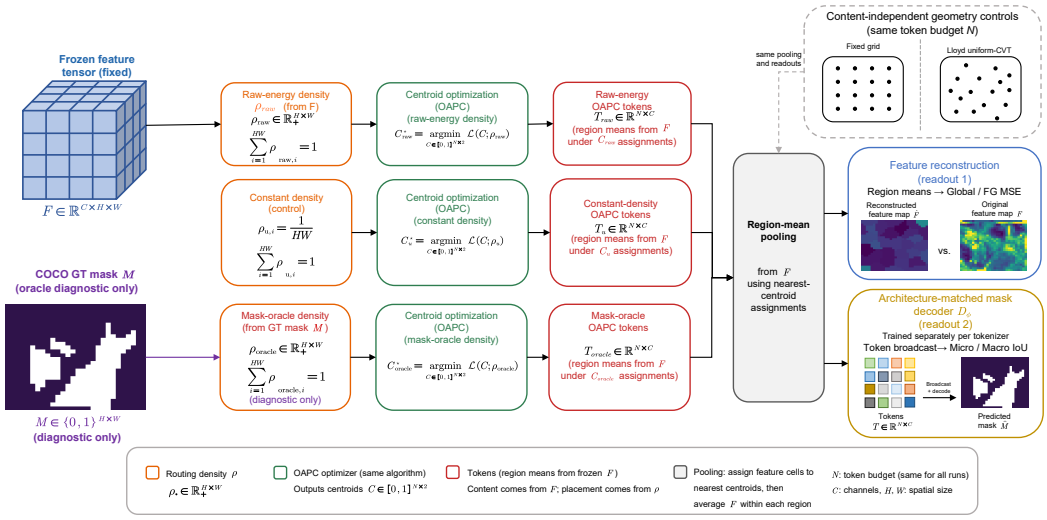

Figure 1: **Diagnostic pipeline.** A frozen feature map F defines a raw-energy density, a constant-density control, or a ground-truth mask oracle (diagnostic only). Each routing density is passed through the same OAPC optimizer under a fixed token budget N; hard nearest-centroid assignments induce regions, whose feature means form the tokens. Fixed-grid and Lloyd uniform-CVT provide content-independent geometry controls and use the same pooling and readout protocols. Mask decoders share architecture and training settings, but are trained separately for each tokenizer condition. The mask-output thumbnail is schematic, not a decoder prediction; feature thumbnails show feature-norm maps.

input's raw feature-energy field improves global reconstruction over constant-density routing. **H3 (semantic routing):** raw-energy-weighted spatial quantization places tokens where they are useful for object-mask prediction. The resulting system, **OAPC**, places N centroid tokens in continuous 2-D feature-map coordinates by minimizing an energy-weighted attraction loss together with a repulsion term. It has no shared learned parameters and uses no labels: centroid coordinates are optimized fresh for each input via 200 gradient steps, guided by the feature-energy field $\rho_i = \|\mathbf{F}_i\|_2/(\varepsilon_\rho + \sum_j \|\mathbf{F}_j\|_2)$. We use OAPC as a controlled probe, not a deployable tokenizer, and test each hypothesis against a matched control. Figure 1 summarizes the pipeline.

Contributions.

- A training-free, per-image continuous-centroid probe and randomized synthetic controls showing that H1 and H2 are budget-dependent rather than universal (§3, §4.1).
- Uniform-CVT and same-optimizer constant-density controls, a projected-gradient implementation check, and a full multi-backbone token-budget sweep that separate density from solver effects: raw energy improves its matched optimizer in 23/24 settings, but projected OAPC trails uniform-CVT in 23/24 (§4.2).
- A 1,500-image, two-backbone semantic stress test crossing three data splits, tokenizer seeds, and decoder seeds, together with a same-protocol, decoder-evaluated GT-mask oracle (§5).
- Cross-representation and normalization diagnostics showing that raw feature magnitude has representation- and normalization-dependent semantic alignment.

2 Related Work

Spatial quantization and centroidal Voronoi tessellations. Lloyd's algorithm (Lloyd, 1982) alternates nearest-center assignment and centroid updates for least-squares quantization; centroidal Voronoi tessellations (CVTs) generalize this view to spatial domains and nonuniform densities (Du et al., 1999), closely connecting it to classical vector quantization (Gersho & Gray, 1992). From this perspective, OAPC is a density-weighted spatial quantizer with differentiable assignments and local repulsion, while our uniform-CVT control applies Lloyd iterations to a uniform spatial density. Our focus is not a new general-purpose CVT solver, but using these complementary free-position constructions and a same-optimizer density control to separate geometry, routing signal, and solver quality.

Fixed-grid tokenization. ViT (Dosovitskiy et al., 2021) tiles the image into non-overlapping 16×16 patches; the Swin Transformer (Liu et al., 2021) begins with 4×4 patch partitioning and applies self-attention in shifted local windows; DETR (Carion et al., 2020) applies the same principle at the feature-map level. All three inherit the rectangular-grid constraint we study here.

Token pruning, merging, and halting. DynamicViT (Rao et al., 2021), AdaViT (Meng et al., 2022), and EViT (Liang et al., 2022) train predictors to drop or reorganize uninformative tokens after the initial grid embedding. A-ViT (Yin et al., 2022) adapts computation time by learning when spatial tokens should halt. Token Merging (ToMe) (Bolya et al., 2023) combines similar tokens using bipartite matching. These methods share a common prior: the grid is fixed and tokens are subsequently removed, reorganized, or merged. OAPC instead optimizes *where* to place tokens before any downstream processing.

Adaptive token selection and learned tokenization. TokenLearner (Ryoo et al., 2021) learns spatial tokenization maps that compress an image or video into a small set of informative tokens. CAT (Shen et al., 2025) instead adapts the total latent-token count to image-level complexity; it changes representation capacity across images rather than reallocating a fixed budget within each image. Adaptive Token Sampling (ATS) (Fayyaz et al., 2022) selects a variable subset of existing tokens inside a transformer using attention-derived scores. STViT (Chang et al., 2023) pools image tokens into learned semantic cluster centers, while MSViT (Havtorn et al., 2023) learns content-conditioned patch scales. PS-ViT (Yue et al., 2021) learns task-supervised sampling offsets, while Quadformer (Ronen et al., 2023) uses saliency-guided quadtree patches. These methods are close in spirit to adaptive tokenization, but they either learn the routing module or sample from an existing grid-token sequence. OAPC uses a fixed token budget, free continuous coordinates, and no offline training; its role here is to diagnose the consequences of routing by raw feature energy.

Free-position and irregular tokenization. DPT (Chen et al., 2021) learns image-dependent patch positions and scales end to end, whereas Token Pooling (Marin et al., 2023) selects irregular tokens to approximate an underlying continuous signal. TCFormer (Zeng et al., 2022) progressively clusters tokens into flexible shapes, and AiluRus (Li et al., 2023) uses spatial-aware density clustering to create adaptive-resolution regions for dense prediction. Superpixel tokenization (Lew et al., 2024) replaces rectangular patches with semantically coherent regions, and subobject-level tokenization (Chen et al., 2025) constructs arbitrary-shaped tokens aligned with object and part morphology. Differentiable hierarchical tokenization (Aasan et al., 2025) learns adaptive superpixel hierarchies, whereas GPSToken (Zhang et al., 2025) represents content-adaptive regions with continuous Gaussian position and shape parameters. Most closely, SPoT (Hjelkrem-Tan et al., 2025) samples fixed-size image patches at continuous subpixel positions and uses label-supervised, per-image gradient search to estimate an oracle classification ceiling. OAPC does not claim continuous placement itself as the novelty: it pools a frozen intermediate feature map and uses a training-free raw-energy objective to diagnose geometric confounds and the gap between spatial quantization and semantic routing.

Adaptive grouping and slot-based representations. Slot Attention (Locatello et al., 2020) iteratively binds a fixed set of learned slots to image regions via cross-attention, producing object-centric representations. Perceiver (Jaegle et al., 2021) cross-attends a small latent array to high-dimensional inputs, decoupling compute cost from input resolution. GroupViT (Xu et al., 2022) progressively merges image tokens into a smaller set of group tokens for open-vocabulary segmentation. These

methods learn *what* to group but start from a fixed spatial grid; OAPC instead optimises *where* to place centroids in continuous 2-D space before any grouping or attention step, and does so per-input without learned parameters.

Differentiable spatial resampling. Spatial Transformer Networks (STN; Jaderberg et al., 2015) learn per-image affine or thin-plate-spline warps to align features; this resamples the input at transformed sampling locations while retaining a regular output lattice and a fixed token count. Deformable convolution (Dai et al., 2017) offsets convolutional sampling points adaptively but is tied to a fixed-stride grid at a macro level. Deformable DETR (Zhu et al., 2021) samples sparse key locations around learned reference points inside a trained detection architecture. OAPC operates at the token-budget level: it places N free-position centroids with no fixed-stride constraint.

3 METHOD

3.1 PROBLEM SETUP

OAPC is a training-free, per-image diagnostic probe rather than a deployable inference-time tokenizer. Its non-amortized optimization is intentionally explicit so that geometry and routing can be controlled independently.

