EVERY SPARKLE GETS A TOKEN: ENTROPY-GUIDED FOVEATED TOKENIZATION FOR FINE-GRAINED VISUAL RECOGNITION

Anonymous authors
Paper under double-blind review

ABSTRACT

Uniform tokenization fixes the number, location, and scale of visual tokens before seeing image content. This is inefficient when fine-grained evidence is concentrated in a small part of the image, yet learned pruning usually decides relevance only after dense patches have entered the backbone. We ask whether a transparent, parameter-free selector can construct the token set directly from pixels. SpiralFovea computes local luminance entropy, selects spatially distributed anchors, and extracts multi-scale rings around them; only the retained crops are processed. The rings also induce a fovea-to-periphery sequence that can be consumed directly by either a transformer or a causal Mamba fusion head. At 224×224, the evaluated configuration uses at most 78 spatial tokens instead of the 196-token ViT-S/16 grid. Across four fine-grained tasks, it improves the reported mean accuracy of frozen ViT/DINO and ResNet–Mamba pipelines while raising measured throughput by 17.8–28.9%. On standard 200-way CUB at a matched sparse budget, SpiralFovea obtains 85.16% top-1 accuracy, within 0.45 points of the dense model and 0.99 points above ToMe. At 518×518, it reduces 1,369 dense tokens to roughly 120–200, with accuracy changes ranging from $+1.90$ points on WikiArt Style to -5.86 on CUB. Annotation analysis exposes the boundary: anchor lift disappears when textured backgrounds dominate. Thus SpiralFovea is a sparse operating point for locally entropic, spatially coherent evidence, with a meaningful order for attention and state-space backbones.

1 INTRODUCTION

A vision transformer (ViT) normally converts a 224×224 image into a fixed 14×14 grid of 196 patches (Dosovitskiy et al., 2021). This design is predictable and hardware-friendly, but it allocates the same number, scale, and locations of tokens to every image. Fine-grained recognition often violates the implied uniformity: a species decision may depend on a bird's wing pattern, a flower label on petal structure, and a pathology label on localized tissue morphology. The surrounding pixels can still provide context, but need not receive the same sampling density as the decisive region.

Existing efficient ViTs usually reduce tokens after grid tokenization (Rao et al., 2021; Liang et al., 2022; Fayyaz et al., 2022; Bolya et al., 2023; Yin et al., 2022). Recent work instead adapts the input representation itself, using mixed scales, elastic sampling, continuous placement, boundaries, or hierarchical model selection (Havtorn et al., 2023; Pardyl et al., 2025; Hjelkrem-Tan et al., 2025; Chen et al., 2025; Aasan et al., 2025). These methods demonstrate that the grid is not sacrosanct, but often introduce a learned selector, an auxiliary objective, or a search procedure. We ask a deliberately narrower question: *how competitive can a parameter-free, pre-backbone selector be?*

We present SpiralFovea, which uses local image entropy as an inexpensive proxy for where high-frequency visual structure is located. It chooses several spatially distributed anchors and samples fixed multi-scale rings around them. Selection finishes before patch projection, so only the retained crops enter the backbone. The method is inspired by the computational allocation of foveated vision, but we do not claim that entropy models biological gaze or semantic relevance. Instead, we test that premise directly through object-overlap and inverted-entropy analyses.

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.

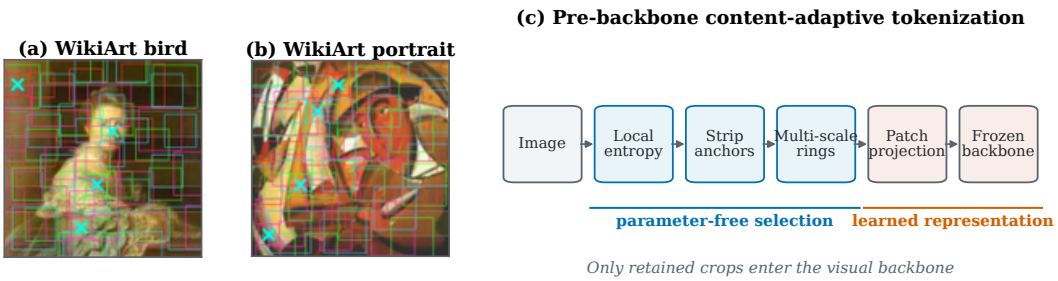


Figure 1: SpiralFovea in practice and in pipeline form. In (a–b), cyan crosses are entropy anchors and colored boxes are multi-scale crops; dense sampling follows textured subjects while uniform regions receive fewer crops. In (c), selection is completed before patch projection, so only retained crops reach the backbone.

The paper is organized around three testable questions. *Can selection finish before the backbone without sacrificing fine-grained accuracy?* We compare dense, random-sparse, merging, pruning, and learned mixed-scale alternatives at controlled token budgets. *Does entropy place anchors on the labelled object?* We measure box and exact-mask overlap, including the adverse cohort in which background entropy is higher. *Do the savings persist as resolution grows?* We evaluate a 1,369-token DINOv2 grid at 518×518. We also show that the spiral is both a sampling geometry and a sequence order: it supports frozen ViT/DINO projections and gives a CNN–Mamba model a deterministic centre-outward scan. The contribution is a measured operating point with an observable scope condition, rather than a universal superiority claim.

