ToMATOS: Tokens Meaningfully Aligned to Object Semantics

Anonymous authors
Paper under double-blind review

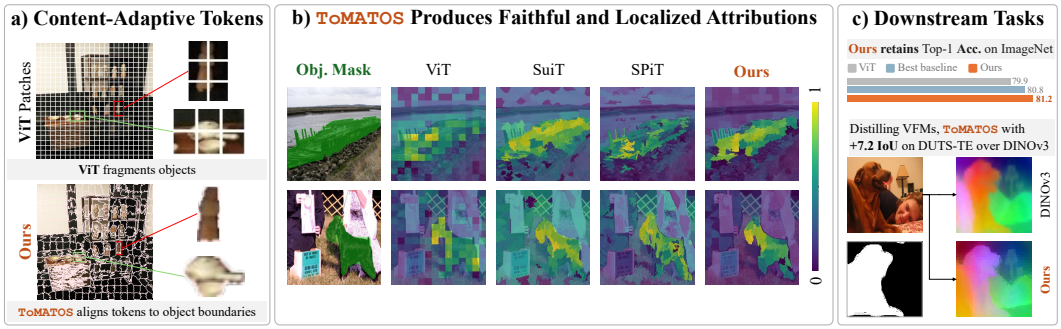

Figure 1: **We introduce ToMATOS, a content-adaptive tokenization method that aligns tokens to object and part boundaries.** **(a)** Standard ViT patches cut across boundaries and fragment objects, whereas ToMATOS tokens follow them. **(b)** Since each token corresponds to a coherent region, attributions computed over ToMATOS tokens are *faithful* and *localized*: they cover the object instead of scattering over a patch grid (ViT) or bleeding into the background (SuiT, SPiT). **(c)** ToMATOS retains ImageNet Top-1 accuracy and, when used as a distillation student, improves TokenCut Segmentation results by $+7.2$ over its DINOv3 teacher.

Abstract

Fixed-grid tokenization in vision transformers partitions images independent of their visual structure and disregard visual boundaries. This mismatch to actual image semantics limits their representation ability, with information of different object and background features bleeding into a single token. We introduce ToMATOS, an efficient, differentiable tokenization framework that represents images using content-adaptive region tokens with granular control. ToMATOS combines a lightweight feature extractor with a GPU-efficient feature clustering, resulting in a flexible and semantically meaningful tokenizer. We evaluate ToMATOS on supervised image classification and distilled vision foundation models of DINO variants over zero-shot segmentation tasks. In experiments ToMATOS demonstrates improved accuracy and attribution faithfulness. Notably, distilled region-token models outperform their foundation-model teachers, e.g. DINOv3, on sematic tasks highlighting the benefits of region-based tokens.

1 Introduction

Vision Transformers (ViTs) represent images as sets of tokens, and how these tokens are formed directly affects the representations learned downstream. The standard approach divides an image into a regular grid of non-overlapping patches and embeds each patch into a token. This design is simple and efficient, supporting increasingly powerful vision foundation models such as the DINO family (Oquab et al., 2023; Siméoni et al., 2025). However, the grid is fixed independently of image content: a patch can cross object boundaries and mix different visual structures, while a coherent region can span multiple patches. Thus, although patch tokens can encode rich semantics, their spatial support is determined by an arbitrary grid rather than by the image structures they represent.

This limitation affects not only representation quality but also interpretability. When an entity (e.g., an object) is distributed across several patches, its representation is spread across multiple tokens

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.

rather than encoded by a coherent spatial unit. Thus, token-level attribution can produce fragmented explanations that do not clearly correspond to the underlying image content. This motivates a natural design goal where *tokens should adapt to image content and correspond to coherent visual regions.* Region-aligned tokens can reduce the mixing of unrelated content within each token while providing a direct spatial correspondence between representations and the structures they encode.

Prior work has explored content-adaptive tokenization using various methods spanning superpixels, learned boundaries, and hierarchical clustering (Lew et al., 2024; Aasan et al., 2024; 2025; Chen et al., 2025). However, existing approaches often sacrifice properties needed for end-to-end representation learning. Some rely on hand-crafted or CPU-based segmentation procedures that are decoupled from the learned features (Lew et al., 2024); others use separate boundary predictors (Chen et al., 2025) or non-differentiable grouping operations (Aasan et al., 2024), preventing gradients from jointly shaping the regions and their representations. Hierarchical approaches further lack direct control over the number of resulting tokens (Aasan et al., 2025). These limitations make it difficult to simultaneously obtain content-adaptive tokens that are efficient to compute, trainable end-to-end, compatible with standard vision architectures, and controllable at inference time.

To address this, we introduce `ToMATOS`, a content-adaptive tokenization framework that constructs spatially coherent region tokens while retaining the simplicity and flexibility of the standard ViT interface. `ToMATOS` combines learnable dense image features with a differentiable region estimation. Inspired by non-iterative superpixel clustering, our tokenizer efficiently constructs regions directly on the GPU and replaces hard pixel-to-region assignments with soft assignments, making the tokenization differentiable. Notably, our representation provides *direct control over token count* rather than obtaining an implicitly determined number of regions from a learned boundary map or hierarchical criterion. This enables the token granularity to be adjusted independently of the input resolution and downstream computational budget. In addition to constructing semantic image regions, `ToMATOS` can retain strong visual representations of modern foundation models. Distilling dense representations from the DINO family into our region-token architecture provides both semantic features and spatially coherent regions.

We evaluate `ToMATOS` from both the representation-learning and interpretability perspectives. Across supervised image classification, zero-shot segmentation, and one-shot segmentation, we examine if content-adaptive region tokens provide useful downstream representations with better interpretability, i.e, more localized and faithful explanations of model predictions. We complement these results with systematic study of the relationship between token count, spatial granularity, latency, and downstream performance, demonstrating how `ToMATOS` exposes an explicit accuracy-efficiency trade-off that is difficult to obtain with existing adaptive tokenization methods. We make code and models publicly available after acceptance. In summary, our contributions are:

- **Learning region-aligned visual representations.** We introduce `ToMATOS`, a GPU-efficient, differentiable tokenizer that constructs spatially coherent regions directly from learned image features, enabling region-aligned tokens without sacrificing the standard ViT compatibility or downstream classification performance.
- **Improved semantic alignment and attribution.** We demonstrate stronger alignment of token boundaries with semantic feature edges, more localized class explainations and attribution faithfulness compared with the evaluated path-based and adaptive tokenizers.
- **Foundation-model representations with adaptive tokens.** We transfer dense representations from DINO-family vision foundation models into our region-token architecture and evaluate their structure alignment capabilities in zero-shot and one-shot segmentation tasks.

2 Related Work

Patch-based Tokenization. ViTs (Dosovitskiy et al., 2020) partition an image into non-overlapping patches, which are essentially a fixed-size grid of visual tokens. This simple design has been widely used, notably in vision foundation models such as DINO (Caron et al., 2021; Oquab et al., 2023; Siméoni et al., 2025), whose patch features emergently encode rich semantics. Nevertheless, the fixed grid determines the spatial boundary of each token independently of image content, and thus a single object can be split into many redundant tokens. `ToMATOS` moves away from this fixed spatial

grid of visual tokens and instead represents image tokens as content-adaptive local regions, which directly translates to improved performance on zero-shot segmentation (c.f. Sec. 4.3).

Superpixel Tokenization. Recent works have explored representing visual tokens as adaptive image regions. SPiT (Aasan et al., 2024) decouples tokenization from feature extraction and derives regions through a hierarchy of parallel graph edge contractions over pixel features, but relies on discrete clustering for region assignment. SuiT (Lew et al., 2024) instead aggregates pixel-level local and positional features within superpixels obtained from Fast-SLIC, which relies on CPU-based computation. SPFormer (Mei et al., 2025) learns soft pixel-superpixel associations with bidirectional cross-attention inside the network, but its iterative association step is computationally heavy. EPOC (Chen et al., 2025) learns object boundaries with a lightweight network and applies watershed segmentation to obtain subobject regions, requiring a separately trained boundary predictor and a non-differentiable segmentation step. ∂HT (Aasan et al., 2025) builds a hierarchy of visual regions through graph-based feature aggregation and merging, introducing differentiable token embeddings for end-to-end training, but does not provide direct control over the number of resulting regions. These approaches share the goal of aligning token boundaries with image content, yet each relies on a discrete or multi-stage region formation step that limits end-to-end differentiability, efficiency, or granular control. `ToMATOS` addresses these limitations by integrating region formation directly into a differentiable tokenizer with GPU-efficient clustering and explicit token count control.

Interpretability and Faithfulness of Visual Tokens. Prior work has investigated whether visual representations provide *faithful* and *interpretable* explanations of model decisions (Aasan et al., 2024; Bousselham et al., 2025; Otsuki et al., 2025; Ribeiro et al., 2016). Emergent attention maps of self-supervised ViTs (Oquab et al., 2023; Siméoni et al., 2025) provide a *direct* view of token-level attributions, yet these tokens cannot represent object boundaries more precisely than the patch grid itself. Post-hoc attribution methods, e.g., LeGrad (Bousselham et al., 2025), instead estimate how much individual input regions contribute to a model's prediction. We evaluate token attributions for both spatial localization and faithfulness, and compare against established region-based attribution protocols (Aasan et al., 2024). We evaluate these attributions for both spatial localization and faithfulness, and compare against established region-based attribution protocols (Aasan et al., 2024).