Let $\mathbf{F} \in \mathbb{R}^{M \times C}$ be a spatial feature map with $M = H \times W$ locations. Each location $\mathbf{x}_i \in [0,1]^2$ is a normalized 2-D coordinate. Define spatial feature energy:

$$\rho_i = \frac{\|\mathbf{F}_i\|_2}{\varepsilon_\rho + \sum_j \|\mathbf{F}_j\|_2}, \quad \varepsilon_\rho = 10^{-8}. \tag{1}$$

The constant guards against all-zero maps; otherwise $\sum_i \rho_i$ is numerically one. Raw norm changes under a common channel-wise feature translation whereas mean-reconstruction MSE does not, so it is a diagnostic proxy rather than a quantity derived from that metric. We place N learnable centroid coordinates $\{\mathbf{C}_j\}_{j=1}^N \subset [0,1]^2$, initialized uniformly at random in $[0,1]^2$, and optimize them to concentrate in high-ρ regions. (Uniform-grid initialization is also supported; see Appendix E for comparison.) The initial forward-clamped implementation stores unconstrained coordinates and applies $\mathbf{C}_j \leftarrow \mathrm{clip}(\mathbf{C}_j, 0, 1)$ in each forward pass. This simple truncated parameterization keeps sampled coordinates valid but can produce zero coordinate gradients outside the interval; we treat it as part of the fixed diagnostic implementation rather than an optimal constrained optimizer. Our projected-gradient control instead projects the centroid parameter into $[0,1]^2$ after every Adam step while leaving the objective and all other hyperparameters unchanged.

3.2 SOFT ASSIGNMENT

For each spatial location $\mathbf{x}_i$ we define the soft responsibility of centroid j:

$$p_{ij} = \frac{\exp(-\|\mathbf{x}_i - \mathbf{C}_j\|^2/\tau)}{\sum_k \exp(-\|\mathbf{x}_i - \mathbf{C}_k\|^2/\tau)}, \tag{2}$$

where τ is the temperature, annealed via a cosine schedule: $\tau(t) = \tau_f + \frac{1}{2}(\tau_0 - \tau_f)(1 + \cos(\pi t/(T-1)))$ from τ_0=1.0 to τ_f=0.10. Here $t = 0, \ldots, T-1$ indexes the T optimization steps. Low temperature sharpens each location's assignment toward its nearest centroid.

3.3 CENTROID OPTIMIZATION

Attraction drives centroids toward high-density regions:

$$\mathcal{L}_{\text{att}} = \sum_i \rho_i \sum_j p_{ij} \|\mathbf{x}_i - \mathbf{C}_j\|^2. \tag{3}$$

Intuitively, $\mathcal{L}_{\text{att}}$ is a density-weighted expected squared distance between each location and its responsible centroid. High-ρ regions exert stronger pull.

Repulsion discourages centroid collapse:

$$\mathcal{L}_{\text{rep}} = \sum_{a<b} \exp\left(-\frac{\|\mathbf{C}_a - \mathbf{C}_b\|^2}{\sigma_r^2}\right), \quad \sigma_r = \frac{\sqrt{2}}{N}. \tag{4}$$

The main experiments use $\sigma_r = \sqrt{2}/N$ as an empirical local-repulsion bandwidth. This is not a 2-D average-spacing estimate, which would scale as $O(N^{-1/2})$; instead, the $O(N^{-1})$ schedule makes repulsion increasingly local as the token budget grows. The intent is to prevent local centroid collapse without enforcing near-uniform spacing at large token budgets. Because this is an empirical solver choice rather than a spacing law, Appendix B reports a direct bandwidth sensitivity check.

Total loss:

$$\mathcal{L} = \mathcal{L}_{\text{att}} + \lambda_{\text{rep}} \cdot \mathcal{L}_{\text{rep}}, \quad \lambda_{\text{rep}} = 0.5. \tag{5}$$

We optimize with Adam (Kingma & Ba, 2015) at learning rate 0.05 for $T{=}200$ steps. No balance loss is used; see §3.5 for justification.

3.4 Training-Free Reconstruction

Centroid token quality is evaluated via a non-learned soft encode–decode cycle.

Encode (soft pooling): $\mathbf{z}_j = \sum_i p_{ij}\mathbf{F}_i / (\sum_i p_{ij} + \varepsilon)$.

Decode (soft broadcast): $\hat{\mathbf{F}}_i = \sum_j p_{ij}\mathbf{z}_j$.

For placement-quality comparison we use **OAPC(hard)**: the same optimized centroid positions with argmax assignment $a_i = \arg\max_j p_{ij}$, then $\mathbf{z}_j = \text{mean}_{a_i=j}(\mathbf{F}_i)$ and $\hat{\mathbf{F}}_i = \mathbf{z}_{a_i}$. This matches fixed-grid encode–decode and isolates centroid *placement* from soft blending. Equations (3)–(5) optimize the placement surrogate directly; reconstruction is evaluated afterward.

3.5 Design Decisions

We use symmetric repulsion and no load-balance penalty, keeping the placement objective limited to density-weighted attraction and local separation. This is a fixed probe configuration, not a claim of optimal solver design. Low-density centroids may contribute little to the placement loss while still owning large hard Voronoi regions; their reconstruction effect must therefore be measured separately. The coordinate-constraint and bandwidth controls below test two specific solver choices.

4 Experiments

4.1 Synthetic Feature-Map Validation

Setup. We generate $28{\times}28{\times}128$ fields inspired by the spatial resolution and channel structure of ResNet-50 C4. Four families vary independently over 12 field seeds: mixtures of 5–12 Gaussian peaks, translated and rotated stripes, translated and rotated crosses, and translated, rotated elliptical rings. We test $N{\in}\{8, 16, 32\}$ and centroid seeds 42–44. Controls are fixed-grid, content-independent uniform-CVT, feature-energy-weighted CVT, the forward-clamped OAPC, and an explicit projected-gradient OAPC control. Table 7 gives the parameter ranges. Reported intervals resample the 12 fields within each shape family 20k times and then average the four family means; positive values denote lower global MSE for the first method.

H1 and H2 are budget-dependent. Figure 2 breaks the aggregate result down by structure family and token budget. At $N{=}32$, uniform-CVT improves over fixed-grid by 0.84%, and projected OAPC improves by 1.79%, supporting free geometry on these randomized fields. Neither result transfers to $N{=}8$ or 16: free geometry is effectively tied at $N{=}8$, and both OAPC variants lose to fixed-grid at $N{=}16$. Content conditioning is similarly conditional. Projected OAPC beats uniform-CVT by 0.92% at $N{=}32$, but not at the lower budgets. Weighted CVT obtains small, consistent 0.67–1.25% gains, showing that the constructed energy field can aid spatial quantization under this reconstruction readout without establishing a representation- or data-independent advantage. This distinction also explains why the synthetic and COCO results need not agree: the synthetic fields are constructed so

Table 1: Randomized synthetic global-MSE controls. F, U, O_c, O_p, and K_ρ denote fixed-grid, uniform-CVT, forward-clamped OAPC, projected OAPC, and feature-energy-weighted CVT. Entries are mean paired reductions (%) with stratified field-bootstrap 95% CIs; positive is better.

N	U–F	O_c–F	O_p–F	O_p–U	K_ρ–U
8	$+.33$ $[-.18, +.89]$	$+.09$ $[-.38, +.56]$	$-.09$ $[-.51, +.31]$	$-.43$ $[-.86, -.05]$	$+.67$ $[+.29, +1.08]$
16	$-.22$ $[-.52, +.04]$	$-.85$ $[-1.37, -.35]$	$-.64$ $[-1.01, -.29]$	$-.42$ $[-.89, +.02]$	$+1.25$ $[+.78, +1.73]$
32	$+.84$ $[+.13, +1.55]$	$+1.39$ $[+.75, +2.03]$	$+1.79$ $[+1.13, +2.43]$	$+.92$ $[+.33, +1.56]$	$+.95$ $[+.37, +1.59]$

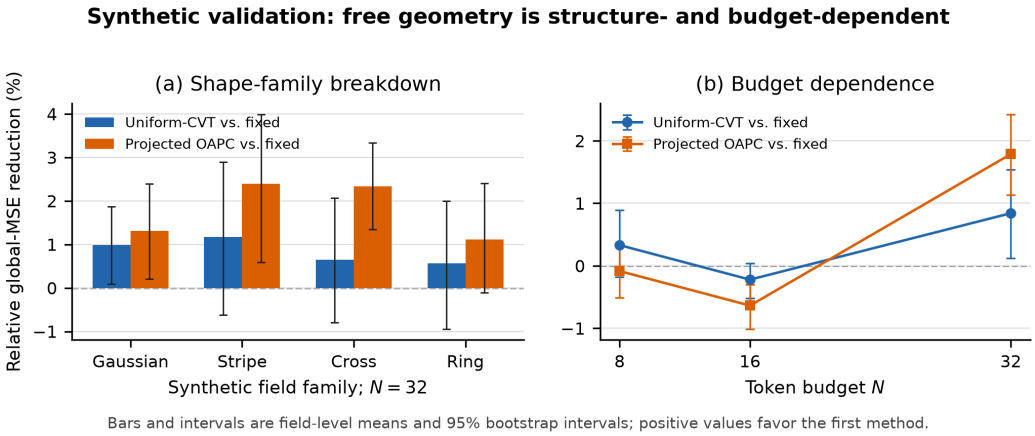

Figure 2: **Synthetic validation across structure families and token budgets.** (a) At N=32, both content-independent uniform-CVT and projected raw-energy OAPC improve over the fixed grid on average, but the magnitude varies by field family. (b) The advantage is budget-dependent and is not stable at N=16. Bars and intervals are field-level means and 95% bootstrap intervals; positive values favor the first method.

that scalar energy is aligned with reconstruction-relevant structure, whereas real backbone feature norms need not preserve that alignment.