2 Related Work

Reducing tokens inside a transformer. DynamicViT, EViT, ATS, A-ViT, and ToMe reduce a sequence after uniform patches have been embedded, through learned pruning, attention-based sampling, adaptive halting, reorganization, or merging (Rao et al., 2021; Liang et al., 2022; Fayyaz et al., 2022; Yin et al., 2022; Bolya et al., 2023). These methods retain the pretrained input interface and can exploit semantic backbone features. Their reduction begins after some input work has already occurred. SpiralFovea uses only pixels to decide the sequence before patch projection; this removes early backbone work but cannot use task-aware features.

Changing the input tokenization. MSViT chooses coarse or fine tokens through a learned gate (Havtorn et al., 2023); elastic sampling and SPoT explore non-grid input locations (Pardyl et al., 2025; Hjelkrem-Tan et al., 2025); EPOC constructs tokens from estimated subobject boundaries (Chen et al., 2025); and differentiable hierarchical tokenization adapts regions with a learned hierarchical criterion (Aasan et al., 2025). SpiralFovea occupies a simpler point in this design space: its selector is deterministic and requires neither task labels nor selector training. This simplicity lowers overhead and makes the failure mode inspectable, but entropy is a weaker relevance signal than learned semantics.

Foveated computation. Foveated sampling has been studied as a saliency-driven neural layer (Recasens et al., 2018), as a learned perceptual organization (Deza & Konkle, 2020), and recently for vision–language systems under limited visual budgets (Gizdov et al., 2025). Our use of "foveated" describes nonuniform sampling density around selected anchors. It is a computational analogy; the anchor policy is not a model of human fixation.

Table 1: High-level positioning of token-reduction strategies. "Input" means the retained set is fixed before transformer block 1. Location and scale indicate whether those token properties adapt to image content.

Method family	Decision stage	Backbone features	Location	Scale	Trained selector
Dense ViT (Dosovitskiy et al., 2021)	input	no	✗	✗	✗
DynamicViT/EViT (Rao et al., 2021; Liang et al., 2022)	inside	yes	✗	✗	✓
ToMe (Bolya et al., 2023)	inside	yes	✗	✗	✗
MSViT (Havtorn et al., 2023)	input	no	✗	✓	✓
Elastic/SPoT (Pardyl et al., 2025; Hjelkrem-Tan et al., 2025)	input	no	✓	partial	✓
EPOC/DHT (Chen et al., 2025; Aasan et al., 2025)	input	no	✓	✓	✓
SpiralFovea	input	no	✓	✓	✗

Table 1 separates two distinctions that are often conflated. Post-hoc methods can exploit semantic backbone features, but they pay for dense patch projection and at least part of the dense transformer sequence. Learned input tokenizers avoid that early cost, but add a trained selection mechanism. SpiralFovea tests the complementary extreme: input-adaptive location and scale with no learned selector. This makes it cheap and auditable, while limiting it to pixel statistics rather than task semantics.

3 METHOD

Let $I \in [0,1]^{3\times H\times W}$ and let a native dense tokenizer produce $N_{\text{grid}} = (H/P)(W/P)$ patches of width P. SpiralFovea replaces this fixed set with a variable sequence $\mathcal{T}(I) = \{(p_i, u_i)\}_{i=1}^{N(I)}$, where crop p_i and positional descriptor u_i are functions of image content. Figure 1 summarizes the pipeline.

Design hypothesis and falsifiable consequences. We assume only that local entropy is positively associated with useful fine structure often enough to be a practical proposal distribution. This is an empirical premise, not a theorem about semantic relevance. It yields three consequences that we test: entropy anchors should overlap annotated foreground more often than an area-based proxy; replacing them with matched random anchors should reduce accuracy; and their advantage should shrink or reverse when the background is more entropic than the object. The ring sampler then converts each anchor into a local multi-scale neighborhood, hedging against entropy maxima that land on a single edge rather than the full object.

3.1 LOCAL ENTROPY AND DISTRIBUTED ANCHORS

We project RGB to BT.601 luminance, downsample to $D\times D$, and quantize intensities into B bins. For a local window $\mathcal{W}_{xy}$, the entropy map is

$$E(x,y) = -\sum_{b=1}^{B} \hat{p}_b(x,y)\log_2\big(\hat{p}_b(x,y)+\epsilon\big), \tag{1}$$

where $\hat{p}_b$ is the empirical bin frequency. An RGB-channel variant changes CUB accuracy by only 0.01 points (Section 7) while tripling entropy-map channel work, so we use luminance.

A single global maximum would collapse all samples onto one textured region. We divide the image into S horizontal strips and select an entropy maximum in each active strip s:

$$c_s = \arg\max_{(x,y)\in\Omega_s} E(x,y), \tag{2}$$

with a boundary margin in Ω_s. In scan order, a candidate is suppressed if its normalized distance from an earlier retained anchor is below τ_{dedup}. This is a diversity heuristic, not a coverage theorem.

3.2 MULTI-SCALE RING SAMPLING

Around every retained anchor, we use four fixed ring templates. At 224×224, their radii are $(0, 26, 58, 96)$ pixels, crop widths are $(24, 28, 36, 48)$, and angular counts are $(1, 5, 10, 13)$. A center for ring k and index j is

$$r_{qkj} = c_q + \rho_k(\cos\theta_{kj}, \sin\theta_{kj}), \qquad \theta_{kj} = 2\pi j/n_k. \tag{3}$$

Crops are obtained with bilinear sampling (Jaderberg et al., 2015). A crop is discarded when more than τ_{oob} of its area lies beyond the canvas. Remaining crops are resized to the backbone's native $P \times P$ patch resolution; for ViT-S/16 and DINO-ViT-S/16, $P = 16$. In the evaluated 224 setting, deduplication and boundary filtering yielded at most 78 spatial tokens, although the unfiltered four-anchor template contains at most 116 candidates.