3 Method

Fixed-grid tokenization ties each visual token to a predetermined spatial partition. A token thus can aggregate pixels across visual boundaries, while a coherent image region may be fragmented across several tokens. We introduce `ToMATOS`, an efficient content-adaptive tokenizer. Given an input image and desired number of regions, `ToMATOS` first extracts lightweight features from an input image section 3.1 and then efficiently constructs a spatial partitioning of these features according to feature similarity sections 3.2 and 3.3. Lastly, initial representations within a region are aggregated and define a region-level token representations section 3.4. We provide an overview in fig. 2.

3.1 From Images to Spatial Features

Let an image be a signal $\mathbf{x} : \mathcal{V} \to \mathbb{R}^3$ on the grid graph $G = (\mathcal{V}, \mathcal{E})$, where $|\mathcal{V}| = N = HW$ and edges join horizontally or vertically adjacent pixels. A labeling $L : \mathcal{V} \to \{1, \dots, K\}$ defines regions $\mathcal{R}_k = \{i : L_i = k\}$, whose nonempty members partition the image. Tokenization associates each region with a descriptor $\mathbf{t}_k \in \mathbb{R}^D$ (D being the feature dimension of the transformer). Unlike a fixed patch grid, this representation permits different region shapes and areas at the same token budget. A lightweight convolutional encoder produces features $\mathbf{h}_i = f_\theta(\mathbf{x})_i$ at the input resolution. Separate projections yield clustering features and features for token construction:

$$\mathbf{f}_i = g_\phi(\mathbf{h}_i), \qquad \mathbf{v}_i = e_\psi(\mathbf{h}_i) + \mathrm{PE}(\mathbf{p}_i), \tag{1}$$

where $\mathbf{p}_i = (y_i, x_i)$ is the pixel position, $\mathbf{f}_i \in \mathbb{R}^{C'}$, and $\mathbf{v}_i \in \mathbb{R}^c$ where C' and c are the channel dimensions for the clustering projection feature and the token construction feature respectively. The Fourier positional encoding PE is evaluated at normalized coordinates before aggregation, allowing a region descriptor to retain information about its spatial support.

3.2 Flash-SNIC: Efficiently learning input-aware image partitioning

Based on the previously defined spatial features, we next assign pixels to a partitioning of the image into local regions. For that, we use local label propagation, inspired by the region-growing approach

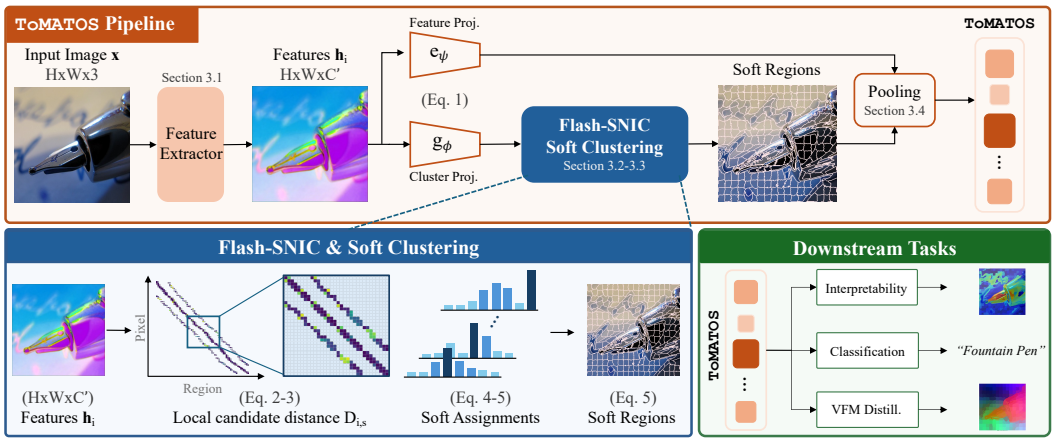

Figure 2: **Overview of ToMATOS. Top:** Our pipeline includes a lightweight feature extractor mapping the input image $\mathbf{x}$ to dense features $\mathbf{h}_i$, which are projected into clustering features $\mathbf{f}_i$ and token features $\mathbf{v}_i$ (Sec. 3.1, Eq. 1). Flash-SNIC partitions the image into K soft regions in clustering-feature space, and pooling within each region yields ToMATOS tokens (Sec. 3.2–3.4). **Bottom left:** Flash-SNIC grows regions from lattice seeds via local parallel label propagation. Candidate distances $D_{i,s}$ are converted into entropy-regularized soft assignments and sparse pixel-to-region associations $\mathbf{A}$, making the partition differentiable (Eqs. 2–5). **Bottom right:** the resulting tokens plug into standard transformer backbones, while $\mathbf{A}$ maps token features back to pixels for region-aligned attributions and dense vision foundation model (VFM) distillation.

of SNIC (Simple Non-Iterative Clustering) (Achanta & Susstrunk, 2017). We define initial seeds distributing on an aspect-ratio-aware lattice. Let s_k denote the pixel of seed k, The assignment cost of pixel i to seed k combines feature similarity and spatial proximity as

$$d(i,k) = \|\mathbf{f}_i - \mathbf{f}_{s_k}\|_2^2 + \frac{m^2}{S^2}\|\mathbf{p}_i - \mathbf{p}_{s_k}\|_2^2, \tag{2}$$

where $S = \max(1, \lfloor\sqrt{N/K}\rfloor)$ is the region spacing and m controls compactness. Smaller m permits stronger adherence to appearance; larger m favors spatially compact regions. Initially, only the seeds carry labels. At sweep t, pixel i considers its current label and the valid labels of its four neighbors, collected in $\mathcal{C}_i^{(t)}$. The parallel update is

$$\ell_i^{(t+1)} \in \arg\min_{k \in \mathcal{C}_i^{(t)}} d(i,k), \tag{3}$$

retaining the current label on ties and leaving a pixel unassigned when its candidate set is empty. Thus, labels are introduced through graph adjacency, and an assigned pixel changes label only when its cost decreases. The parallel operations are implemented in a fast triton kernel (Yang et al., 2025; 2026) with a fixed-budget in our *Flash-SNIC*. Pixels left unassigned are attached to the spatially nearest seed and we denote the resulting labeling by ℓ.

3.3 Differentiable Region Assignments

To learn a model together with the tokenizer end-to-end, the pixel to region assignment needs to be differentiable. The discrete labels above are not differentiable. To recover a gradient path to the feature stem we thus instead build a soft assignment. We collect the labels in a 3×3 window around pixel i into nine candidate slots $I_{i,1}, \ldots, I_{i,9}$, retaining repeated labels and replicating border pixels. Let $D_{i,s} = d(i, I_{i,s})$. We then define the soft assignment as an entropy-regularized local decision

$$\mathbf{Q}_i = \arg\min_{\mathbf{a} \in \Delta^8} \left\{ \sum_{s=1}^{9} a_s D_{i,s} + \tau \sum_{s=1}^{9} a_s \log a_s \right\}, \tag{4}$$

where $\Delta^8 = \{\mathbf{a} \in \mathbb{R}^9_{\geq 0} : \sum_s a_s = 1\}$, $\tau > 0$. Its solution Q and pixel-to-region association A are

$$Q_{i,s} = \frac{\exp(-D_{i,s}/\tau)}{\sum_{r=1}^{9} \exp(-D_{i,r}/\tau)}, \quad A_{k,i} = \sum_{s=1}^{9} Q_{i,s}\, \mathbf{1}[I_{i,s} = k]. \tag{5}$$

The matrix $\mathbf{A} \in \mathbb{R}^{K \times N}$ is non-negative, sparse, and column-stochastic, with at most nine non-zero entries per pixel. Because the softmax operates on the candidate slots, repeated labels contribute their combined probability mass. Candidate indices are held fixed during differentiation, while their weights remain differentiable with respect to the clustering features. If all candidates identify the same region, the association is exactly one regardless of the distances. Assignment learning is therefore concentrated near boundaries where different regions compete. The soft-assignments can be converted to discrete labels at any time with $L_i = I_{i,\arg\max_s Q_{i,s}}$.

3.4 Region Representations

A simple aggregation would summarize a region without explicitly preserving its internal spatial arrangement. To retain this arrangement, we adapt the bounding-box resampling and learned background blending of ∂HT (Aasan et al., 2025) in our framework.