The constraint implementation is not the main explanation. Forward clipping places 70.7/45.3/34.6% of centroids exactly on the boundary at N=8/16/32; explicit projection reduces these rates to 21.1/7.8/0.0%. Nevertheless, projection changes global MSE relative to forward clipping by only $-0.19/+0.19/+0.39\%$, respectively, and the first two intervals include zero (Appendix Table 8). Thus zero-gradient sticking is a real implementation weakness, but removing it does not create a uniform OAPC advantage.

4.2 COCO Feature-Map Validation

Setup. We evaluate frozen ResNet-50 C4 (1024 channels; He et al. 2016), ConvNeXt-Tiny C4 (384 channels; Liu et al. 2022), and ViT-B/16 block-8 features (768 channels; Dosovitskiy et al. 2021). All have stride 16 and spatial size 28×28 after directly resizing each image to 448×448; this diagnostic preprocessing does not preserve the original aspect ratio. Features are extracted from the same stratified 300-image COCO val2017 subset (Lin et al., 2014) (100 randomly sampled per category: *single* = 1 annotation, *few* = 2–3, *crowded* $\geq$ 8; seed 42). Images with 4–7 annotations are outside this deliberately separated diagnostic subset, so our COCO claims refer to this subset rather than to the full validation distribution. All methods use the same feature maps and reconstruct each spatial feature by the mean feature of its assigned token region. We compare five primary placement rules: (i) fixed-grid patches; (ii) *uniform-CVT*, a content-independent centroidal Voronoi tessellation; (iii) OAPC(hard), using optimized centroids with hard final assignment; (iv) ρ top-k, which places centers on the N highest-energy cells and then uses nearest-center regions; and (v) weighted k-means in 2-D coordinates, using ρ_i as the sample weight. We additionally rerun OAPC with post-step projected gradients to isolate the coordinate-constraint implementation. To isolate content density from all

Table 2: COCO global-MSE comparison at $N{=}32$ on the same 300 images. Fixed* is the favorable in-sample global orientation and is never selected per image; its comparisons are descriptive. $\Delta\mathrm{Fixed}^\star$ and ΔU-CVT are mean per-image relative error reductions; positive is better. OAPC(hard) uses forward-clamped coordinates. Stochastic entries are seed-level mean$\pm$sample SD (divisor $3-1$); the prose CIs instead average seeds per image before paired bootstrapping. Absolute MSE and ρ top-k are in Appendix Table 17.

Backbone	Method	win vs. Fixed*	$\Delta\mathrm{Fixed}^\star$	ΔU-CVT
ResNet-50	uniform-CVT	36.7%	-0.36%	—
	OAPC(hard)	$40.2\pm3.02\%$	$-0.39\pm0.14\%$	$-0.03\pm0.14\%$
	weighted k-means	$44.7\pm0.88\%$	$-0.21\pm0.03\%$	$+0.15\pm0.03\%$
ConvNeXt-T	uniform-CVT	66.3%	$+3.17\%$	—
	OAPC(hard)	$58.8\pm2.04\%$	$+1.80\pm0.24\%$	$-1.68\pm0.22\%$
	weighted k-means	$59.0\pm1.53\%$	$+1.42\pm0.25\%$	$-2.03\pm0.28\%$
ViT-B/16	uniform-CVT	48.3%	$+0.04\%$	—
	OAPC(hard)	$45.0\pm3.84\%$	$-0.31\pm0.45\%$	$-0.38\pm0.45\%$
	weighted k-means	$54.7\pm1.53\%$	$+0.67\pm0.11\%$	$+0.61\pm0.11\%$

other OAPC choices, we also run a matched constant-density control, denoted O_u: it replaces only ρ_i by $1/(HW)$ while retaining projected gradients, the same attraction–repulsion objective, temperature schedule, 200 Adam steps, and initialization seeds. Unlike Lloyd uniform-CVT, O_u changes the routing density without changing the solver. Uniform-CVT is the critical geometric control: it removes grid factorization while remaining independent of image content; the final two baselines test alternative energy-aware placement rules. Specifically, uniform-CVT applies Lloyd iterations to the same discrete $H\times W$ coordinate set with equal point weights, k-means++ initialization, at most 50 iterations, and tolerance 10^{-5}. We use 10 fixed-seed restarts $(7919,\ldots,7928)$, select the layout minimizing mean squared nearest-centroid distance, and cache it by (H,W,N). We report paired comparisons against both fixed-grid and uniform-CVT for a full 300-image sweep over $N\in\{16,25,30,32,36,42,49,64\}$. For nonsquare grids ($N\in\{30,32,42\}$), Fixed* denotes the lower-mean of the two transposed grids per backbone–budget pair on this diagnostic set, selected globally rather than per image; square budgets are unchanged. Because the same diagnostic set selects and evaluates this orientation, Fixed* is an intentionally favorable in-sample control: comparisons against it are descriptive, not inferential. O_u is the direct content-conditioning test because it changes only ρ; uniform-CVT is the stronger content-independent geometric benchmark, selected without image features (Appendix Table 6). For the stochastic placement methods, all reported reconstruction results are averaged over centroid-initialization seeds 42, 43, and 44; fixed-grid and ρ top-k are deterministic, as is the cached uniform-CVT layout.

The $N{=}32$ comparison is orientation-sensitive. Against the original 4×8 grid, OAPC appears successful on all three backbones. Transposing it to 8×4 lowers fixed-grid MSE by 1.77% on ResNet-50 and 1.65% on ViT-B/16; ConvNeXt retains 4×8. Against Fixed*, all free-position methods are worse on ResNet-50, and ViT's uniform-CVT result is a numerical tie ($+0.04\%$). ConvNeXt is the only backbone showing a clear free-geometry benefit: uniform-CVT improves by 3.17% and OAPC by 1.80%. These Fixed* comparisons remain descriptive because orientation was selected on the same diagnostic set. Relative to uniform-CVT, OAPC changes global MSE by -0.03% (paired-bootstrap 95% CI $[-0.13,+0.06]\%$), -1.68% ($[-2.34,-1.03]\%$), and -0.38% ($[-0.67,-0.08]\%$) on the three backbones, respectively. For these intervals, the three centroid-seed reconstructions are averaged per image before a 20k paired image bootstrap; positive means lower error. Thus, both factorization and orientation can make the fixed baseline appear weak; none of the three backbones provides evidence that feature-conditioned routing outperforms the stronger Lloyd-optimized content-independent geometry. The same-optimizer test below separates this baseline ranking from the contribution of the density signal itself.

Full N-scan. The stronger fixed-grid control sharpens the result: across 24 backbone–budget settings, forward-clamped OAPC beats Fixed* only at ConvNeXt, $N{=}32$, where U-CVT gains 3.17% versus 1.80%. Forward-clamped OAPC(hard) is lower than U-CVT only once (ResNet-50, $N{=}42$, by 0.02%, effectively a tie; Appendix Table 14).

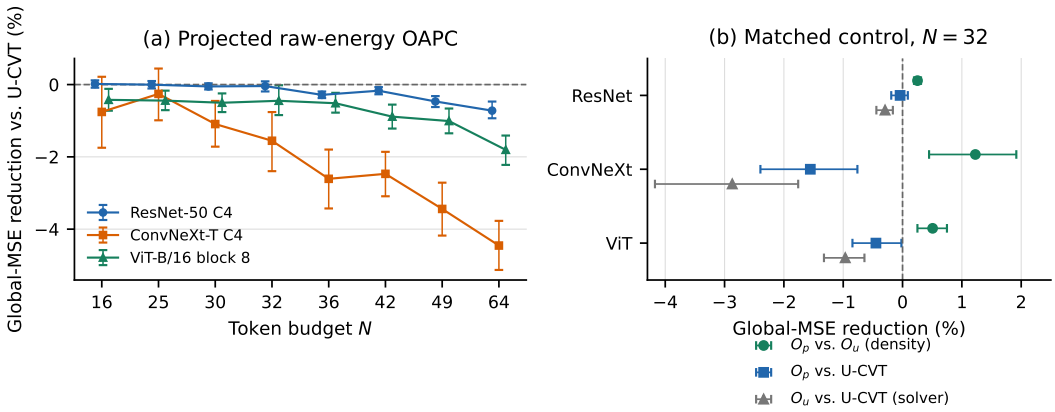

Figure 3: **COCO reconstruction controls.** (a) Full token-budget sweep: relative global-MSE reduction of projected raw-energy OAPC over Lloyd uniform-CVT across 24 backbone–budget settings. (b) Matched density control at N=32: the same projected optimizer gains modestly from raw-energy density over constant density ($O_p - O_u$), but the constant-density optimizer remains below the stronger Lloyd control ($O_u - U$). Positive values favor the first method; crossed intervals resample images and the three observed centroid seed IDs and describe sensitivity to those seeds.

The matched control separates density from solver quality. Projected raw-energy OAPC (O_p) improves over constant-density O_u in 23/24 settings, with 21 positive crossed CIs and 20 Holm-significant effects (Figure 3; Table 16); at N=32 the gains are 0.25%, 1.23%, and 0.51% for ResNet, ConvNeXt, and ViT. Yet O_u trails Lloyd U-CVT in all 24 settings, and O_p beats U only once (ResNet-50, N=16, by 0.02%), with no positive crossed CI; 18 crossed intervals are negative, of which 16 remain below the Holm-adjusted .05 threshold (Table 15). These intervals describe sensitivity to the three observed seeds, not uncertainty over a well-sampled population of initializations. All setting counts above use mean per-image relative reductions, weighting each image equally in percentage terms. Ranking by absolute dataset-mean MSE instead gives O_p two wins over U-CVT (ResNet N=16, ConvNeXt N=25); neither has a positive relative-reduction crossed interval. The high-energy metric favors OAPC mainly on ResNet-50 (7/8 budgets, 0.05–0.54%), loses consistently on ConvNeXt, and is mixed on ViT. Because this region is defined by the same raw ρ field, it is target-aligned; global and foreground-mask MSE are less method-aligned controls. Weighted k-means beats U-CVT at N=32 on ResNet-50 and ViT, while direct ρ top-k loses by 3.98%, 36.91%, and 7.03% on the three backbones (Appendix Table 17); full per-budget values are in Appendix Table 14.