3.3 Embedding and spatial encoding

For ViT/DINO, every retained crop is resized to $P \times P$ and passed through the pretrained shared patch projection ϕ. For crop $i = (q, k, j)$, we form

$$z_i = \phi(\text{resize}(p_i, P, P)) + g[\hat{x}_q, \hat{y}_q, \rho_k/H, \log(\sigma_k/P), \sin\theta_{kj}, \cos\theta_{kj}], \tag{4}$$

where g is a two-layer MLP with 64 hidden units and output dimension equal to the backbone width. Normalized anchor coordinates $\hat{x}_q, \hat{y}_q$ identify the foveal centre; radius, crop scale, and angle distinguish tokens around that centre. The original CLS token is prepended, and the sequence is processed by the frozen transformer. Only g and the linear classifier are trained.

Spiral order and Mamba fusion. For the CNN setting, a shared ResNet-18 independently maps the $N(I)$ crops to $X \in \mathbb{R}^{N(I)\times d}$. The token order is lexicographic in strip, anchor, ring radius, and angle:

$$\pi(I) = \big[(q, k, j) : q = 1{:}Q,\ k = 0{:}K-1,\ j = 0{:}n_k - 1\big]. \tag{5}$$

Consequently, tokens from one fovea remain contiguous, the centre is observed before its progressively coarser context, and anchors are visited from the uppermost active strip downward. This ordering is meaningful for Mamba because its selective state-space update is directional: in simplified form, $h_t = A_t h_{t-1} + B_t x_{\pi_t}$ and $y_t = C_t h_t + D x_{\pi_t}$. The state first receives high-resolution evidence near an entropy maximum, then integrates wider context as radius grows. A raster scan would instead interleave unrelated grid cells and has no explicit foveal hierarchy.

Two stacked Mamba blocks (Gu & Dao, 2024) preserve sequence length, after which the outputs are mean-pooled and classified. Mamba accepts the variable $N(I)$ produced by deduplication and boundary filtering, and its recurrent cost is linear in sequence length. The spiral is therefore not merely a patch-placement pattern: it supplies the one-dimensional order required by a state-space model. We do not claim that this is the unique optimal scan; the linear-layout ablation in Section 7 provides the empirical evidence for the chosen order.

3.4 Computation

With N spatial tokens, the quadratic interaction term scales as $(N+1)^2$ including CLS. Thus 78 versus 196 spatial tokens corresponds to a peak reduction of

$$1 - \left(\frac{79}{197}\right)^2 = 83.9\%. \tag{6}$$

This is an attention-interaction calculation, not a claim about total model FLOPs or wall-clock speed. Entropy computation, resampling, MLPs, memory movement, and padding remain. We therefore report measured throughput alongside token and attention statistics.

4 Experimental Protocol

Datasets. We evaluate WikiArt genre and style classification (Saleh & Elgammal, 2016), Oxford Flowers-102 (Nilsback & Zisserman, 2008), PatchCamelyon (PCam) histopathology (Veeling et al., 2018), and the standard 200-species CUB-200-2011 task (Wah et al., 2011). CUB and Flowers provide boxes and masks used only for post-hoc localization analysis; annotations never enter the selector. These tasks span artwork, natural objects, and histopathology. They do not substitute for a large-scale multiclass benchmark: PCam is large but binary and patch-based, whereas CUB and Flowers are comparatively small. We treat large-scale generalization as unresolved rather than extrapolating from these datasets.

Models and training. Inputs are 224×224 unless stated. ViT-S/16, ViT-B/16, and DINO-ViT-S/16 use frozen transformer weights with a trained positional MLP and classifier. The CNN branch uses ResNet-18 and a two-block Mamba fusion head. We use AdamW, learning rate 10^{-4}, weight decay 0.01, label smoothing 0.1, batch size 32, and 30 epochs (20 on PCam). Selection uses $B = 16$, a 15×15 entropy window, $D = 112$, $\tau_{\text{oob}} = 0.60$, and $\tau_{\text{dedup}} = 0.15$.

Comparisons. Dense models receive all 196 uniform patches. Matched sparse controls include a random subset of 78 grid patches and random anchors with the same ring sampler. On standard CUB, all methods use the same ViT-S/16 backbone, split, classifier recipe, and approximately 78-token endpoint. We compare against ToMe (Bolya et al., 2023), DynamicViT (Rao et al., 2021), and MSViT (Havtorn et al., 2023). ToMe can be applied directly to the same pretrained transformer and is therefore the cleanest comparator. DynamicViT and MSViT contain learned selectors normally trained with method-specific schedules, distillation, or budget losses; their shared frozen-backbone results are diagnostic stress tests, not claims about the methods under their recommended recipes.