Learnable Box Pooling. Let B_k denote the tight axis-aligned box of $\mathcal{R}k$, and $\mathcal{S}B_k$ bilinearly sample it on a fixed $P \times P$ lattice. Applying $\mathcal{S}_{B_k}$ to the feature map and region indicator gives

$$\mathbf{P}_k = \mathcal{S}_{B_k}(\mathbf{V}), \qquad M_k = \mathcal{S}_{B_k}(\mathbf{1}[L = k]). \tag{6}$$

Here, $M_k \in [0,1]^{P \times P}$ is occupancy under bilinear sampling. A learned scalar $\beta = \mathrm{sigmoid}(b)$ controls the contribution of the area outside the region:

$$\widetilde{\mathbf{P}}_k = M_k \odot \mathbf{P}_k + (1 - M_k) \odot \big[\beta \mathbf{P}_k + (1 - \beta)\mathbf{m}\big], \tag{7}$$

where $\mathbf{m} \in \mathbb{R}^{c \times P \times P}$ is a learned background prototype shared across regions. This preserves region content while learning how much surrounding context to retain. A linear readout produces the token:

$$\mathbf{t}_k = \mathbf{W}_{\mathrm{box}}[\mathrm{vec}(\widetilde{\mathbf{P}}_k) \,\|\, \boldsymbol{\gamma}_k] + \mathbf{b}_{\mathrm{box}}, \tag{8}$$

where the optional descriptor $\boldsymbol{\gamma}_k$ collects occupancy, an absolute position histogram, or normalized box geometry. These descriptors retain shape and extent information that resampling alone may discard. Box pooling is differentiable through feature values and learned parameters, while its hard memberships and box coordinates are treated as fixed.

The retained associations A also map output token features $\mathbf{Z} \in \mathbb{R}^{K \times D}$ back to pixel space:

$$\mathbf{u}_i^{\mathrm{hard}} = \mathbf{z}_{L_i}, \qquad \mathbf{U}^{\mathrm{soft}} = \mathbf{A}^\top \mathbf{Z}. \tag{9}$$

Hard/discrete allocation is piecewise constant on the regions, so output discontinuities can occur only along region boundaries. Soft allocation interpolates among the local candidate regions and reduces to a gather where all candidates coincide. Both preserve an explicit correspondence between pixel locations and token features without reconstructing a regular patch grid.

4 Experiments

We evaluate whether ToMATOS aligns with meaningful visual units while maintaining task performance. We first evaluate downstream performance of our tokenizer on image classification and compare latency with other region-based tokenizers (c.f. Sec. 4.1). We then examine if the learned tokens align with semantically meaningful regions in Sec. 4.2. This includes attribution faithfulness, localization, and token boundary alignment with foundation-model semantic boundaries. Then we look at the transfer of foundation representations to region tokens, using frozen-feature segmentation to assess spatial grouping and correspondence (section 4.3). Finally, in section 4.4, we study the accuracy–computation trade-off obtained by changing token count and image resolution, showing that ToMATOS is computationally efficient despite non-grid token shapes.

Experimental Setup. We compare Small and Base ViT backbones for supervised classification and attribution, and Base encoders for foundation-model transfer. Comparisons distinguish supervised models from frozen DINO teachers and their distilled ToMATOS counterparts. Classification uses 224×224 inputs unless specified otherwise. Semantic-boundary evaluation uses the training compactness $m = 0.3$ for ToMATOS and approximately matched token budgets, with realized counts reported explicitly. Latency columns refer to model inference including tokenization; attribution computation and downstream segmentation algorithms are excluded. The latency configuration and remaining evaluation details are collected in Appendix B and C.

Table 1: **ToMATOS maintains on-par or better accuracy as ViT models.** Supervised classification top-1 and kNN accuracy (%), parameter counts, and latency. †Results reported by ∂HT, not reproduced in our evaluation. **Best** and second best are highlighted within each group.

Method	Size	#Params	Latency (ms) ↓	ImageNet-1K		ImageNet-ReaL		ImageNet-v2	
				top-1 ↑	kNN ↑	top-1 ↑	kNN ↑	top-1 ↑	kNN ↑
ViT	Small	22.1M	5.7	79.90	77.50	83.90	84.90	68.10	57.40
SuiT	Small	21.7M	16.1	80.82	79.99	86.09	**86.33**	69.53	68.30
SPiT	Small	22.2M	33.2	73.25	70.94	79.22	78.17	59.21	56.61
∂HT†	Small	22.3M	57.8	80.00	77.60	84.10	85.90	69.20	57.20
ToMATOS	Small	22.1M	12.7	**81.24**	**80.07**	**86.23**	85.52	**70.26**	**68.70**
ViT	Base	86.8M	6.5	81.80	80.40	86.70	87.30	71.10	62.10
SuiT	Base	86.0M	25.3	82.24	**81.73**	87.00	87.47	70.85	**70.14**
SPiT	Base	86.8M	35.5	75.96	71.51	82.28	79.20	63.16	58.03
∂HT†	Base	86.9M	58.3	**83.20**	80.30	87.40	**88.20**	**72.80**	62.10
ToMATOS	Base	86.7M	14.7	82.66	80.22	**87.43**	85.26	71.52	68.95

4.1 Image Classification

We test whether region tokens retain discriminative image representations on ImageNet-1K (Russakovsky et al., 2014), comparing ToMATOS with ViT (Dosovitskiy et al., 2021; Touvron et al., 2021), SuiT (Lew et al., 2024), SPiT (Aasan et al., 2024), and ∂HT (Aasan et al., 2025). Table 1 reports top-1 and k-nearest-neighbour (kNN) accuracy for Small and Base models. ImageNet-ReaL (Beyer et al., 2020) evaluates revised annotations of the original validation images, while the matched-frequency ImageNet-v2 split (Recht et al., 2019) tests independently collected images. These evaluations reuse the trained encoders without additional backbone training. Full code of ∂HT was not available, their reported results are marked separately and make for limited comparability.

Results. Region-based tokenizers are on-par on the Image-classification task, with ToMATOS often outperforming the other approaches. The top-1 accuracy results on ImageNet-1K, ImageNet-ReaL, and ImageNet-V2 show a similar pattern, while kNN results are less uniform, with SuiT outperforming ToMATOS at Base capacity. Thus, the spatial benefits of region tokenization coexist with competitive supervised recognition, without a consistent advantage on every classification measure.

Overall, ToMATOS maintains or improves top-1 accuracy over patch ViT and other adaptive tokenizers at both model sizes while reducing inference latency relative to adaptive tokenizers.

4.2 Interpretability and Semantic Alignment

We evaluate three complementary properties: whether attributed content affects the classifier's output, whether class evidence localizes to the annotated object, and whether token boundaries follow changes in a common foundation representation. The first two use LeGrad (Bousselham et al., 2024) as attribution method; the third evaluates the token partitions directly, all summarized in table 2.

Attribution Faithfulness. On ImageNet-1K validation images where all models are in agreement of the predictions, we evaluate comprehensiveness and sufficiency through pixel-space perturbations, following the protocol considered in SPiT (Aasan et al., 2024). Comprehensiveness measures the target-class confidence drop after removing highly attributed content. Sufficiency measures the drop when only that content is retained. We use mean-colour occlusion and area fractions $Q = \{0.01, 0.05, 0.2, 0.5\}$, matching perturbed pixel area across tokenizers. The class being explained is also used to score the perturbed images and is held fixed throughout the perturbations.

Attribution Localization. We compare target-class LeGrad attribution maps with annotated masks from 4 276 images in ImageNet-Segmentation (IN-Seg) (Guillaumin et al., 2014). The energy-based pointing game (EPG) measures the fraction of positive attribution energy inside the foreground (Rao et al., 2023). Foreground IoU, binary mIoU, and pixel accuracy evaluate thresholded maps; binary mean IoU averages foreground and background IoU. These metrics measure spatial localization, complementing the perturbation-based evaluation.

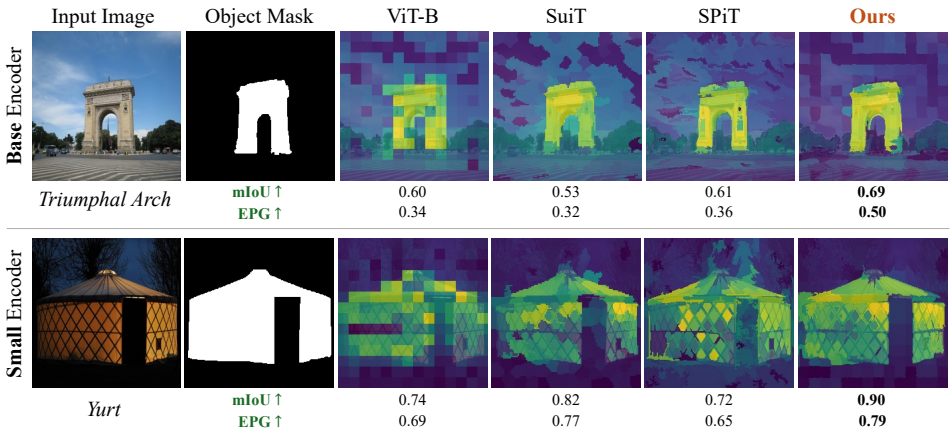

Figure 3: **ToMATOS produces more localized attributions.** LeGrad attribution maps for the target class on ImageNet-Segmentation images, using Base (top) and Small (bottom) encoders. Per-image binary mIoU and EPG (section 4.2) are reported below each map. Patch-grid attributions (ViT-B) are coarse and extend beyond the object, while superpixel-based tokenizers (SuiT, SPiT) better follow object boundaries but retain background evidence. ToMATOS concentrates attribution on the annotated object with minimal background response, yielding the highest mIoU and EPG in both examples and reflecting the localization gains in table 2. See Appendix E for more visualizations.