Repulsion-bandwidth sensitivity. On ResNet-50 C4, the canonical $\sqrt{2}/N$ scale is near-tied with uniform-CVT at N=32 but remains below it at N=49; the nominal $O(N^{-1/2})$ alternative causes boundary concentration and empty regions. The full sensitivity check is in Appendix Table 4; conclusions remain conditional on this solver protocol.

5 Discussion and Misalignment Analysis

For each tokenizer condition, we separately train an architecture-matched token-broadcast decoder for union-of-instance masks on frozen ResNet-50 C4 and ConvNeXt-T C4 features. The stress test uses three fixed 999/501 train/validation repartitions of one sampled 1,500-image pool. At N=32, it compares fixed orientation, uniform-CVT, and projected OAPC with raw, standardized, or diagnostic GT-mask density. All use hard pooling, 20 epochs, and seeds 42–44 (198 runs; Appendix Table 11). We report micro-IoU, nonempty-image macro-IoU, and empty-mask false-positive rates. Crossed intervals (20k replicates) resample images and all stochastic axes; they measure sensitivity within this pool, not confidence over independent dataset replications. Appendix Tables 20–21 give the condition means, empty-mask errors, and paired intervals.

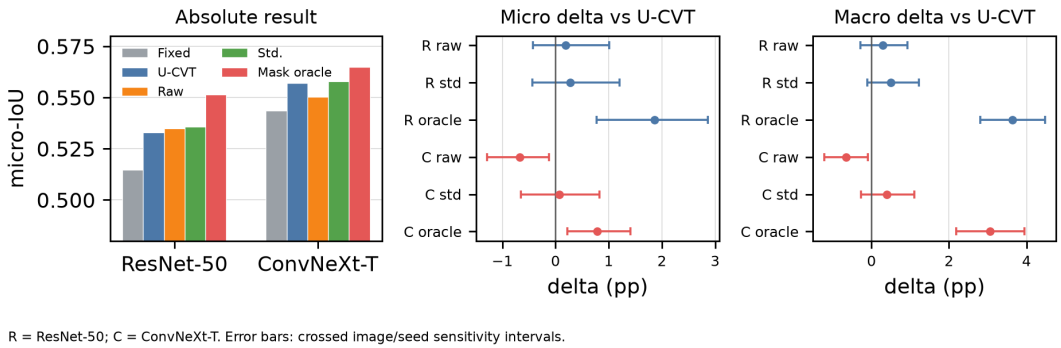

Figure 4: Decoder-connected semantic stress test at $N{=}32$: absolute micro-IoU (left) and changes from uniform-CVT with crossed image–split–centroid–decoder sensitivity intervals (center/right). Raw-energy routing is representation-dependent; mask-aligned routing improves both backbones.

Raw-energy OAPC shows no clear advantage over content-independent free geometry (Figure 4): its ResNet-50 difference from uniform-CVT is unresolved ($+.0019$ micro-IoU, CI $[-.0042, +.0101]$), while ConvNeXt-T loses .0067 (CI $[-.0128, -.0012]$; macro-IoU loss .0066). Standardized routing also has intervals spanning zero on both backbones, so the effect depends on representation and normalization.

Oracle routing provides a foreground-placement reference. Diagnostic GT-mask density (foreground/background weights 1/.01) improves micro/macro IoU over uniform-CVT by $+.0186/+.0364$ on ResNet-50 and $+.0079/+.0305$ on ConvNeXt-T; all four crossed CIs are above zero. This mask-defined route is an achievable-performance witness, not a deployable optimum; the geometric oracle improves foreground MSE by 7.7% at nearly unchanged global MSE (Appendix Table 22).

Centroid energy preference is representation-dependent. OAPC raises centroid energy percentile by 0.028–0.096 on the ResNet and ConvNeXt representations, while ViT is essentially unchanged (-0.001) relative to fixed anchors (Appendix Table 12). This forward-clamped placement diagnostic describes energy preference; it does not by itself establish semantic utility.

Energy/object alignment depends on representation and normalization. Top-5% hotspots are enriched on ResNet C3–C5 (44.9–50.1% hit versus 24.9–25.2% mask area), anti-aligned on ConvNeXt C4 (8.0%), and below mask area on ViT block 8 (14.4%). The matched standardization diagnostic changes hit rates from 48.7/8.0/14.2% to 50.7/42.1/23.8% for ResNet/ConvNeXt/ViT, respectively. For ViT, improved hotspot alignment does not yield a clear foreground-coverage gain over uniform-CVT, and global MSE remains 0.71% higher (Appendix Table 13). Alignment therefore depends on representation and normalization, while placement also depends on the objective (Appendix Figure 5).

6 Conclusion

Using OAPC, we separate free geometry, content conditioning, and semantic routing. On the 300-image COCO diagnostic subset, raw energy improves its matched control in 23/24 settings but surpasses Lloyd uniform-CVT in only 1/24 by mean per-image relative MSE reduction, with no positive crossed CI. The semantic stress test is unresolved on ResNet-50 and negative on ConvNeXt-T; standardization only closes the latter gap, whereas mask-oracle routing improves both micro- and macro-IoU. These results favor task-aligned routing over raw feature norm.

Limitations. This non-task-scale stress test uses one layer per backbone, a stratified square-resized COCO subset, and a target-leaking oracle. Its matched control isolates ρ, not solver quality; the bandwidth check further shows that the exact repulsion scale affects the optimizer. Conclusions are therefore limited to the tested raw-L_2 spatial quantizer and its explicitly reported solver protocol.

AI USE STATEMENT

During the preparation of this work, the authors used OpenAI Codex to assist with language editing, manuscript organization, related-work review, consistency checking, code review, and the refinement and extension of revision-stage experimental configurations and analysis scripts. The research questions, core methodology, experimental framework, and original software implementation were developed by the authors. Codex also assisted in reviewing experimental methodology, interpreting results, and designing additional diagnostic controls, including the design and implementation of seeded mathematical synthetic-field generators. These fields were produced by procedural programs rather than sampled from a generative model. All AI-assisted code, formulas, analyses, citations, and text were reviewed and verified by the authors, who take full responsibility for all claims and artifacts in this submission. Quantitative figures were plotted from experimental data, and the conceptual workflow figure was manually drawn by the authors as a vector graphic.

REPRODUCIBILITY STATEMENT

The supplementary code archive contains the OAPC implementation, frozen-feature extraction and reconstruction pipelines, semantic-control pipeline, statistical summaries, and figure-generation scripts. The main text and Appendix Tables 7, 11, 16, and 4 specify the data construction, preprocessing, seeds, optimizer settings, and uncertainty protocols; the archive also contains the commands used to reproduce the reported controls.

REFERENCES

Marius Aasan, Martine Hjelkrem-Tan, Nico Catalano, Changkyu Choi, and Adín Ramírez Rivera. Differentiable hierarchical visual tokenization. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Daniel Bolya, Cheng-Yang Fu, Xiaoliang Dai, Peizhao Zhang, Christoph Feichtenhofer, and Judy Hoffman. Token merging: Your ViT but faster. In *International Conference on Learning Representations*, 2023.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision*, pp. 213–229, 2020.

Shuning Chang, Pichao Wang, Ming Lin, Fan Wang, David Junhao Zhang, Rong Jin, and Mike Zheng Shou. Making vision transformers efficient from a token sparsification view. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6195–6205, 2023.

Delong Chen, Samuel Cahyawijaya, Jianfeng Liu, Baoyuan Wang, and Pascale Fung. Subobject-level image tokenization. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 7719–7738, 2025.

Zhiyang Chen, Yousong Zhu, Chaoyang Zhao, Guosheng Hu, Wei Zeng, Jinqiao Wang, and Ming Tang. DPT: Deformable patch-based transformer for visual recognition. In *Proceedings of the 29th ACM International Conference on Multimedia*, pp. 2899–2907, 2021. doi: 10.1145/3474085.3475467.

Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 764–773, 2017.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *International Conference on Learning Representations*, 2021.

Qiang Du, Vance Faber, and Max Gunzburger. Centroidal Voronoi tessellations: Applications and algorithms. *SIAM Review*, 41(4):637–676, 1999. doi: 10.1137/S0036144599352836.

Mohsen Fayyaz, Soroush Abbasi Koohpayegani, Farnoush Rezaei Jafari, Sarthak Sengupta, Hamid Reza Vaezi Joze, Eric Sommerlade, Hamed Pirsiavash, and Juergen Gall. Adaptive token sampling for efficient vision transformers. In *European Conference on Computer Vision*, 2022.

Allen Gersho and Robert M. Gray. *Vector Quantization and Signal Compression*. Kluwer Academic Publishers, 1992. doi: 10.1007/978-1-4615-3626-0.

Jakob Drachmann Havtorn, Amélie Royer, Tijmen Blankevoort, and Babak Ehteshami Bejnordi. MSViT: Dynamic mixed-scale tokenization for vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pp. 838–848, 2023.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

Martine Hjelkrem-Tan, Marius Aasan, Gabriel Y. Arteaga, and Adín Ramírez Rivera. SPoT: Subpixel placement of tokens in vision transformers. In *ICCV Workshop on Efficient Computing under Limited Resources: Visual Computing*, 2025.