Metrics. We report top-1 accuracy, spatial tokens, attention-interaction reduction, optimization loss when available, and measured images/s on a Tesla T4 at batch 32. Throughput includes selection, crop extraction, padding, and backbone inference in the implemented pipeline. Attention percentages isolate the quadratic interaction term and should not be read as end-to-end FLOP reductions. High-resolution tests use DINOv2-L/14 (Oquab et al., 2024) at 518×518, whose dense grid has $37^2 = 1{,}369$ tokens; these results are a separate scaling study and are not directly pooled with the 224-resolution experiments.

5 Accuracy and Efficiency

Table 2: Top-1 accuracy (%) at 224×224. Each sparse transformer configuration uses at most 78 spatial tokens; the corresponding dense grid uses 196. Deltas are relative to the matched dense/uniform backbone.

Backbone	Tokenization	WikiArt-G	WikiArt-S	Flowers	PCam	Mean
ResNet-18	uniform Mamba	75.4	85.3	91.4	88.1	85.1
ResNet-18	SpiralFovea+Mamba	77.5	87.1	93.2	89.3	86.8
ViT-S/16	dense	76.2	85.8	91.5	88.5	85.5
ViT-S/16	SpiralFovea	78.4	87.9	93.8	90.1	87.6
ViT-B/16	dense	77.1	86.4	92.8	89.4	86.4
ViT-B/16	SpiralFovea	79.2	88.5	94.6	90.5	88.2
DINO-S/16	dense	78.0	87.6	93.6	90.2	87.4
DINO-S/16	SpiralFovea	**80.3**	**89.5**	**95.4**	**91.0**	**89.1**

Table 2 summarizes the original four-task evaluation. Across the three frozen transformer families, the sparse representation improves the reported mean by 1.7–2.1 points. The ResNet–Mamba branch also improves on every task ($+2.1$, $+1.8$, $+1.8$, and $+1.2$ points; $+1.7$ mean), showing that the result is not confined to attention backbones. Its ordered sequence is especially relevant here: Mamba has no permutation invariance, so the centre-outward spiral defines how fine evidence and peripheral context accumulate in the recurrent state. Gains are smaller on PCam than on the natural-image datasets. These results do not isolate entropy from ring geometry or ordering; Section 7 does so.

Table 3: Matched-budget standard CUB-200-2011 comparison with ViT-S/16. Arrow notation denotes progressive reduction. Progressive-method attention savings are ranges because length changes by layer; SpiralFovea's 83.9% is the 78-token peak and its empirical average across datasets is 79%.

Method	Final	Token red.	Train loss	Val. loss	Top-1	Attn. red.
Dense ViT-S/16	196	0.0%	1.31	**1.38**	**85.61**	0%
ToMe	$\rightarrow 76$	61.2%	1.76	1.48	84.17	45–55%
DynamicViT†	$\rightarrow 78$	60.2%	1.64	1.79	71.78	50–60%
MSViT†	80	59.1%	2.29	2.49	51.00	$\sim$83%
SpiralFovea	≤ 78	60.2%	**1.26**	1.39	85.16	$\leq$ 83.9%

†Shared frozen-backbone protocol; not the method's full task-specific training recipe.

On standard CUB, SpiralFovea is within 0.45 points of the dense model and 0.99 points above ToMe at a similar final token count (Table 3). It also reaches the lowest training loss and a validation loss within 0.01 of the dense reference. Its computational distinction is the placement of reduction: the sequence is sparse before block 1, whereas ToMe progressively merges it. The large gaps to DynamicViT and MSViT under the shared frozen protocol show that learned selectors cannot simply be transplanted without their training machinery; we do not interpret these rows as intrinsic rankings of those methods.

Table 4: End-to-end throughput at batch 32 on one Tesla T4. Measurements include the implemented selector and crop-extraction path.

Pipeline	Dense/uniform	SpiralFovea	Gain
ResNet-18 + Mamba	681	803	+17.9%
ViT-S/16	598	714	+19.4%
ViT-B/16	412	531	+28.9%
DINO-ViT-S/16	612	721	+17.8%

The measured gains in Table 4 are far smaller than the peak quadratic interaction reduction. This gap is expected: patch projection, entropy calculation, crop resampling, batching, MLPs, and non-attention layers remain. The ResNet–Mamba result also shows that the selector's benefit is not restricted to self-attention, although its compute advantage there comes from fewer encoded crops and a shorter state-space sequence rather than quadratic attention.

5.1 High-resolution scaling

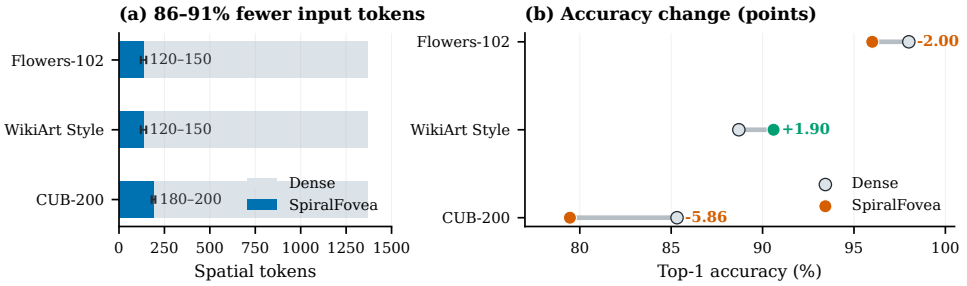


Figure 2: High-resolution trade-off at 518×518 with DINOv2-L/14. Dense models use 1,369 tokens; sparse points use the midpoint of their observed token ranges for visualization. Large compute savings coexist with task-dependent accuracy changes.