Table 2: **ToMATOS shows strong attribution localization and semantic boundary alignment.** Faithfulness: Attribution comprehensiveness and sufficiency on IN-1K val. Localization: Attribution maps against IN-Seg masks. Semantic boundaries: precision, recall, and F-score of each model's token borders against TIPSv2-AnyUp feature edges on IN-Seg images. ViT, SuiT and SPiT tokenizers do not depend on model size, so their boundary scores are shared by the Small and Base rows; ToMATOS uses each model's training compactness ($m = 0.3$ Small, $m = 0.7$ Base). Scores are scaled by 100; **best** and second best are highlighted within each group.

Method	Size	Faithfulness: IN-1K		Localization: IN-Seg.				Semantic boundaries: IN-Seg.		
		Comprehensive. ↑	Sufficiency ↓	EPG ↑	FG IoU ↑	Binary mIoU ↑	Pixel Acc. ↑	SB-P ↑	SB-R ↑	SB-F ↑
ViT	Small	44.93	74.31	48.79	44.68	56.90	75.28	34.9	59.4	43.7
SuiT	Small	40.31	**48.58**	52.04	**48.57**	**60.93**	78.67	43.5	91.7	58.8
SPiT	Small	51.25	66.00	49.98	43.04	56.38	75.35	42.5	**95.5**	58.5
ToMATOS	Small	**54.44**	68.77	**55.69**	46.74	60.62	**80.48**	**45.6**	93.2	**61.1**
ViT	Base	39.96	70.04	49.13	45.82	56.75	74.61	34.9	59.4	43.7
SuiT	Base	39.13	**49.03**	43.14	42.84	53.32	71.52	43.5	91.7	58.8
SPiT	Base	46.40	55.82	49.51	44.44	56.14	74.42	42.5	**95.5**	58.5
ToMATOS	Base	**49.66**	63.20	**51.73**	**46.86**	**60.77**	**80.73**	**45.0**	90.3	**59.9**

Semantic-Boundary Alignment. We compare token borders with feature edges from a fixed TIPSv2-L/14–AnyUp reference on ImageNet-Segmentation (Cao et al., 2026; Wimmer et al., 2026). Reference features are extracted at 448-pixel resolution and ℓ_2-normalized after upsampling. Edges are defined by thresholding each pixel's maximum cosine dissimilarity to its four-connected neighbours at the per-image 90th percentile. We report boundary precision (SB-P), recall (SB-R), and F-score (SB-F), averaged over images, using a two-pixel Chebyshev matching tolerance on the common evaluation grid. These scores provide feature-based proxies for semantic alignment. Full definitions and evaluation details are provided in Appendix C.3.

Results. ToMATOS achieves the highest comprehensiveness at both model sizes, exceeding the next-best method by 3.19 points for Small and 3.26 points for Base. EPG also improves by 3.65 and 2.22 points, respectively. The Base model leads all reported localization metrics; Small leads EPG and pixel accuracy while remaining close to SuiT in foreground and binary mean IoU. SuiT performs betterin in suffiency, indicating that the content selected by ToMATOS is consequential when removed but does not preserve the prediction as well when shown in isolation. Since perturbation scores also depend on the classifier's confidence and response to occlusion, we interpret them jointly with localization rather than as an accuracy-independent ranking of models.

Table 3: **ToMATOS remains competitive or improves zero-shot grouping over every DINO teacher.** We compare frozen ViT-B teachers with their ToMATOS students after distillation on YFCC15M. TokenCut measures within-image grouping on ECSSD, DUTS-TE, and DUT-OMRON at 576-pixel resolution; INSID3 measures one-shot correspondence on PASCAL-Part (foreground IoU across all folds). ToMATOS raises TokenCut IoU for all three teachers, with the largest gains on DINOv2 (+10.3 IoU on ECSSD), and improves one-shot correspondence for DINOv1 and DINOv2. All scores are percentages; bold marks the better result within each teacher–student pair.

Encoder	Zero-shot grouping: TokenCut									One-shot:INSID3
	ECSSD			DUTS-TE			DUT-OMRON			PASCAL-Part
	IoU ↑	$F_{\max}$ ↑	Acc. ↑	IoU ↑	$F_{\max}$ ↑	Acc. ↑	IoU ↑	$F_{\max}$ ↑	Acc. ↑	FG IoU ↑
DINOv1	77.05	**88.37**	**93.22**	62.03	**80.91**	**91.11**	56.14	73.73	87.71	47.60
ToMATOS–DINOv1	**78.66**	87.77	93.10	**63.67**	80.48	91.07	**57.84**	**74.38**	**88.04**	**49.19**
DINOv2	64.55	78.68	87.04	51.78	71.75	86.19	43.80	62.58	80.86	50.45
ToMATOS–DINOv2	**74.84**	**86.39**	**90.84**	**55.95**	**76.26**	**86.57**	**49.28**	**68.37**	**82.14**	**52.45**
DINOv3	72.90	86.46	89.94	57.35	78.23	86.82	48.46	69.41	81.73	**55.80**
ToMATOS–DINOv3	**77.10**	**86.71**	**91.47**	**64.58**	**82.37**	**90.72**	**52.90**	**69.87**	**84.62**	51.74

At the main boundary setting, ToMATOS obtain a higher SB-F score than all other methods; SPiT attains higher recall while ToMATOS improves precision. ToMATOS leads SB-F among the available baselines at every reported budget under TIPSv2 with AnyUp at both resolutions.

Overall, ToMATOS aligns token boundaries more closely with semantic feature edges and concentrates class evidence within the target object, with improved comprehensiveness.

4.3 Improving Foundation Representations with Region Tokens

We next check whether we can improve token-level representations and structure alignment of vision foundation models from content-adaptive spatial tokenization. We distill DINOv1 (Caron et al., 2021), DINOv2 (Oquab et al., 2023), and DINOv3 (Siméoni et al., 2025) into ToMATOS using images from YFCC15M dataset Gu et al. (2024). Following the label-free alignment strategy in (Maser et al., 2026), the teacher remains frozen while cosine feature alignment supervises the student's global and intermediate representations (Our student uses region tokens). Table 3 compares each Base teacher with its corresponding distilled student through two frozen-encoder tasks that require spatial grouping or cross-image correspondence as the representative tasks for our analysis.

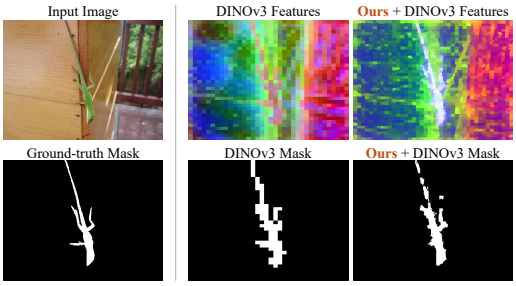

Figure 4: **Region tokens sharpen distilled DINOv3 features.** the DINOv3 teacher's patch features mix information from the lizard and the wall, resulting in a blocky and noisy mask; the student (ours)'s region tokens cleanly separate these, resulting in a crisp mask following the thin limbs and tail, matching ground-truth annotation.

Within-image Grouping (Zero-Shot Segmentation). TokenCut (Wang et al., 2022) constructs a token-similarity graph and applies a normalized cut to discover a salient foreground object. We evaluate on ECSSD (Yan et al., 2013), DUTS-TE (Wang et al., 2017), and DUT-OMRON (Yang et al., 2013), using 576-pixel inputs and reporting IoU, maximum F-measure ($F_{\max}$, $\beta^2 = 0.3$), and accuracy. Token-level predictions are mapped to the image through their patch or region support.

Cross-image Correspondence. We use INSID3 (Cuttano et al., 2026) for **one-shot segmentation** on PASCAL-Part (Chen et al., 2014). Given a support image and binary annotation, INSID3 uses feature correspondence and query-image grouping to recover the indicated part. We report foreground IoU, with both teacher and student frozen and support annotation provided only at inference.

Results. The distilled region models improve TokenCut IoU over their respective teachers in all comparisons. The DINOv2 student, for example, improves ECSSD IoU by about 10 points. These

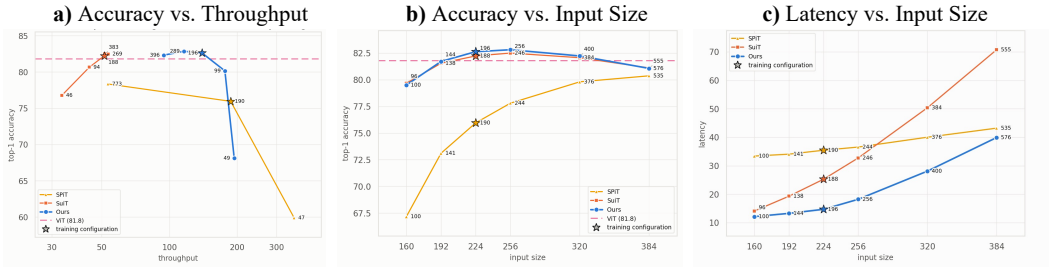

Figure 5: **Controllable tokenization lets a fixed checkpoint trade accuracy for speed, whereas scaling resolution raises cost without a matching gain.** (a) Accuracy–throughput trade-off for Base models at 224×224 as the token budget is varied: ToMATOS retains most of its accuracy over a wide throughput range and dominates SuiT near their standard operating points, while SPiT gains throughput only at a large accuracy cost. (b, c) Accuracy and latency of the same checkpoints across input resolutions at the patch-equivalent token budget: accuracy peaks at intermediate resolutions while latency keeps rising. Labels give realized tokens per image.

results show improved within-image grouping after adaptation to region tokens. They assess the combined effect of distillation and region representation.