Max Jaderberg, Karen Simonyan, Andrew Zisserman, and Koray Kavukcuoglu. Spatial transformer networks. In *Advances in Neural Information Processing Systems*, 2015.

Andrew Jaegle, Felix Gimeno, Andrew Brock, Andrew Zisserman, Oriol Vinyals, and João Carreira. Perceiver: General perception with iterative attention. In *International Conference on Machine Learning*, pp. 4651–4664, 2021.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *International Conference on Learning Representations*, 2015.

Zhenglun Kong, Peiyan Dong, Xiaolong Ma, Xin Meng, Wei Li, Wei Niu, Mengshu Sun, Xuan Shen, Geng Yuan, Bin Ren, et al. SPViT: Enabling faster vision transformers via latency-aware soft token pruning. In *European Conference on Computer Vision*, 2022.

Jaihyun Lew, Soohyuk Jang, Jaehoon Lee, Seungryong Yoo, Eunji Kim, Saehyung Lee, Jisoo Mok, Siwon Kim, and Sungroh Yoon. Superpixel tokenization for vision transformers: Preserving semantic integrity in visual tokens. *arXiv preprint arXiv:2412.04680*, 2024.

Jin Li, Yaoming Wang, Xiaopeng Zhang, Bowen Shi, Dongsheng Jiang, Chenglin Li, Wenrui Dai, Hongkai Xiong, and Qi Tian. AiluRus: A scalable ViT framework for dense prediction. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Youwei Liang, Chongjian Ge, Zhan Tong, Yibing Song, Jue Wang, and Pengtao Xie. Not all patches are what you need: Expediting vision transformers via token reorganizations. In *International Conference on Learning Representations*, 2022.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common objects in context. In *European Conference on Computer Vision*, pp. 740–755, 2014.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10012–10022, 2021.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A ConvNet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11976–11986, 2022.

Stuart P. Lloyd. Least squares quantization in PCM. *IEEE Transactions on Information Theory*, 28 (2):129–137, 1982. doi: 10.1109/TIT.1982.1056489.

Francesco Locatello, Dirk Weissenborn, Thomas Unterthiner, Aravindh Mahendran, Georg Heigold, Jakob Uszkoreit, Alexey Dosovitskiy, and Thomas Kipf. Object-centric learning with slot attention. In *Advances in Neural Information Processing Systems*, 2020.

Dmitrii Marin, Jen-Hao Rick Chang, Anurag Ranjan, Anish Prabhu, Mohammad Rastegari, and Oncel Tuzel. Token pooling in vision transformers for image classification. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 12–21, 2023.

Lingchen Meng, Hengduo Li, Bor-Chun Chen, Shanghang Lan, Zuxuan Wu, Yu-Gang Jiang, and Ser-Nam Lim. AdaViT: Adaptive vision transformers for efficient image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12309–12318, 2022.

Yongming Rao, Wenliang Zhao, Benlin Liu, Jiwen Lu, Jie Zhou, and Cho-Jui Hsieh. DynamicViT: Efficient vision transformers with dynamic token sparsification. In *Advances in Neural Information Processing Systems*, 2021.

Tomer Ronen, Omer Levy, and Avram Golbert. Vision transformers with mixed-resolution tokenization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 4613–4622, 2023.

Michael S. Ryoo, AJ Piergiovanni, Anurag Arnab, Mostafa Dehghani, and Anelia Angelova. TokenLearner: What can 8 learned tokens do for images and videos? In *Advances in Neural Information Processing Systems*, 2021.

Junhong Shen, Kushal Tirumala, Michihiro Yasunaga, Ishan Misra, Luke Zettlemoyer, Lili Yu, and Chunting Zhou. CAT: Content-adaptive image tokenization. In *Advances in Neural Information Processing Systems*, volume 38, pp. 830–860, 2025. doi: 10.52202/085713-0026.

Jiarui Xu, Shalini De Mello, Sifei Liu, Wonmin Byeon, Thomas Breuel, Jan Kautz, and Xiaolong Wang. GroupViT: Semantic segmentation emerges from text supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18134–18144, 2022.

Hongxu Yin, Arash Vahdat, Jose M. Alvarez, Arun Mallya, Jan Kautz, and Pavlo Molchanov. A-ViT: Adaptive tokens for efficient vision transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

Xiaoyu Yue, Shuyang Sun, Zhanghui Kuang, Meng Wei, Philip H. S. Torr, Wayne Zhang, and Dahua Lin. Vision transformer with progressive sampling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 387–396, 2021.

Wang Zeng, Sheng Jin, Wentao Liu, Chen Qian, Ping Luo, Wanli Ouyang, and Xiaogang Wang. Not all tokens are equal: Human-centric visual analysis via token clustering transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11101–11111, 2022.

Zhengqiang Zhang, Rongyuan Wu, Lingchen Sun, and Lei Zhang. GPSToken: Gaussian parameterized spatially-adaptive tokenization for image representation and generation. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable DETR: Deformable transformers for end-to-end object detection. In *International Conference on Learning Representations*, 2021.

A HYPERPARAMETER TABLE

Table 3: Default OAPC hyperparameters used in the main experiments. Appendix B reports the explicit repulsion-bandwidth sensitivity.

Parameter	Value	Notes
τ_0 (initial temperature)	1.0	cosine schedule start
τ_f (final temperature)	0.10	cosine schedule end
σ_r (repulsion bandwidth)	$\sqrt{2}/N$	empirical local-repulsion bandwidth
λ_{rep} (repulsion weight)	0.5	
λ_{bal} (balance weight)	0.0	disabled (§3.5)
Learning rate	0.05	Adam
T (optimization steps)	200	

B REPULSION-BANDWIDTH SENSITIVITY

The default $\sigma_r = \sqrt{2}/N$ is an empirical local-repulsion scale, not a claim about the typical spacing of N points in two dimensions. We therefore reran the ResNet-50 C4 reconstruction protocol on the same 300 images at $N \in \{32, 49\}$, using three centroid seeds and the same optimizer, temperature schedule, repulsion weight, and 50-step Lloyd control. We tested multiplicative scales $0.5\sqrt{2}/N$, $\sqrt{2}/N$, and $2\sqrt{2}/N$, together with the alternative $\sqrt{2/N}$. No hyperparameter was retuned for this check.

Table 4 reports mean relative MSE reductions over the three centroid seeds; positive values favor OAPC. The canonical scale is close to uniform-CVT at $N{=}32$ but remains below it at $N{=}49$. The $O(N^{-1/2})$ alternative produces severe boundary concentration and empty regions under the forward-clamped implementation. The smaller and larger $O(N^{-1})$ constants change the effect size, but none yields a systematic advantage over Lloyd uniform-CVT.

Table 4: Repulsion-bandwidth sensitivity on the 300-image ResNet-50 C4 diagnostic subset. Entries average per-image relative global-MSE reductions of OAPC against the reference; positive is better for OAPC. Values average centroid seeds 42–44. Boundary is the mean fraction of centroids at the coordinate boundary and empty is the mean number of empty hard-assignment regions.

N	σ_r	OAPC–Fixed	OAPC–U-CVT	boundary	empty
32	$0.5\sqrt{2}/N$	$+0.68\%$	-0.79%	26.1%	0.00
32	$\sqrt{2}/N$	$+1.43\%$	-0.03%	34.8%	0.00
32	$2\sqrt{2}/N$	$+1.01\%$	-0.46%	53.1%	0.00
32	$\sqrt{2/N}$	-2.61%	-4.13%	80.2%	6.00
49	$0.5\sqrt{2}/N$	-1.87%	-1.26%	23.7%	0.04
49	$\sqrt{2}/N$	-0.96%	-0.35%	36.9%	0.00
49	$2\sqrt{2}/N$	-0.66%	-0.06%	43.6%	0.00
49	$\sqrt{2/N}$	-5.01%	-4.39%	71.4%	7.67

C Fixed-Grid Factorization and Bad-N Values

The fixed-grid baseline decomposes N as follows: set $r = \lfloor\sqrt{N}\rfloor$, decrement r until $N \bmod r = 0$, then $c = N/r$. For a 28×28 feature map, the patch aspect ratio is $\text{AR} = \max(H/r, W/c)/\min(H/r, W/c)$.

Table 5: Fixed-grid factorization for selected N values on a 28×28 input. "Bad" N values (AR ≥ 2) highlighted.

N	rows	cols	patch size	AR
16	4	4	7.0×7.0	1.00
25	5	5	5.6×5.6	1.00
30	5	6	5.6×4.7	1.20
32	4	8	7.0×3.5	2.00
36	6	6	4.7×4.7	1.00
42	6	7	4.7×4.0	1.17
49	7	7	4.0×4.0	1.00
64	8	8	3.5×3.5	1.00
8	2	4	14.0×7.0	2.00
18	3	6	9.3×4.7	2.00
40	5	8	5.6×3.5	1.60

Continuous centroid positions are not factorized into rows and columns. However, our controls show that a fixed-grid comparison depends not only on factorization and patch AR, but also on which global orientation is chosen for the same factor pair.

Table 6: Global fixed-grid orientation control on the same 300 images. "Swap Δ" is the mean per-image relative MSE reduction (%) from transposing the default grid; positive favors the transpose. Fixed* is the lower-mean global orientation for each backbone–budget pair, never an orientation selected per image. Because it is selected on these same images, Fixed* is a favorable descriptive control. O denotes forward-clamped OAPC with hard pooling.