At 518×518, 8–12 anchors yield approximately 120–200 tokens instead of 1,369 (Figure 2). WikiArt Style improves from 88.70% to 90.60% with roughly 120–150 tokens. Flowers drops from 98.00% to 96.00%, and CUB drops from 85.32% to 79.46% with 12 anchors (8 anchors reach 73.25%). Thus resolution magnifies the computational advantage, but a constant sparse budget can discard fine detail on articulated objects. The result motivates budget curves rather than a single universal setting.

6 When Does Entropy Help?

Local entropy detects edges and texture, not semantic relevance. We therefore measure whether anchors reach the labelled foreground. Across all CUB images, 57.3% of individual anchors fall in the box and 45.0% in the exact mask, compared with area proxies of 36.2% and 16.3%. At least one anchor reaches the box in 95.4% of images and the mask in 87.8%. Flowers yields 83.9% anchor overlap with boxes and 47.7% with masks, compared with area proxies of 67.6% and 37.3%; per-image success is 98.6% and 84.0%.

Table 5: Foreground localization over complete annotated datasets. The area proxy is the mean target-area fraction. Lift is overlap divided by that proxy; success is the fraction of images with at least one anchor in the region.

Dataset	Region	Area proxy	Anchor overlap	Lift	Success
CUB-200-2011	box	36.2	57.3	1.58×	95.4
CUB-200-2011	exact mask	16.3	45.0	2.76×	87.8
Flowers-102	box	67.6	83.9	1.24×	98.6
Flowers-102	exact mask	37.3	47.7	1.28×	84.0

The area proxy is intentionally simple and is not a matched random-anchor experiment: strip constraints, margins, deduplication, and boundary filtering alter the null distribution. It nevertheless gives a transparent scale reference. The random-anchor accuracy ablation in Section 7 provides a separate controlled test using the same ring sampler.

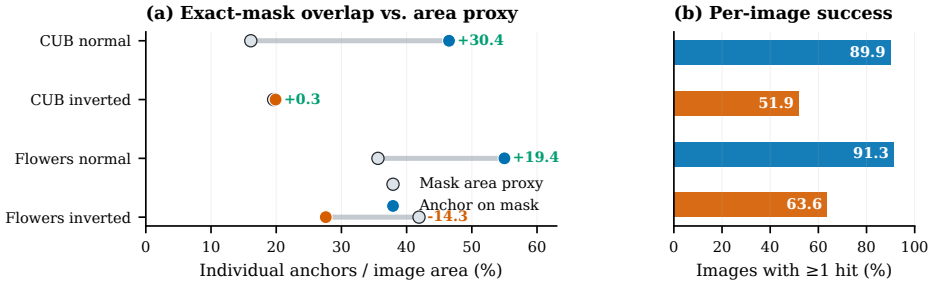


Figure 3: Exact-mask stress test. "Inverted" means background entropy exceeds foreground entropy. The chance proxy is the mask's image-area fraction; it is descriptive rather than a matched random-anchor control. Entropy anchors remain likely to touch the object, but their lift collapses in inverted cases.

We split images by whether mean foreground entropy exceeds background entropy. In the CUB inverted cohort (5.7% of images), mask overlap falls to 19.9%, essentially the 19.6% area proxy, although 51.9% of images retain at least one anchor on the mask. In the Flowers inverted cohort (26.5%), anchor mask overlap is 27.6%, below the 41.9% area proxy, while per-image mask success is 63.6% (Figure 3). Strip-wise selection prevents all anchors from collapsing onto one cluttered region, but it does not make entropy task-aware. This is the method's clearest boundary.

6.1 Object size, spatial coherence, and throughput

Object area alone does not determine accuracy. CUB and Flowers have similar median foreground fractions (approximately 35% and 38%), yet ViT-B/16 accuracy changes by only $+0.50$ points on CUB and $+1.80$ on Flowers. The likely distinction is geometry: flowers are usually compact and contiguous, while birds contain thin articulated structures, varied poses, occlusion, and cluttered backgrounds. This interpretation is consistent with the localization gap in Table 5, but remains an observational explanation rather than a controlled causal result.

Table 6: Change relative to the matched dense ViT-B/16 within object-area terciles. Each dataset is independently split into three equal-sized groups by annotated foreground fraction.

Area tercile	CUB acc.	CUB throughput	Flowers acc.	Flowers throughput
Lowest third	$+0.23$ pp	$+19.4\%$	$+1.32$ pp	$+15.6\%$
Middle third	$+0.24$ pp	$+21.3\%$	$+1.34$ pp	$+17.2\%$
Highest third	$+0.27$ pp	$+22.4\%$	$+1.38$ pp	$+18.9\%$

Throughput gain increases monotonically with object-area tercile in both datasets (Table 6), whereas accuracy is nearly flat within a dataset. Thus annotated area predicts efficiency better than accuracy in

these experiments. Accuracy appears to depend additionally on whether discriminative structure can be captured by a small set of coherent multi-scale neighborhoods. We use this two-factor description—area for expected efficiency and spatial coherence for likely fidelity—as a scope hypothesis, not as a general performance law.

7 Ablations and Analysis

Table 7: DINO-S/16 ablations, averaged across the four 224-resolution tasks. All rows share the frozen backbone. Deltas are relative to full SpiralFovea.