One-shot transfer is more sensitive to the teacher representation. DINOv1 and DINOv2 students yield smaller foreground-IoU gains, while the DINOv3 student performs worse relative to its teacher. Among the original encoders, DINOv1 achieves the strongest TokenCut IoU, whereas DINOv3 performs best on INSID3, suggesting that spatial grouping and cross-image correspondence probe different properties of the learned representations. Fig. 4 complement these aggregate results with dense feature maps and paired segmentation examples.

Region-token distillation improves zero-shot segmentation IoU across all comparisons, while improvements in one-shot correspondence depend on the teacher.

4.4 Token Control and Inference Cost

We evaluate how token count and input resolution affect the operating point of a fixed checkpoint, comparing ToMATOS with SuiT and SPiT. Fig. 5 shows that ToMATOS offers a more favourable accuracy–throughput trade-off than SuiT. Reducing the token count initially improves throughput while retaining most of the classification accuracy, but aggressive reduction leads to substantial accuracy loss. SPiT achieves higher throughput at low token counts, but with substantially lower accuracy. Increasing resolution and the corresponding token budget does not consistently improve a fixed checkpoint: accuracy peaks at intermediate resolutions while inference cost continues to increase. These results highlight the value of controllable tokenization for selecting an operating point under a given computational budget.

Explicit token-count control lets a fixed checkpoint operate at different computational budgets, maintaining strong accuracy under moderate token reduction.

5 Conclusion

We introduced ToMATOS, a content-adaptive tokenizer with GPU-efficient region construction and explicit token-count control. Our tokens improve semantic alignment and attribution faithfulness while maintaining competitive classification performance with traditional patch-based ViTs and other adaptive tokenizers. DINO-family distillation consistently improves zero-shot segmentation, showing that spatially structured tokens can support strong transfer within standard vision transformers. We further discuss our limitation and future work in Appendix section F and section G.

AI USE STATEMENT

We used large language models (LLMs) to improve the grammar, clarity, and readability of the manuscript. All research ideas, methods, experiments, analyses, and conclusions are the authors' own. Every AI generated content was thoroughly reviewed and authors take full responsibility for all content of this paper.

REPRODUCIBILITY STATEMENT

We describe the tokenization framework and representation-learning objectives in Section 3, and the experimental setup and comparisons in Section 4. Additional architecture and implementation details, including region construction, soft assignments, and pooling, are provided in Appendix A. Training hyperparameters and distillation configurations are documented in Appendix B, while datasets, preprocessing, and evaluation protocols are described in Appendix C. These details specify the settings used in our experiments and support reproducibility of the reported results. We will release the source code and trained model checkpoints upon acceptance.

REFERENCES

Marius Aasan, O. Kolbjørnsen, A. Solberg, and Adín Ramírez Rivera. A spitting image: Modular superpixel tokenization in vision transformers. In *ECCV Workshops*, pp. 124–142. Springer, Springer Nature Switzerland, 2024. doi: 10.1007/978-3-031-93806-1_11. URL https://arxiv.org/abs/2408.07680.

Marius Aasan, Martine Hjelkrem Tan, Nico Catalano, Changkyu Choi, and Adín Ramírez Rivera. Differentiable hierarchical visual tokenization. In *Advances in Neural Information Processing Systems 38*, pp. 97391–97420. Neural Information Processing Systems Foundation, Inc. (NeurIPS), 2025. doi: 10.52202/085713-2930. URL https://arxiv.org/abs/2511.02652.

Radhakrishna Achanta and Sabine Susstrunk. Superpixels and polygons using simple non-iterative clustering. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4895–4904. IEEE, 2017. doi: 10.1109/cvpr.2017.520. URL https://openaccess.thecvf.com/content_cvpr_2017/html/Achanta_Superpixels_and_Polygons_CVPR_2017_paper.html.

Lucas Beyer, Olivier J. Hénaff, Alexander Kolesnikov, Xiaohua Zhai, and Aäron van den Oord. Are We Done with ImageNet?, 2020. URL https://arxiv.org/abs/2006.07159.

Walid Bousselham, Angie Boggust, Sofian Chaybouti, Hendrik Strobelt, and Hilde Kuehne. LeGrad: An Explainability Method for Vision Transformers via Feature Formation Sensitivity, 2024. URL https://arxiv.org/abs/2404.03214.

Walid Bousselham, Angie Boggust, Sofian Chaybouti, Hendrik Strobelt, and Hilde Kuehne. Legrad: An explainability method for vision transformers via feature formation sensitivity. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1–10. IEEE, IEEE, 2025. doi: 10.1109/iccv51701.2025.01891.

Bingyi Cao, Koert Chen, Kevis-Kokitsi Maninis, Kaifeng Chen, Arjun Karpur, Ye Xia, Sahil Dua, Tanmaya Dabral, Guangxing Han, Bohyung Han, et al. TIPSv2: Advancing vision-language pretraining with enhanced patch-text alignment. In *arXiv.org*, 2026. doi: 10.48550/arXiv.2604.12012. URL https://arxiv.org/abs/2604.12012.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *2021 IEEE/CVF international conference on computer vision (ICCV)*, pp. 9630–9640. IEEE, 2021.

Delong Chen, Samuel Cahyawijaya, Jianfeng Liu, Baoyuan Wang, and Pascale Fung. Subobject-level image tokenization. In *International Conference on Machine Learning*, 2025. doi: 10.48550/arXiv.2402.14327. URL https://openreview.net/forum?id=imkFoKwFwd.

Xianjie Chen, Roozbeh Mottaghi, Xiaobai Liu, Sanja Fidler, Raquel Urtasun, and Alan Yuille. Detect what you can: Detecting and representing objects using holistic models and body parts. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1979–1986. IEEE, 2014. doi: 10.1109/CVPR.2014.254. URL https://arxiv.org/abs/1406.2031.

Claudia Cuttano, Gabriele Trivigno, Christoph Reich, Daniel Cremers, Carlo Masone, and Stefan Roth. INSID3: Training-Free In-Context Segmentation with DINOv3. In *arXiv.org*, 2026. doi: 10.48550/arXiv.2603.28480. URL https://arxiv.org/abs/2603.28480.

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. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=YicbFdNTTy.

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. In *International Conference on Learning Representations*, 2021. URL https://arxiv.org/abs/2010.11929.

Tiancheng Gu, Kaicheng Yang, Xiang An, Ziyong Feng, Dongnan Liu, Weidong Cai, and Jiankang Deng. RWKV-CLIP: A robust vision-language representation learner. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 4799–4812. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.emnlp-main.276.

M. Guillaumin, Dániel Küttel, and V. Ferrari. Imagenet auto-annotation with segmentation propagation. *International Journal of Computer Vision*, 110(3):328–348, 2014. doi: 10.1007/s11263-014-0713-9.

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.org*, 2024. doi: 10.48550/arXiv.2412.04680. URL https://arxiv.org/abs/2412.04680.

Raphael Maser, Siddhartha Gairola, Sukrut Rao, and Bernt Schiele. Align once to explain: Feature alignment for scalable b-cosification of foundational vision transformers. In *MPG.PuRe (Max Planck Society)*. Max Planck Society, 2026. URL https://rmaser.github.io/aloe_project/.

Jieru Mei, Liang-Chieh Chen, Alan Yuille, and Cihang Xie. SPFormer: Enhancing vision transformer with superpixel representation. *arXiv (Cornell University)*, 2025. ISSN 2835-8856. doi: 10.48550/arxiv.2401.02931. URL https://openreview.net/forum?id=nu1SjVgSuy.

Maxime Oquab, Timothée Darcet, Theo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision, 2023. URL https://arxiv.org/abs/2304.07193.

Seitaro Otsuki, Tsumugi Iida, Félix Doublet, Tsubasa Hirakawa, Takayoshi Yamashita, Hironobu Fujiyoshi, and Komei Sugiura. Layer-wise relevance propagation with conservation property for resnet. In *Lecture Notes in Computer Science*, pp. 349–364. Springer, Springer Nature Switzerland, 2025. doi: 10.1007/978-3-031-72775-7_20.

Sukrut Rao, Moritz D Boehle, Amin Parchami-Araghi, and B. Schiele. Studying how to efficiently and effectively guide models with explanations. In *IEEE International Conference on Computer Vision*, pp. 1922–1933. IEEE, 2023. doi: 10.1109/ICCV51070.2023.00184. URL https://openaccess.thecvf.com/content/ICCV2023/papers/Rao_Studying_How_to_Efficiently_and_Effectively_Guide_Models_with_Explanations_ICCV_2023_paper.pdf.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do ImageNet Classifiers Generalize to ImageNet?, 2019. URL https://arxiv.org/abs/1902.10811.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pp. 97–101. Association for Computational Linguistics, 2016. doi: 10.18653/v1/n16-3020.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2014. doi: 10.1007/s11263-015-0816-y. URL https://arxiv.org/abs/1409.0575.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michael Ramamonjisoa, et al. DINOv3, 2025. URL https://arxiv.org/abs/2508.10104.

Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, pp. 10347–10357. PMLR, 2021.

Lijun Wang, Huchuan Lu, Yifan Wang, Mengyang Feng, Dong Wang, Baocai Yin, and Xiang Ruan. Learning to detect salient objects with image-level supervision. In *Computer Vision and Pattern Recognition*, pp. 3796–3805. IEEE, 2017. doi: 10.1109/CVPR.2017.404. URL https://openaccess.thecvf.com/content_cvpr_2017/html/Wang_Learning_to_Detect_CVPR_2017_paper.html.

Yangtao Wang, Xi Shen, Shell Hu, Yuan Yuan, James Crowley, and Dominique Vaufreydaz. Self-Supervised Transformers for Unsupervised Object Discovery Using Normalized Cut. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14523–14533. IEEE, 2022. doi: 10.1109/cvpr52688.2022.01414. URL https://arxiv.org/abs/2202.11539.

Thomas Wimmer, Prune Truong, Marie-Julie Rakotosaona, Michael Oechsle, Federico Tombari, Bernt Schiele, and Jan Eric Lenssen. AnyUp: Universal feature upsampling. In *arXiv (Cornell University)*. Cornell University, 2026. doi: 10.48550/arxiv.2510.12764. URL https://arxiv.org/abs/2510.12764.

Qiong Yan, Li Xu, Jianping Shi, and Jiaya Jia. Hierarchical saliency detection. In *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1155–1162. IEEE, 2013. doi: 10.1109/CVPR.2013.153. URL https://openaccess.thecvf.com/content_cvpr_2013/html/Yan_Hierarchical_Saliency_Detection_2013_CVPR_paper.html.

Chuan Yang, Lihe Zhang, Huchuan Lu, Xiang Ruan, and Ming-Hsuan Yang. Saliency detection via graph-based manifold ranking. In *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3166–3173. IEEE, 2013. doi: 10.1109/cvpr.2013.407. URL https://openaccess.thecvf.com/content_cvpr_2013/html/Yang_Saliency_Detection_via_2013_CVPR_paper.html.

Shuo Yang, Haocheng Xi, Yilong Zhao, Mu-Yang Li, Jintao Zhang, Han Cai, Yujun Lin, Xiuyu Li, Chen-Feng Xu, Kelly Peng, et al. Sparse videogen2: Accelerate video generation with sparse attention via semantic-aware permutation. *Advances in Neural Information Processing Systems 38*, pp. 107568–107594, 2025. doi: 10.52202/085713-3243.

Shuo Yang, Haocheng Xi, Yilong Zhao, Mu-Yang Li, Xiao-yun Fan, Jintao Zhang, Han Cai, Yujun Lin, Xiu-Yu Li, Kurt Keutzer, et al. Flash-kmeans: Fast and memory-efficient exact k-means. *arXiv.org*, 2026. doi: 10.48550/arXiv.2603.09229.

ToMATOS: Tokens Meaningfully Aligned to Object Semantics

Appendix

This supplement provides additional method and implementation details, training settings, evaluation protocols, and qualitative comparisons for ToMATOS. We further examine semantic-boundary alignment across token budgets and reference representations, the contribution of teacher features on region supports, and one-shot segmentation performance across folds.

(A) Method and Implementation Details **14**
- (A.1) Architecture 14
- (A.2) Flash-SNIC Region Construction 14
- (A.3) Rolling PCA and Sparse Soft Assignments 15
- (A.4) Region Pooling Variants 15
- (A.5) Mapping Token Features to Pixels 16

(B) Training Details **16**
- (B.1) ImageNet-1K Classification 16
- (B.2) Distillation from DINO Teachers 16

(C) Evaluation Protocols **19**
- (C.1) Classification 19
- (C.2) Attribution Faithfulness and Localization 19
- (C.3) Semantic-Boundary Alignment 19
- (C.4) TokenCut 20
- (C.5) INSID3 20
- (C.6) Datasets 21

(D) Additional Quantitative Results **21**
- (D.1) Semantic-Boundary Alignment across Budgets and References 21
- (D.2) Teacher Features on Region Supports 22
- (D.3) One-Shot Segmentation by Fold 23

(E) Additional Qualitative Comparisons **23**

(F) Limitations **25**

(G) Future Work **26**

Table A.1: ToMATOS architectures. Parameter counts exclude teachers.

Component	Small classifier	Base classifier	Base student
Feature extractor	Two 3×3 convolutions, 64 channels, full resolution	Same	Same
Clustering map g_ϕ	1×1 rolling PCA, $64 \to 6$	Same	Same
Token features e_ψ	1×1, $64 \to D/2$	1×1, $64 \to D/2$	1×1, $64 \to 48$
Position encoding	Fourier, $\sigma = 10$, added before pooling	Same	Same
Pooling	Soft mean + hard max	Same	Box pooling, $P = 8$
Training tokens	$K = 196$	$K = 196$	Curriculum, 64–1369
Compactness	$m = 0.3$	$m = 0.7$	Curriculum, $1.3 \to 0.3$
Transformer	$D = 384$, 12 blocks, 6 heads	$D = 768$, 12 blocks, 12 heads	$D = 768$, 12 blocks, 12 heads, q/k norm
Prefix tokens	Class + distillation	Class + distillation	Class + 4 registers
Attention position	None	None	2-D RoPE at region centroids
Parameters	22.1 M	86.7 M	88.5 M

A Method and Implementation Details

A.1 Architecture

Table A.1 summarizes the classification models and distilled students. The tokenizer uses a full-resolution convolutional feature extractor, a low-dimensional clustering projection, and a separate feature stream for token construction. Classification models use mean–max pooling, whereas distilled students use learnable box pooling. These variants share the region-construction procedure but provide different paths from the training objective to the tokenizer.

The classification backbones contain a class token and a distillation token. Distilled Base students instead use a class token and four register tokens, together with query/key normalization and two-dimensional rotary positional embeddings evaluated at region centroids. The rotary temperature is 100. Fourier positional features use scale $\sigma = 10$ and are incorporated before region pooling.

A.2 Flash-SNIC Region Construction

Given an $H \times W$ image, a requested token budget K, compactness m, and propagation budget T, seeds are distributed on an aspect-ratio-aware lattice. The nominal spacing is $S = \max(1, \lfloor\sqrt{HW/K}\rfloor)$. For pixel features $\mathbf{f}_i$ and positions $\mathbf{p}_i$, the seed-based assignment cost used in the main formulation is

$$d(i,k) = \|\mathbf{f}_i - \mathbf{f}_{s_k}\|_2^2 + \frac{m^2}{S^2}\|\mathbf{p}_i - \mathbf{p}_{s_k}\|_2^2. \tag{A.1}$$

Only seeds are initially labeled. At each sweep, every pixel considers the valid labels at its own position and its four-connected neighbours. Updates are synchronous: every new label is computed from the preceding sweep. A labeled pixel changes its label only when the cost decreases, preserving its current label on ties. An unlabeled pixel remains unlabeled if no valid candidate is available. After T sweeps, remaining unlabeled pixels are assigned to the spatially nearest seed.

Parallel procedure. For each image, the computation is:

1. Construct the seed lattice and initialize $\ell_i^{(0)}$ to its seed label, or to an unassigned sentinel elsewhere.
2. For $t = 0, \ldots, T-1$, collect valid candidate labels $\mathcal{C}_i^{(t)}$ from pixel i and its four neighbours, and compute

$$\ell_i^{(t+1)} \in \arg\min_{k \in \mathcal{C}_i^{(t)}} d(i,k)$$

in parallel, with the empty-set and tie rules above.

3. Attach remaining unlabeled pixels to their nearest seed in position space.
4. Collect the nine candidate slots $I_{i,1}, \ldots, I_{i,9}$ from the resulting 3×3 label neighbourhood, retaining repeated labels and replicating labels at image borders.

The propagation sweeps are implemented in Triton over the batch. We use $T = 60$ for classification and $T = 45$ for distilled students. This fixes the number of propagation steps rather than testing for convergence; it does not imply identical end-to-end runtime for every image or configuration.

The requested budget controls the seed lattice, while its integer dimensions and subsequent region construction determine the realized number of tokens. We therefore distinguish requested budgets from realized counts in our reports.

A.3 Rolling PCA and Sparse Soft Assignments

The clustering projection is a rolling PCA map to $C' = 6$ channels. Its weights are refreshed every 500 training steps using the leading six eigenvectors of a running feature covariance, with exponential-moving-average momentum 0.95. The running mean is absorbed into the projection bias. The PCA projection is implemented as a linear layer, whose weights can optionally be updated by the optimizer between PCA refreshes. Gradients can pass through the projection to the convolutional feature extractor whenever the loss depends on the soft assignments.