Backbone	N	Default → swap	Fixed*	Swap Δ	U–F*	O–F*
ResNet-50	30	5×6 → 6×5	6×5	+0.60	−0.27	−0.50
	32	4×8 → 8×4	8×4	+1.77	−0.36	−0.39
	42	6×7 → 7×6	7×6	+0.39	−0.52	−0.50
ConvNeXt-T	30	5×6 → 6×5	6×5	+1.14	−0.78	−2.41
	32	4×8 → 8×4	4×8	−1.77	+3.17	+1.80
	42	6×7 → 7×6	6×7	−2.43	−2.13	−3.81
ViT-B/16	30	5×6 → 6×5	6×5	+0.78	−0.32	−0.96
	32	4×8 → 8×4	8×4	+1.65	+0.04	−0.31
	42	6×7 → 7×6	6×7	−0.03	+0.13	−0.40

D PER-EXPERIMENT SETUP

Table 7: Synthetic feature-map validation (§4.1) setup.

Parameter	Value
Feature size $H \times W \times C$	$28 \times 28 \times 128$
Field families	Gaussian mixtures, stripe, cross, elliptical ring
Random fields	12 per family (seeds 1000–1011; 48 total)
Gaussian parameters	$K \in [5, 12]$; $\sigma \in [.035, .11]$; random centers/amplitudes
Structured-shape parameters	random translation, rotation, width, radius/aspect
Token budgets N	$\{8, 16, 32\}$
ρ definition	$\|\mathbf{F}_i\|_2/(\Sigma + \varepsilon_\rho)$ (Eq. 1)
Metrics	global/high-energy/foreground MSE; foreground geometry; boundary rate
Methods	fixed, U-CVT, weighted CVT, OAPC-clip, OAPC-projected
τ_0/τ_f / anneal	1.0 / 0.10 / cosine
λ_{rep}	0.5
Learning rate / T	0.05 / 200
Centroid seeds	42–44
Uncertainty	20k field bootstrap, stratified by family

Table 8: Synthetic coordinate-constraint control over 48 fields. Boundary is the fraction of centroids exactly on the square boundary. Δ is projected OAPC relative to forward-clamped OAPC in global MSE; positive favors projection. CIs use the stratified field bootstrap from Table 1.

N	forward-clamp boundary	projected boundary	Δ MSE (95% CI)
8	70.7%	21.1%	$-.19\%$ $[-.50, +.12]$
16	45.3%	7.8%	$+.19\%$ $[-.23, +.62]$
32	34.6%	0.0%	$+.39\%$ $[+.04, +.72]$

Table 9: COCO coordinate-constraint control. Entries are projected-gradient OAPC relative to uniform-CVT in mean per-image global-MSE reduction (%), with 20k paired image-bootstrap 95% CIs after averaging centroid seeds within image; positive favors projection. No interval is significantly positive, while 18 negative intervals remain significant after Holm correction over the 24 settings.

N	ResNet-50 C4	ConvNeXt-T C4	ViT-B/16 block 8
16	$+0.02$ $[-0.05, +0.09]$	-0.76 $[-1.52, +0.01]$	-0.42 $[-0.61, -0.23]$
25	-0.00 $[-0.07, +0.07]$	-0.26 $[-0.90, +0.38]$	-0.44 $[-0.64, -0.25]$
30	-0.05 $[-0.12, +0.02]$	-1.09 $[-1.67, -0.53]$	-0.50 $[-0.72, -0.29]$
32	-0.04 $[-0.12, +0.04]$	-1.55 $[-2.15, -0.95]$	-0.45 $[-0.67, -0.23]$
36	-0.28 $[-0.36, -0.21]$	-2.61 $[-3.20, -2.03]$	-0.52 $[-0.75, -0.27]$
42	-0.17 $[-0.25, -0.09]$	-2.47 $[-2.99, -1.94]$	-0.89 $[-1.15, -0.63]$
49	-0.47 $[-0.56, -0.38]$	-3.44 $[-4.02, -2.88]$	-1.01 $[-1.32, -0.70]$
64	-0.72 $[-0.83, -0.62]$	-4.45 $[-5.00, -3.91]$	-1.81 $[-2.19, -1.44]$

Across budgets, projected centroids place on average 0.99%, 1.19%, and 0.92% of centers within 10^{-7} of the boundary for ResNet-50, ConvNeXt-T, and ViT-B/16; the largest per-setting means are 7.02%, 7.45%, and 6.03%. Relative to forward clipping, projection has lower MSE in 11/24 settings, but the effects are representation- and budget-dependent: seven image-bootstrap CIs favor projection and eight favor clipping. Thus the constraint implementation affects optimization without explaining why raw-energy OAPC does not outperform the Lloyd-optimized uniform geometry.

Table 10: COCO feature-map validation (§4.2) setup.

Parameter	Value
Backbones	ResNet-50, ConvNeXt-Tiny, ViT-B/16 (torchvision, ImageNet pretrained)
Features	C4 / C4 / block 8; stride 16; $28^2 \times (1024/384/768)$
ViT forward path	CLS and patch positional embeddings added before encoder dropout; pretrained patch positions bicubically interpolated to 28×28; output after block 8, before final encoder normalization
Input preprocessing	direct bilinear resize to 448×448 (aspect ratio not preserved)
Dataset subset	COCO val2017; 300 images, seed 42; same images for all backbones and budgets
Strata	100 each: single (1), few (2–3), crowded (≥ 8); counts 4–7 excluded
Token budgets	main N=32; scan $N \in \{16, 25, 30, 32, 36, 42, 49, 64\}$
Placement methods	fixed-grid, uniform-CVT, OAPC(hard), OAPC(projected), matched constant-ρ OAPC, ρ top-k, weighted k-means
Fixed-grid orientation	both transposes for $N = 30, 32, 42$; in-sample lower-mean global orientation defines descriptive Fixed*
Uniform-CVT	equal-weight Lloyd iterations; 50-step cap; tolerance 10^{-5}; fixed seeds 7919–7928; best of 10; cached by H, W, N
Statistical tests	paired Wilcoxon; paired image bootstrap: 10k for descriptive orientation checks; 20k crossed image–seed bootstrap for OAPC–U-CVT and matched-density controls; Holm correction within each 24-setting comparison family; Wilson-score CIs; rank-biserial
ρ / metrics	Eq. 1; MSE_{all} and target-aligned MSE_{hi} (top-25% raw ρ)
Comparison mode	paired against fixed-grid and uniform-CVT at matched N
OAPC optimization	$\tau_0/\tau_f = 1.0/0.10$ cosine; $\lambda_{\text{rep}} = 0.5$; Adam, lr .05; $T = 200$; forward-clamp main and post-step projection control
Matched density control	replace only ρ_i by $1/(HW)$ in projected OAPC; same objective, schedule, steps, and seeds
Initialization	random OAPC/weighted-k-means method seeds 42, 43, 44

Table 11: Semantic mask-prediction stress-test setup used in §5.

Parameter	Value
Backbones / feature layer	Frozen ResNet-50 and ConvNeXt-T C4, stride 16
Input / feature / mask size	448×448 / $28 \times 28 \times (1024/384)$ / 28×28
Dataset pool	COCO val2017, 1,500 images (500 per category), sampling seed 42
Pilot splits	seeds 42–44; each 999 train / 501 validation; same sampled pool repartitioned
Categories	single (1 annotation), few (2–3), crowded (≥ 8)
Mask target	Union of non-crowd instance masks at feature resolution
Tokenizers	favorable fixed grid; uniform-CVT; raw/standardized/mask-oracle OAPC
Assignment broadcast	matched hard pooling and hard broadcast for all methods
Token budget / init	N=32, random centroid init seeds 42, 43, 44
OAPC tokenizer settings	projected; T=200, τ_f=0.10, λ_{rep}=0.5
Decoder	token MLP + broadcast + $1 \times 1/3 \times 3/1 \times 1$ conv head
Decoder hidden dim / parameters	256 / 508,801 (ResNet), 344,961 (ConvNeXt)
Optimizer / schedule	Adam, learning rate 10^{-3}, 20 epochs
Batch size / workers	32 / 4
Loss	BCEWithLogits (pos_weight 4.0) + Dice loss
Model selection	final epoch; no best-epoch selection
Metrics	micro-IoU/Dice, nonempty-image macro-IoU/Dice, empty-mask false positives
Uncertainty	20k paired image–split–centroid–decoder crossed bootstrap
Factorial protocol	per backbone: fixed/U-CVT 3×3 each; each OAPC density $3 \times 3 \times 3$ (99 runs)

Table 12: Cross-backbone zero-training energy/mask diagnostic on 300 COCO images. Hit is the fraction of top-5% energy cells inside the union mask; enrichment is hit divided by mask area; mask ρ pct. is the mean within-mask energy percentile; centroid bias is OAPC minus fixed-grid mean energy percentile. Each quantity is computed per image and then averaged; enrichment is the reported mean hit divided by the reported mean mask area. Mask-percentile means omit masks that become empty after feature-resolution downsampling (valid $n = 299/295/276/295/295$ in row order); other columns use all 300.

Backbone	layer / size	hotspot hit	mask area	enrichment	mask ρ pct.	centroid bias
ResNet-50	C3 / 56×56×512	44.9%	25.2%	1.78×	.711	+.096
ResNet-50	C4 / 28×28×1024	48.7%	25.2%	1.93×	.746	+.028
ResNet-50	C5 / 14×14×2048	50.1%	24.9%	2.01×	.738	+.034
ConvNeXt-T	C4 / 28×28×384	8.0%	25.2%	0.32×	.265	+.061
ViT-B/16	block 8 / 28×28×768	14.4%	25.2%	0.57×	.519	−.001

Table 13: Channel-standardization routing control at three 28×28 representations. Hit rates average the same 300 images, selecting $\mathrm{round}(.05HW) = 39$ cells; Table 12 uses $\lceil .05HW \rceil = 40$ at this resolution, explaining the small raw-hit difference. Standardized L2 first centers and scales each channel over spatial locations. The last three columns compare forward-clamped standardized-L2 OAPC with uniform-CVT at N=32, averaging three centroid seeds; positive is better. Paired image-bootstrap CIs are reported below.