Configuration	Tokens	Mean Δ
Full entropy + multi-scale rings + polar PE	≤ 78	0.0
Random 78 grid patches	78	-1.7
Random anchors, same ring sampler	≤ 78	-1.5
Single-scale rings	≤ 78	-1.1
Linear spatial layout	≤ 78	-0.9
No positional encoding	≤ 78	-0.8
Sinusoidal positional encoding	≤ 78	-0.5
No anchor deduplication	≤ 94	-0.2

Entropy-guided placement is the largest measured component: replacing entropy anchors with matched random anchors costs 1.5 points (Table 7). Randomly selecting 78 uniform grid patches costs 1.7 points, showing that a smaller count alone does not explain the gain. Multi-scale extraction contributes 1.1 points. Replacing the centre-outward spiral layout with a linear spatial layout costs 0.9 points, consistent with the claim that order carries useful structure for sequential fusion. Positional encoding contributes a further 0.5–0.8 points relative to sinusoidal or absent encoding.

On standard CUB with ViT-S/16, luminance entropy reaches 85.16% and independent RGB-channel entropy reaches 85.17%. The result motivates luminance as a cheaper default; it should be read as practical parity in this setting, not a universal statistical-equivalence claim.

The method also produces variable sequence lengths because anchors can be deduplicated and outer crops can leave the canvas. This helps single-image inference but introduces padding in conventional batches. The throughput measurements include the implemented batching path, while the attention percentages describe idealized sequence lengths. Bucketed batching could narrow this gap but is not studied here.

8 Discussion

What supports pre-backbone selection? No single comparison isolates every difference between tokenizers, but the evidence is consistent across complementary controls. Randomly retaining 78 grid patches performs 1.7 points worse than SpiralFovea, so sequence length alone does not explain the result. Random anchors with the same ring sampler lose 1.5 points, isolating the value of entropy-guided placement from the geometry of the rings. At a matched final budget, SpiralFovea is 0.99 points above ToMe and within 0.45 points of the dense grid on standard CUB. Because ToMe uses the same pretrained representation without a separately trained selector, this is the cleanest evidence that choosing the input set can be competitive with merging an already embedded grid. The DynamicViT and MSViT rows answer a different question—whether their selectors transfer to the shared frozen protocol—and should not be used to claim general dominance.

Why does ordering matter? The selector changes more than count: it reallocates samples toward local structure, uses multiple crop scales, and converts the resulting set into a coherent sequence. For transformers, the learned descriptor distinguishes its non-raster geometry. For Mamba, order is part of the computation itself: each state summarizes the previously visited foveal centre, rings, and strips. The centre-outward scan lets fine evidence seed the state before progressively wider context arrives, while the -0.9-point linear-layout ablation suggests that an arbitrary traversal is weaker. This does not establish a unique causal mechanism, but it explains why the same tokenization can benefit both attention and state-space fusion for different architectural reasons.

A practical scope rule. The combined results suggest two separate deployment questions. Foreground area predicts how much compute can be saved, while entropy alignment and spatial coherence predict whether the sparse representation will preserve accuracy. A compact, textured subject on a simpler background is favorable; thin articulated parts or high-entropy clutter require more anchors or a dense fallback. The inverted-cohort statistic can be computed when annotations are available for analysis, but it is not an inference-time oracle. A practical extension would estimate uncertainty from the entropy map itself—for example, diffuse or multi-modal entropy could trigger a larger budget—and test whether this recovers the CUB high-resolution loss without surrendering most of the savings.

9 LIMITATIONS

Entropy is task-agnostic and can prioritize textured clutter, as the inverted-cohort analysis demonstrates. Fixed horizontal strips provide diversity but are not rotation invariant and can miss vertically compact structures. The ring schedule is hand designed, and its hyperparameters do not adapt to object scale. Sparse crops cover some background at coarse resolution; the method reduces background sampling density but does not guarantee that every background pixel is excluded.

The empirical scope is also limited. CUB and Flowers are small fine-grained datasets; WikiArt Style is larger, and PCam contains many samples, but PCam is a binary patch-classification task. We do not evaluate ImageNet-1K, iNaturalist, detection, segmentation, or VQA, so the paper does not establish large-scale multiclass or task-general superiority. The high-resolution study uses DINOv2-L/14 rather than the 224-resolution backbones, making it a scaling probe rather than a controlled resolution-only comparison. Learned selectors evaluated under a frozen shared protocol are disadvantaged relative to their standard training recipes, and the reported point estimates do not quantify uncertainty across extensive multi-seed sweeps.

Finally, theoretical interaction savings overstate end-to-end gains because entropy computation, resampling, patch projection, padding, MLPs, and non-attention layers remain. Variable sequence lengths are convenient for single-image inference but can waste work under padded batching. Future work should learn a lightweight fallback for inverted-entropy images, adapt the number of anchors to image complexity, and test query-conditioned foveation in vision–language tasks, where the relevant region depends on the question rather than the image alone.

10 CONCLUSION

SpiralFovea asks which visual tokens should exist before the backbone is run and in what order they should be processed. A local-entropy selector and multi-scale ring sampler create a short, centre-outward sequence without training a selection network. That sequence nearly preserves dense CUB accuracy at a matched budget, improves measured throughput, and supports both transformer attention and linear-time Mamba fusion. Entropy anchors localize annotated foreground well above its area fraction on ordinary CUB and Flowers images, explaining why the selector can work. The advantage remains conditional: high-resolution savings can grow even as accuracy falls, and anchor lift collapses when background entropy exceeds foreground entropy. Parameter-free foveation is therefore an auditable operating point for spatially coherent fine-grained evidence and a useful baseline for learned or query-conditioned tokenizers.