For candidate distances $D_{i,s} = d(i, I_{i,s})$, the local association probabilities are

$$Q_{i,s} = \frac{\exp(-D_{i,s}/\tau)}{\sum_{r=1}^{9} \exp(-D_{i,r}/\tau)}, \qquad A_{k,i} = \sum_{s=1}^{9} Q_{i,s} \mathbf{1}[I_{i,s} = k]. \tag{A.2}$$

Candidate indices are treated as fixed during differentiation. Because labels may occur in several slots, their probabilities are summed to obtain the region association. Selecting the largest slot probability is therefore different from selecting the largest aggregated region probability. The hard labeling used in the supplied implementation notes is

$$L_i = I_{i,s_i^\star}, \qquad s_i^\star = \arg\max_s Q_{i,s}. \tag{A.3}$$

For classifier pooling, the recorded temperature is $\tau = 0.1$. The implementation retains the six largest candidate-slot weights and renormalizes them before aggregation. If $\mathcal{S}_i$ contains these six slots, the retained weights and associations are

$$\widetilde{Q}_{i,s} = \frac{Q_{i,s} \mathbf{1}[s \in \mathcal{S}_i]}{\sum_{r \in \mathcal{S}_i} Q_{i,r}}, \qquad \widetilde{A}_{k,i} = \sum_{s} \widetilde{Q}_{i,s} \mathbf{1}[I_{i,s} = k]. \tag{A.4}$$

Selection of the retained slots is discrete; gradients propagate through their values. Truncation reduces aggregation work, but equality with nine-slot pooling is not a general mathematical guarantee.

A.4 Region Pooling Variants

Classification: mean–max pooling. Let $\mathbf{v}_i \in \mathbb{R}^{D/2}$ denote a pixel's projected feature with positional encoding. The classification token concatenates a soft weighted mean and a hard-region channel-wise maximum:

$$\boldsymbol{\mu}_k = \frac{\sum_i \widetilde{A}_{k,i} \mathbf{v}_i}{\sum_i \widetilde{A}_{k,i}}, \qquad \boldsymbol{\nu}_k = \max_{i: L_i = k} \mathbf{v}_i, \qquad \mathbf{t}_k = [\boldsymbol{\mu}_k \| \boldsymbol{\nu}_k] \in \mathbb{R}^D. \tag{A.5}$$

The mean branch propagates gradients through both feature values and association weights; the maximum branch propagates gradients through selected feature values. The classification objective can therefore influence the clustering features through the soft mean branch.

Distillation: learnable box pooling. Distilled students use the bounding-box pooling construction in the main paper, adapted from differentiable hierarchical tokenization (Aasan et al., 2025). Each region's tight bounding box is resampled onto an 8×8 lattice. The recorded inputs comprise a 48-channel projection of dense features, RGB values, and a region occupancy mask. A learned background prototype controls the contribution of content outside the region. A normalized histogram of absolute region position and box geometry are appended, and a linear projection produces a D-dimensional token.

Hard memberships and box coordinates are fixed during differentiation in the stated box-pooling formulation. Gradients through sampled features still train the shared feature extractor, whose changing statistics also affect subsequent PCA refreshes. A gradient through the soft region weights requires an explicit use of those weights in pooling or a loss; box pooling with fixed hard masks alone does not provide that path.

A.5 Mapping Token Features to Pixels

For output tokens $\mathbf{Z} \in \mathbb{R}^{K_{\text{out}} \times D}$, hard and soft map-back are

$$\mathbf{u}_i^{\text{hard}} = \mathbf{z}_{L_i}, \qquad \mathbf{U}^{\text{soft}} = A^\top \mathbf{Z}. \tag{A.6}$$

Attribution maps, TokenCut masks, and the region visualizations use hard map-back. For INSID3, the DINOv1- and DINOv3-derived students use hard map-back, while the DINOv2-derived student uses soft map-back. When every candidate slot identifies the same region, the corresponding column of A is one-hot and the two mappings coincide.

For uncertainty visualization, we use entropy over aggregated region probabilities,

$$H_i = -\sum_k A_{k,i} \log A_{k,i}, \tag{A.7}$$

with $0 \log 0 = 0$. This entropy vanishes when all candidates identify one region. Slot entropy over $Q_{i,s}$ does not have this property, because repeated copies of the same region can share probability mass.

B Training Details

B.1 ImageNet-1K Classification

Table B.1 summarizes the checkpoint settings. Small and Base models use 224×224 inputs and a nominal training budget of 196 tokens, with compactness 0.3 and 0.7, respectively. Small is trained for 100 epochs and Base for 300, each with five warm-up epochs. Both use hard-label distillation from RegNetY-16GF with weight $\alpha = 0.5$ and a distillation token.

CNN distillation supplies additional supervision beyond image labels. The ViT accuracies reproduced in the main comparison are non-distilled values, so those comparisons do not isolate the contribution of tokenization under identical training. The Base model initializes from our patch-token ViT-B trained on ImageNet-1K, while the Small model initializes randomly.

B.2 Distillation from DINO Teachers

We adapt label-free feature alignment to region-token students, following the global and intermediate alignment strategy of ALOE (Maser et al., 2026). Teachers remain frozen. The recorded objective combines global alignment, intermediate region alignment, a transformer weight-norm penalty, and a student variance term:

$$\begin{aligned}\mathcal{L} = \lambda_g &\left(1 - \cos(\mathbf{z}_S^{\text{cls}}, \mathbf{z}_T^{\text{cls}})\right) \\ &+ \lambda_l \sum_{t \in \mathcal{T}} w_t \frac{1}{|\mathcal{K}|} \sum_{k \in \mathcal{K}} \left(1 - \cos(\widehat{\mathbf{z}}_{S,k}^{(t)}, \widehat{\mathbf{z}}_{T,k}^{(t)})\right) \\ &+ \alpha \sum_{(\mathbf{W}_S, \mathbf{W}_T) \in \mathcal{W}} \left(\|\mathbf{W}_T\|_F - \|\mathbf{W}_S\|_F\right)^2 + \lambda_v \mathcal{L}_{\text{var}}.\end{aligned} \tag{B.1}$$

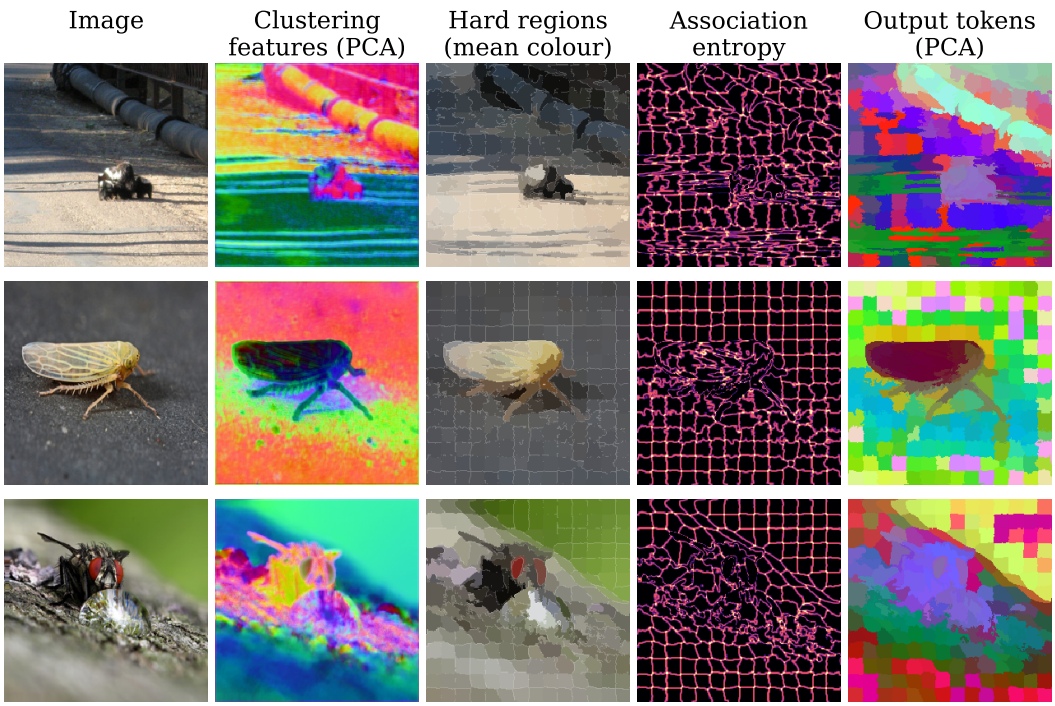

Figure A.1: Intermediate representations of ToMATOS. Panels show the clustering features, hard partition, entropy of the aggregated region associations, and output token features mapped to pixels. PCA colours visualize feature variation and do not denote semantic class labels.

Table B.1: ImageNet-1K classification training settings.

Setting	ToMATOS-S	ToMATOS-B
Input	224×224	Same
Epochs (warm-up)	100 (5)	300 (5)
Initialization	Random	Random
Optimizer	AdamW; linear warm-up, then cosine decay	Same
Learning rate / weight decay	10^{-4} / 0.05	Same
Hardware	4× H100	4× H100
Batch per GPU × accumulation (global)	128×2 (1024)	64×4 (1024)
Augmentation	3-Augment (DeiT III): random resized crop, flip, one of grayscale/solarize/blur, color jitter	Same
Label smoothing / drop path	0.1 / 0.1	Same
Knowledge distillation	Hard labels from RegNetY-16GF, $\alpha = 0.5$, distillation token	Same
Tokens K / compactness m / τ_{SNIC}	196 / 0.3 / 0.1	196 / 0.7 / 0.1
Flash-SNIC sweeps	60	60

Here $\mathcal{K}$ contains valid, nonempty regions, $\mathcal{T} = \{4, 8, 12\}$, $w = (0.25, 1.0, 1.75)$, and $\lambda_g = \lambda_l = 1$. The weight-norm coefficient is $\alpha = 8 \times 10^{-7}$ over 48 corresponding transformer weight matrices. The variance coefficient is $\lambda_v = 0.05$.