Representation	raw hit	std. hit	Δcenter-in-mask	ΔFG coverage	Δglobal MSE
ResNet-50 C4	48.7%	50.7%	+0.66 pp	+3.03 pp	−0.02%
ConvNeXt-T C4	8.0%	42.1%	+0.49 pp	+2.62 pp	−3.84%
ViT-B/16 block 8	14.2%	23.8%	−0.80 pp	+0.46 pp	−0.71%

Paired image-bootstrap 95% CIs (ResNet/ConvNeXt/ViT) are $[+1.67, +4.44]$, $[+1.14, +4.08]$, and $[-0.83, +1.75]$ pp for foreground coverage; global-MSE reduction intervals are $[-0.11, +0.08]\%$, $[-4.61, -3.09]\%$, and $[-0.98, -0.46]\%$. These 20k-resample intervals average the three fixed seeds within image and measure image uncertainty only. Foreground coverage omits the five empty downsampled masks; center-in-mask and global MSE use all 300 images. Standardization is performed per image and channel with a spatial standard deviation floor of 10^{-6}; pooled feature vectors remain unstandardized.

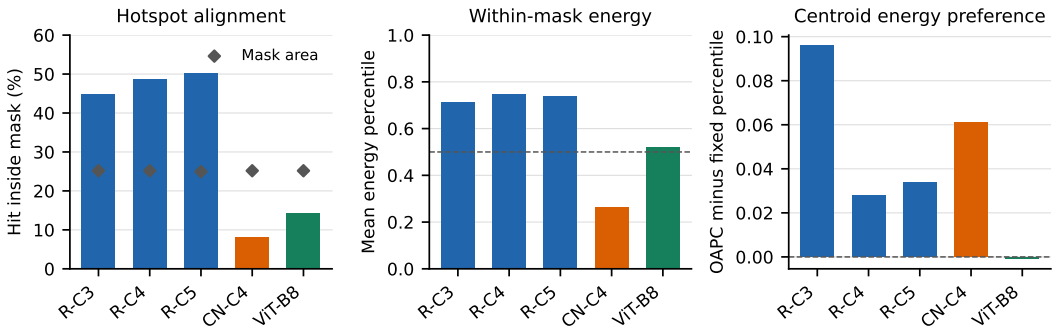

Figure 5: Zero-training feature-energy/semantic-mask misalignment on the 300-image COCO diagnostic subset. The three panels show top-5% energy hit rate inside the union mask, mean energy percentile inside the mask, and the OAPC minus fixed-grid centroid-percentile difference. OAPC centroids sample slightly higher-energy locations than fixed-grid anchors for the CNN representations, while the ViT comparison is effectively tied. Gray markers indicate mask area in the hit-rate panel. The figure therefore shows that target following is representation-dependent even when the target is imperfectly aligned with foreground semantics.

Table 14: Full token-count sweep on the same 300 COCO images. Entries are mean relative MSE reductions (%); positive is better. U, O, K, and F denote uniform-CVT, forward-clamped OAPC with hard pooling, weighted k-means, and fixed-grid. F* uses the better in-sample global orientation for each nonsquare grid, whose rows are highlighted; its columns are descriptive. Square budgets have F*=F. Stochastic methods average seeds 42–44.

Backbone	N	U–F$^\star_{\text{all}}$	O–F$^\star_{\text{all}}$	O–U$_{\text{all}}$	O–U$_{\text{hi}}$	K–U$_{\text{all}}$
ResNet-50	16	-0.02	-0.19	-0.17	-0.19	$+0.11$
	25	-0.05	-0.14	-0.09	$+0.08$	$+0.08$
	30	-0.27	-0.50	-0.23	$+0.08$	$+0.09$
	32	-0.36	-0.39	-0.03	$+0.37$	$+0.15$
	36	$+0.13$	-0.09	-0.22	$+0.23$	$+0.04$
	42	-0.52	-0.50	$+0.02$	$+0.54$	$+0.38$
	49	-0.60	-0.96	-0.35	$+0.22$	$+0.49$
	64	-0.64	-1.32	-0.68	$+0.05$	$+0.82$
ConvNeXt-T	16	-0.31	-1.86	-1.56	-3.39	-1.29
	25	$+0.52$	-0.31	-0.91	-1.91	-1.67
	30	-0.78	-2.41	-1.72	-3.02	-1.99
	32	$+3.17$	$+1.80$	-1.68	-2.91	-2.03
	36	-1.38	-3.61	-2.30	-3.59	-2.06
	42	-2.13	-3.81	-1.71	-2.61	-2.20
	49	-1.33	-4.08	-2.81	-3.80	-1.52
	64	$+1.35$	-2.37	-3.88	-4.50	-1.59
ViT-B/16	16	-0.06	-1.15	-1.09	-1.04	$+0.15$
	25	-0.26	-1.02	-0.76	-0.22	$+0.28$
	30	-0.32	-0.96	-0.65	-0.35	$+0.40$
	32	$+0.04$	-0.31	-0.38	$+0.27$	$+0.61$
	36	$+0.20$	-0.31	-0.53	$+0.00$	$+0.80$
	42	$+0.13$	-0.40	-0.56	-0.02	$+0.88$
	49	-0.38	-0.98	-0.62	$+0.01$	$+1.29$
	64	$+0.39$	-1.40	-1.88	-2.22	$+0.94$

Table 15: Projected OAPC versus uniform-CVT global-MSE uncertainty across the full sweep, matching Figure 3(a). Both use hard region-mean reconstruction. Δ is the mean paired per-image reduction (%; positive favors OAPC). Crossed 95% CIs independently resample 300 paired images and the three global centroid seed IDs 20k times. "Seeds 42/43/44" gives the three descriptive full-dataset effects. p_{H} is the crossed-bootstrap sign probability adjusted across all 24 settings by Holm's procedure. With only three seed IDs, these intervals are a sensitivity analysis, not a precise population-level seed CI.

Backbone	N	Δ	crossed 95% CI	seeds 42/43/44	p_{H}
ResNet-50 C4	16	$+0.02$	$[-0.09, +0.12]$	$+0.07/+0.05/-0.07$	1
	25	-0.00	$[-0.11, +0.10]$	$+0.04/-0.10/+0.05$	1
	30	-0.05	$[-0.14, +0.03]$	$-0.10/-0.02/-0.04$	1
	32	-0.04	$[-0.19, +0.09]$	$+0.01/-0.18/+0.04$	1
	36	-0.28	$[-0.37, -0.19]$	$-0.34/-0.27/-0.25$	0.0024
	42	-0.17	$[-0.27, -0.07]$	$-0.12/-0.22/-0.16$	0.014
	49	-0.47	$[-0.62, -0.32]$	$-0.44/-0.36/-0.61$	0.0024
	64	-0.72	$[-0.93, -0.47]$	$-0.81/-0.89/-0.46$	0.0024
ConvNeXt-T C4	16	-0.76	$[-1.75, +0.22]$	$-0.23/-1.43/-0.61$	0.787
	25	-0.26	$[-0.99, +0.44]$	$-0.47/-0.02/-0.28$	1
	30	-1.09	$[-1.72, -0.45]$	$-0.88/-1.22/-1.17$	0.0121
	32	-1.55	$[-2.40, -0.76]$	$-1.33/-2.17/-1.17$	0.0028
	36	-2.61	$[-3.43, -1.80]$	$-2.03/-2.62/-3.17$	0.0024
	42	-2.47	$[-3.09, -1.86]$	$-2.67/-2.22/-2.52$	0.0024
	49	-3.44	$[-4.18, -2.71]$	$-3.58/-2.98/-3.77$	0.0024
	64	-4.45	$[-5.13, -3.77]$	$-4.34/-4.20/-4.83$	0.0024
ViT-B/16 block 8	16	-0.42	$[-0.72, -0.12]$	$-0.16/-0.47/-0.64$	0.0592
	25	-0.44	$[-0.71, -0.17]$	$-0.23/-0.54/-0.56$	0.0279
	30	-0.50	$[-0.76, -0.24]$	$-0.36/-0.55/-0.60$	0.0039
	32	-0.45	$[-0.84, -0.02]$	$-0.04/-0.75/-0.56$	0.276
	36	-0.52	$[-0.78, -0.23]$	$-0.45/-0.62/-0.47$	0.0108
	42	-0.89	$[-1.22, -0.55]$	$-0.85/-1.09/-0.73$	0.0024
	49	-1.01	$[-1.35, -0.66]$	$-1.01/-1.09/-0.92$	0.0024
	64	-1.81	$[-2.22, -1.41]$	$-1.83/-1.86/-1.73$	0.0024

Table 16: Matched raw-energy versus constant-density projected OAPC control. Δ is the mean paired per-image global-MSE reduction (%; positive favors raw energy). Crossed 95% CIs independently resample 300 paired images and the three matched global seed IDs 20k times. Seed columns are descriptive paired effects; p_{H} is the crossed-bootstrap sign probability adjusted across all 24 settings by Holm's procedure. With only three seed IDs, these intervals are a sensitivity analysis, not a precise population-level seed CI.