REPRODUCIBILITY STATEMENT

The appendix records the exact selector procedure, reference pseudocode, token geometry, filtering thresholds, optimization settings, dataset tasks, baseline-adaptation conditions, and cost conventions used in the experiments. The main text separates measured throughput from quadratic attention calculations and distinguishes the shared frozen-backbone protocol from methods' original task-specific training recipes. The datasets and pretrained backbone checkpoints used in the study are publicly available; the provided specifications are intended to support independent reimplementation.

AI USE STATEMENT

Generative AI tools were used to help organize reviewer feedback, identify consistency checks, suggest manuscript structure, and assist with writing and editing. They were not used to generate experimental measurements. The authors manually checked the technical claims, calculations, citations, and final manuscript and take responsibility for all content.

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*, 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.

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.

Arturo Deza and Talia Konkle. Emergent properties of foveated perceptual systems. *arXiv preprint arXiv:2006.07991*, 2020.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

Mohsen Fayyaz, Soroush Abbasi Koohpayegani, Farnoush Rezaei Jafari, et al. Adaptive token sampling for efficient vision transformers. In *European Conference on Computer Vision*, 2022.

Andrey Gizdov, Shimon Ullman, and Daniel Harari. Seeing more with less: Human-like representations in vision models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. In *Conference on Language Modeling*, 2024.

Jakob Drachmann Havtorn, Amélie Royer, Tijmen Blankevoort, and Babak Ehteshami Bejnordi. MSViT: Dynamic mixed-scale tokenization for vision transformers. In *ICCV Workshop on New Ideas in Vision Transformers*, 2023.

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.

Youwei Liang, Chongjian Ge, Zhan Tong, et al. Not all patches are what you need: Expediting vision transformers via token reorganizations. In *International Conference on Learning Representations*, 2022.

Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *Sixth Indian Conference on Computer Vision, Graphics and Image Processing*, pp. 722–729, 2008. doi: 10.1109/ICVGIP.2008.47.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, et al. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024.

Adam Pardyl, Grzegorz Kurzejamski, Jan Olszewski, Tomasz Trzciński, and Bartosz Zieliński. Beyond grids: Exploring elastic input sampling for vision transformers. In *IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 8536–8545, 2025. doi: 10.1109/WACV61041.2025.00827.

Yongming Rao, Wenliang Zhao, Benlin Liu, et al. DynamicViT: Efficient vision transformers with dynamic token sparsification. In *Advances in Neural Information Processing Systems*, 2021.

Adrià Recasens, Petr Kellnhofer, Simon Stent, Wojciech Matusik, and Antonio Torralba. Learning to zoom: A saliency-based sampling layer for neural networks. In *European Conference on Computer Vision*, pp. 51–66, 2018.

Babak Saleh and Ahmed Elgammal. Large-scale classification of fine-art paintings: Learning the right metric on the right feature. *International Journal for Digital Art History*, 2:70–93, 2016. doi: 10.11588/dah.2016.2.23376.

Bastiaan S. Veeling, Jasper Linmans, Jim Winkens, Taco Cohen, and Max Welling. Rotation equivariant CNNs for digital pathology. In *Medical Image Computing and Computer Assisted Intervention*, 2018.

Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-UCSD birds-200-2011 dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011.

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

A Implementation Details

For the 224-resolution configuration, the dense ViT grid contains $14 \times 14 = 196$ patches because the native patch width is 16 pixels. The number 14 refers to the number of grid cells along one axis, not the crop size passed to the patch projection. Every retained variable-size crop is resampled to 16×16 before the frozen projection. Four active horizontal strips propose at most four anchors. Each anchor proposes 29 crops across the four rings, producing at most 116 candidates before deduplication and boundary filtering. The observed retained sequence is no longer than 78 spatial tokens under the evaluated 224-resolution configuration.

For high-resolution DINOv2-L/14 experiments, dense tokenization produces $37 \times 37 = 1{,}369$ spatial tokens. The sparse configurations use 8 anchors for Flowers and WikiArt and 8 or 12 for CUB, with reported token ranges of approximately 120–150 and 180–200 after filtering. The high-resolution attention-reduction percentages in the rebuttal are quadratic interaction estimates derived from these ranges; Figure 2 therefore reports token counts directly.

A.1 Exact selector procedure

For each image, the implementation performs the following deterministic steps before evaluating the backbone:

1. Convert RGB to BT.601 luminance, resize the map to 112×112, quantize it into 16 bins, and compute 15×15 local histograms and Shannon entropy.
2. Divide the entropy map into the configured horizontal strips. Within the valid margin of each strip, retain the maximum-entropy location.
3. Scan candidates from the first strip to the last and suppress a candidate if its normalized Euclidean distance to an earlier anchor is below 0.15.
4. Instantiate the four ring templates around each retained anchor using radii $(0, 26, 58, 96)$, crop widths $(24, 28, 36, 48)$, and angular counts $(1, 5, 10, 13)$ at 224 resolution.
5. Bilinearly sample every crop. Discard it if more than 60% of its area is outside the image, resize each survivor to the backbone's native patch resolution, and preserve anchor–ring–angle metadata for Equation 4.
6. Concatenate tokens in strip, anchor, ring, and angular order; prepend CLS for transformers or pass the ordered feature sequence to the Mamba fusion head.