Backbone	N	Δ	crossed 95% CI	seeds 42/43/44	p_{H}
ResNet-50 C4	16	$+0.05$	$[+0.01, +0.09]$	$+0.04/+0.06/+0.05$	0.0628
	25	$+0.15$	$[+0.08, +0.21]$	$+0.10/+0.14/+0.20$	0.0024
	30	$+0.17$	$[+0.06, +0.28]$	$+0.18/+0.06/+0.28$	0.0032
	32	$+0.25$	$[+0.19, +0.31]$	$+0.28/+0.25/+0.23$	0.0024
	36	$+0.24$	$[+0.17, +0.32]$	$+0.23/+0.20/+0.29$	0.0024
	42	$+0.32$	$[+0.15, +0.46]$	$+0.43/+0.39/+0.14$	0.0042
	49	$+0.37$	$[+0.14, +0.61]$	$+0.34/+0.62/+0.14$	0.0032
	64	$+0.27$	$[+0.12, +0.40]$	$+0.35/+0.31/+0.13$	0.011
ConvNeXt-T C4	16	-0.28	$[-1.13, +0.37]$	$-0.22/+0.03/-1.01$	1
	25	$+1.29$	$[+0.58, +1.83]$	$+0.64/+1.44/+1.59$	0.011
	30	$+1.32$	$[+0.82, +1.77]$	$+1.14/+1.07/+1.50$	0.0024
	32	$+1.23$	$[+0.45, +1.92]$	$+0.86/+1.86/+0.72$	0.0144
	36	$+1.24$	$[+0.71, +1.70]$	$+1.12/+0.94/+1.42$	0.0024
	42	$+1.75$	$[+0.95, +2.47]$	$+1.04/+1.66/+2.33$	0.0024
	49	$+1.56$	$[+0.50, +2.61]$	$+0.58/+2.60/+1.20$	0.0144
	64	$+0.22$	$[-0.43, +0.76]$	$+0.50/+0.05/-0.21$	1
ViT-B/16 block 8	16	$+0.05$	$[-0.37, +0.39]$	$+0.34/+0.21/-0.43$	1
	25	$+0.30$	$[+0.13, +0.49]$	$+0.42/+0.23/+0.22$	0.0091
	30	$+0.36$	$[+0.14, +0.58]$	$+0.44/+0.15/+0.46$	0.0144
	32	$+0.51$	$[+0.25, +0.75]$	$+0.69/+0.50/+0.30$	0.0096
	36	$+0.36$	$[+0.17, +0.54]$	$+0.43/+0.28/+0.34$	0.011
	42	$+0.75$	$[+0.52, +1.02]$	$+0.71/+0.66/+0.85$	0.0024
	49	$+0.98$	$[+0.51, +1.53]$	$+0.60/+1.52/+0.76$	0.0024
	64	$+0.45$	$[+0.15, +0.73]$	$+0.30/+0.43/+0.52$	0.0255

Table 17: Absolute mean global reconstruction MSE at $N{=}32$ on the same 300 images. Fixed* is the favorable in-sample orientation. Stochastic methods report mean$\pm$sample SD (divisor $3-1$) over seeds 42–44. This table includes the direct ρ top-k baseline omitted from the compact main-text comparison.

Backbone	Fixed*	U-CVT	OAPC(hard)	ρ top-k	weighted k-means
ResNet-50 C4	0.007024	0.007044	0.007048 ± 0.000009	0.007312	0.007034 ± 0.000002
ConvNeXt-T C4	0.248138	0.239147	0.242171 ± 0.000583	0.328612	0.243900 ± 0.000696
ViT-B/16 block 8	0.232794	0.232698	0.233349 ± 0.001195	0.248880	0.231119 ± 0.000246

Table 18: Measured placement and reconstruction cost and token geometry on 30 ResNet-50 C4 images ($N{=}32$, RTX 3090), excluding backbone extraction. Distances are in normalized coordinates. Uniform-CVT mean latency includes its first one-time 286.9-ms layout construction; its steady median is 4.2 ms.

Method	mean ms	median ms	center NN	coverage radius	region-area CV
Fixed-grid	3.64	3.60	.125	.140	.143
Uniform-CVT	13.65	4.20	.169	.135	.106
OAPC(hard)	954.34	893.15	.166	.184	.229
ρ top-k	6.24	7.50	.060	.479	1.173
Weighted k-means	30.13	28.95	.166	.143	.181

E Initialization Sensitivity

All experiments initialize centroids uniformly at random in $[0, 1]^2$; COCO reconstruction reports seeds 42–44 and other controlled experiments use seed 42 unless stated otherwise. Reduced checks compare random with a boundary-margined uniform-grid initialization on three configurations (Table 19). They use the main optimizer settings but a standalone script and separately generated synthetic realizations, so we interpret only within-configuration initialization differences, not absolute MSE or another fixed-grid baseline.

Table 19: Initialization sensitivity: reduced verification checks. MSE values $\times 10^{-3}$; Δ = relative change from random to uniform-grid.

Test	Init	MSE_{all}	MSE_{hi}	Δ_{all}	Δ_{hi}
Gauss $K{=}8$, $N{=}8$	random	3.08	4.13	—	—
	uniform-grid	3.06	4.09	-0.6%	-1.0%
Gauss $K{=}8$, $N{=}16$	random	2.86	3.44	—	—
	uniform-grid	2.87	3.47	$+0.2\%$	$+0.8\%$
Stripe-v, $N{=}16$	random	3.04	4.18	—	—
	uniform-grid	3.31	5.51	$+9.0\%$	$+\mathbf{31.8}\%$

The two modes differ by at most 1% on isotropic peaks, while uniform-grid initialization degrades stripe MSE_{hi} by 32%. This is consistent with persistent lattice bias on anisotropic structure, so random initialization remains the default.

F Semantic Routing Diagnostics

Table 20: Semantic readout means over the crossed runs on the 1,500-image pool. Each tokenizer condition has a separately trained decoder of the same architecture. Fixed and U-CVT average nine runs per backbone; each OAPC density averages 27. Macro-IoU excludes empty masks. Empty pixel FPR is the fraction of pixels predicted foreground on empty-mask images; empty image FPR is the fraction of those images with any foreground prediction. These are descriptive means, not paired effects.

Backbone	Tokenizer	micro-IoU	macro-IoU	empty pixel FPR	empty image FPR
ResNet-50	Fixed	0.5148	0.4534	4.38%	40.7%
	U-CVT	0.5328	0.4751	4.24%	42.1%
	Raw	0.5347	0.4780	3.28%	55.8%
	Standardized	0.5356	0.4800	3.39%	52.0%
	Mask oracle	0.5514	0.5115	3.41%	50.1%
ConvNeXt-T	Fixed	0.5436	0.4736	2.50%	50.2%
	U-CVT	0.5571	0.4876	2.41%	37.9%
	Raw	0.5504	0.4810	2.34%	45.6%
	Standardized	0.5578	0.4915	2.24%	51.9%
	Mask oracle	0.5649	0.5181	2.63%	53.4%

Table 21: Semantic differences from uniform-CVT. Entries are mean paired changes and 95% crossed image–split–centroid–decoder sensitivity intervals (20k replicates). The three splits repartition the same sampled pool; intervals do not quantify generalization across independently sampled datasets. Positive is better.

Backbone	OAPC density	Δmicro-IoU (95% interval)	Δmacro-IoU (95% interval)
ResNet-50	Raw	$+0.0019$ $[-0.0042, +0.0101]$	$+0.0029$ $[-0.0028, +0.0092]$
	Standardized	$+0.0028$ $[-0.0043, +0.0120]$	$+0.0049$ $[-0.0011, +0.0122]$
	Mask oracle	$+0.0186$ $[+0.0077, +0.0287]$	$+0.0364$ $[+0.0281, +0.0448]$
ConvNeXt-T	Raw	-0.0067 $[-0.0128, -0.0012]$	-0.0066 $[-0.0122, -0.0009]$
	Standardized	$+0.0007$ $[-0.0065, +0.0083]$	$+0.0040$ $[-0.0026, +0.0110]$
	Mask oracle	$+0.0079$ $[+0.0022, +0.0140]$	$+0.0305$ $[+0.0219, +0.0395]$

Geometric oracle. To contrast observed routing with an oracle foreground-placement reference, we compare feature-energy and mask-oracle densities with the same optimizer, $N{=}32$, 200 steps, seeds 42–44, feature maps, and 300 images. Non-crowd mask unions are downsampled to $28{\times}28$ with foreground/background weights 1/.01; feature routing uses Eq. 1. Foreground metrics omit five empty masks, while global MSE uses all images. Centroid-in-mask, nearest-centroid distance, 1.5-cell coverage, and foreground MSE measure capacity near semantic foreground, not supervised-tokenizer performance.

Table 22: Per-seed relative change from feature-energy to mask-oracle routing at $N{=}32$. Positive is better for fraction/coverage and negative for distance/MSE. Per seed, paired Wilcoxon $p < 10^{-40}$ for all four foreground metrics.

Seed	centroid-in-mask	FG mean distance	FG coverage	FG MSE	global MSE
42	$+44.05\%$	-23.39%	$+69.16\%$	-8.26%	$+0.31\%$
43	$+36.27\%$	-21.84%	$+67.30\%$	-7.24%	$+0.03\%$
44	$+40.34\%$	-22.26%	$+66.64\%$	-7.70%	$+0.54\%$
Mean	$+40.22\%$	-22.50%	$+67.70\%$	-7.73%	$+0.29\%$

Interpretation. The oracle improves every foreground metric under all seeds while changing global MSE by less than 0.6% (Table 22). Continuous placement can cover semantic foreground, but global reconstruction scarcely rewards it. Together with Tables 12–13, this supports the narrower conclusion that raw ρ is representation- and normalization-dependent energy, not architecture-independent semantic salience.