The entropy map is $O(HW)$ for fixed bins and window size; crop extraction is linear in the number of proposed samples. The method does not evaluate the backbone to select tokens.

A.2 COST CONVENTION

For a transformer with L blocks and token length n, we report the normalized quadratic interaction term $\sum_{\ell=1}^{L} n_\ell^2/(Ln_{\text{dense}}^2)$. Dense and SpiralFovea sequences are constant across blocks, while ToMe and DynamicViT use their layer-wise lengths. This statistic isolates attention interactions; it excludes linear projections, MLPs, convolutional patch embedding, sampling, and selector cost. Reported T4 throughput is the corresponding end-to-end measurement and is the primary systems result.

A.3 REFERENCE PSEUDOCODE

Listing 1 gives framework-level pseudocode for the pre-backbone path described in Section 3. It makes the order of operations explicit: entropy and crop selection are evaluated from pixels, retained crops pass through the shared patch projection, and only the resulting sparse sequence enters the transformer blocks. Vectorized batching and library-specific tensor reshaping are omitted for readability.

Listing 1: Reference pseudocode for the 224-resolution SpiralFovea transformer path.

```
RINGS = [(0, 24, 1), (26, 28, 5),
         (58, 36, 10), (96, 48, 13)]

def local_entropy(image, bins=16, window=15):
    # image is unnormalized RGB in [0, 1]
    y = 0.2989*image[0] + 0.5870*image[1] + 0.1140*image[2]
    y = resize(y, (112, 112))
    q = quantize(y, bins)
    hist = sliding_histogram(q, bins=bins, window=window)
    prob = hist / window**2
    return resize(-(prob * log2(prob + 1e-7)).sum(-1), image.shape[-2:])

def select_anchors(entropy, active_strips=(0, 2, 4, 6),
                   num_strips=8, margin=16, dedup=0.15):
    anchors = []
    for strip in active_strips:          # top-to-bottom scan order
        region = valid_strip(entropy, strip, num_strips, margin)
        candidate = normalize_xy(argmax_2d(region), entropy.shape)
        if all(l2(candidate, prev) >= dedup for prev in anchors):
            anchors.append(candidate)
    return anchors

def make_sparse_tokens(image, patch_embed, position_mlp, native_patch=16):
    entropy = local_entropy(image)
    anchors = select_anchors(entropy)
    tokens = []
    for q, anchor in enumerate(anchors):
        for k, (radius, crop_width, count) in enumerate(RINGS):
            for j in range(count):
                theta = 0.0 if count == 1 else 2*pi*j/count
                center = anchor + pixel_offset(radius, theta, image.shape)
                if out_of_bounds_fraction(center, crop_width, image.shape) > 0.60:
                    continue
                crop = bilinear_crop(image, center, crop_width)
                crop = resize(crop, (native_patch, native_patch))
                descriptor = [anchor.x, anchor.y, radius/image.height,
                              log(crop_width/native_patch),
                              sin(theta), cos(theta)]
                tokens.append(patch_embed(crop) + position_mlp(descriptor))
    return stack(tokens)                # at most 78 in evaluated setting

def forward(image, vit, position_mlp, classifier):
    sparse = make_sparse_tokens(image, vit.patch_embed.proj, position_mlp)
    sequence = concat([vit.cls_token, sparse], axis=0)
    sequence = vit.blocks(sequence)      # dense grid is never instantiated
    return classifier(vit.norm(sequence)[0])

def forward_mamba(image, resnet, mamba_blocks, classifier):
    # Uses the same (strip, anchor, ring, angle) order as above.
    ordered_crops = make_ordered_crops(image)
    sequence = stack([resnet(crop) for crop in ordered_crops])
    for block in mamba_blocks:
        sequence = block(sequence)       # directional state-space scan
    return classifier(sequence.mean(axis=0))
```

B Additional Protocol Notes

The matched-budget CUB table compares all methods on the standard 200-way species task. ToMe progressively reduces the sequence, whereas SpiralFovea fixes the retained sequence before transformer block 1. DynamicViT and MSViT include learned selectors and normally require method-specific training; their shared frozen-protocol results are retained for completeness but are not used as the principal evidence of superiority. A definitive leaderboard comparison would retrain each method with its recommended optimization, distillation, and budget regularization while reporting total training and inference cost.

The foreground analyses use object boxes and exact masks only after inference. "Chance" refers to mean annotated area fraction, which is an intuitive proxy but does not exactly match strip-constrained random anchors. A stricter follow-up would sample random anchors under identical strips, margins, deduplication, and boundary filtering, then bootstrap the difference in overlap and accuracy.

C Detailed Stress-Test Results

Table 8: Foreground localization in normal and inverted entropy cohorts. Overlap is the percentage of individual anchors inside the exact mask; success is the percentage of images with at least one such anchor.

Dataset	Cohort	Share	Mask overlap	Area proxy	Success
CUB	all	100.0	45.0	16.3	87.8
CUB	normal	94.3	46.5	16.1	89.9
CUB	inverted	5.7	19.9	19.6	51.9
Flowers	all	100.0	47.7	37.3	84.0
Flowers	normal	73.5	55.0	35.6	91.3
Flowers	inverted	26.5	27.6	41.9	63.6