Table B.2: Distillation curriculum. Each teacher column gives the image size R and the requested region count K. The compactness is sampled uniformly from the stated range for every batch; τ_{SNIC} is the soft-assignment temperature.

Stage	Start epoch	DINOv1 (R, K)	DINOv2 (R, K)	DINOv3 (R, K)	m range	τ_{SNIC}
1	0	128, 64	126, 81	128, 64	$[1.2, 1.3]$	0.05
2	5	192, 144	196, 196	192, 144	$[1.0, 1.2]$	0.07
3	11	256, 256	252, 324	256, 256	$[0.8, 1.0]$	0.10
4	18	384, 576	378, 729	384, 576	$[0.5, 0.8]$	0.15
5	26	384, $\{432, 576\}$	518, $\{1024, 1369\}$	512, $\{768, 1024\}$	$[0.3, 0.5]$	0.30

Table B.3: Distillation runs. An epoch is a fixed number of batches, not a pass over the data.

Setting	DINOv1 student	DINOv2 student	DINOv3 student
Data	YFCC15M	YFCC100M subset	YFCC100M subset
Hardware	10× V100S	2× H100	2× H100
Batch per GPU × accumulation (global)	16 × 4 (640)	32 × 16 (1024)	32 × 16 (1024)
Batches per epoch	4000	8000	8000
Learning rate / weight decay / clip	$5 \cdot 10^{-4}$ / 0 / 1.0	Same	Same
Precision	FP16 mixed	BF16 mixed	BF16 mixed
Crop scale	$[0.4, 1.0]$	Same	Same
Checkpoint epoch	32	38	76

Patch-to-region targets. Let P_p denote the support of teacher patch p, and define overlap counts $C_{k,p} = |R_k \cap P_p|$. The teacher target for region k at tap t is

$$\mathbf{y}_{T,k}^{(t)} = \frac{\sum_p C_{k,p} \mathbf{h}_{T,p}^{(t)}}{\sum_p C_{k,p}}. \tag{B.2}$$

We use C for patch-overlap counts to distinguish them from pixel-to-region probabilities A. This operation pools piecewise-constant patch features using their spatial overlap. It is not mathematically identical to bilinear interpolation followed by pooling. The implementation notes report cosine agreement of approximately 0.999 in a diagnostic comparison, with 0.2 versus 8 ms per image; these are diagnostic measurements rather than a universal equivalence or an end-to-end runtime comparison.

Before intermediate cosine alignment, student predictions and teacher targets are divided by the same per-dimension running target scale, with momentum 0.99 and numerical stabilizer 10^{-4}. This is diagonal rescaling rather than full covariance whitening. It increases the relative contribution of low-variance target directions. Empty regions are excluded. Students initialize from the supervised `ToMATOS`-Base model, with newly introduced components requiring their own initialization; even the DINOv1 student contains four register tokens.

Multi-scale curriculum. Stages begin at epochs 0, 5, 11, 18, and 26. Each stage determines image resolution, region budget, and a compactness interval sampled uniformly per batch. Table B.2 gives the recorded schedule. DINOv2 uses resolutions aligned with its patch size of 14. The soft-assignment temperature increases from 0.05 in the first stage to 0.30 in the final stage.

Run configurations. Table B.3 lists teacher identifiers and run settings. The students use YFCC15M dataset.

C Evaluation Protocols

C.1 Classification

We report top-1 accuracy on the full 50,000-image ImageNet-1K validation set, its ReaL relabeling, and the 10,000-image matched-frequency ImageNet-v2 split. Images are resized by the shorter side and center-cropped with crop ratio 0.875. Weighted cosine kNN uses $k = 20$, temperature 0.07, and pre-classifier embeddings. Its reference bank contains 100 training images per class, for $100,000$ images in total.

C.2 Attribution Faithfulness and Localization

Attribution maps. LeGrad aggregates relevance from a selected start block to the final block. Attribution maps are min-max normalized per image. Region models map token relevance through hard region labels, while patch models use their patch support.

Perturbation scores. For an image x, let $c = \arg\max_j p_j(x)$ be the unperturbed predicted class and let M_q retain high-attribution pixels at nominal fraction q. Comprehensiveness and sufficiency are

$$\mathrm{Comp}(x, q) = p_c(x) - p_c(x_{\mathrm{remove}(M_q)}), \tag{C.1}$$

$$\mathrm{Suff}(x, q) = p_c(x) - p_c(x_{\mathrm{keep}(M_q)}). \tag{C.2}$$

The explained class remains fixed after perturbation. We average over $q \in \{0.01, 0.05, 0.2, 0.5\}$ and images. Occluded pixels are replaced by the ImageNet mean colour, with the same resized image supplied to each model.

The mask threshold is the per-image $(1-q)$ attribution quantile, including ties. Thus, equal nominal fractions do not imply equal realized perturbed areas. Piecewise-constant token maps can add entire tied regions at once.

Localization. We use the 4,276-image ImageNet-Segmentation evaluation described by Guillaumin et al. and used in transformer attribution work (LeGrad). Images are resized directly to 224×224; masks use nearest-neighbour interpolation. The predicted class is explained, and the map is binarized at its own mean. EPG is the fraction of attribution energy inside the annotated foreground. Foreground IoU and binary mIoU are calculated from dataset-accumulated foreground/background counts, with binary mIoU averaging the two classes. Pixel accuracy measures the fraction of correctly classified pixels.

C.3 Semantic-Boundary Alignment

We evaluate the spatial partition directly, independently of classifier predictions and attribution maps. All $4,276$ ImageNet-Segmentation images are used. The main reference is TIPSv2-L/14 with AnyUp; additional references use DINOv3-L/16 and DINOv2-L/14 with registers, each with AnyUp or bilinear upsampling (Cao et al., 2026; Wimmer et al., 2026). TIPSv2-L/14 and DINOv2-L/14 receive 448×448 inputs, while DINOv3-L/16 receives 512×512 inputs. All references initially produce 32×32 last-layer patch grids. These are upsampled to the common evaluation resolution, 224×224 or 384×384, and then ℓ_2-normalized per pixel.

Reference feature edges. Let Ω denote the evaluation grid and $\mathbf{f}(p)$ its normalized reference feature. We define local feature variation and the reference edge set as

$$e(p) = \max_{q \in \mathcal{N}_4(p)} \left(1 - \mathbf{f}(p)^\top \mathbf{f}(q)\right), \tag{C.3}$$

$$E = \left\{p \in \Omega : e(p) \geq \mathrm{Quantile}_{0.9}\left(\{e(q) : q \in \Omega\}\right)\right\}. \tag{C.4}$$

Here $\mathcal{N}_4(p)$ contains valid four-connected neighbours. The threshold is computed per image; ties at the threshold are included, so the set need not contain exactly 10% of the pixels.

Table C.1: TokenCut settings of the Base encoders: the block and $q/k/v$ feature of the graph, the graph threshold τ_{TC}, and for ToMATOS the compactness m and soft-assignment temperature τ_{SNIC}. ToMATOS requests $K = 1693$ regions, which the seed lattice realizes as 1681. Negative block indices count back from the final block. —: not applicable.

Encoder	Block	Feature	τ_{TC}	m	τ_{SNIC}
DINOv1	-3	k	0.2	—	—
ToMATOS–DINOv1	-2	k	0.6	1.0	0.1
DINOv2	-2	v	0.3	—	—
ToMATOS–DINOv2	-3	k	0.3	0.05	0.05
DINOv3	-2	v	0.2	—	—
ToMATOS–DINOv3	-2	k	0.5	1.0	0.1

Token boundaries and tolerance. For a hard labeling L, let

$$B = \{p \in \Omega : \exists q \in \mathcal{N}_4(p) \text{ with } L(p) \neq L(q)\}. \tag{C.5}$$

For a pixel set X, define its Chebyshev dilation by

$$X^{(r)} = \left\{ p \in \Omega : \min_{q \in X} \|p - q\|_\infty \leq r \right\}. \tag{C.6}$$

We use $r = 2$ pixels on the evaluation grid at both resolutions. Precision, recall, and F-score are

$$\text{SB-P} = \frac{|B \cap E^{(r)}|}{|B|}, \qquad \text{SB-R} = \frac{|E \cap B^{(r)}|}{|E|}, \qquad \text{SB-F} = \frac{2\,\text{SB-P}\,\text{SB-R}}{\text{SB-P} + \text{SB-R}}. \tag{C.7}$$

Precision measures how many token-border pixels lie near a feature edge; recall measures coverage of reference-edge pixels. Each score is calculated per image and then averaged, so the reported F-score is not computed from the dataset-average precision and recall. The supplied results report image-level standard errors no larger than 0.003 on the $[0, 1]$ scale. Appendix boundary tables retain this scale; multiply by 100 to compare with the main table.

Tokenizer settings. Main-paper comparisons use each ToMATOS model's training compactness. The appendix comparisons use $m = 0.3$ throughout. SuiT uses half-resolution SLIC with compactness 10, followed by nearest-neighbour label upsampling. SPiT uses an available merge level within 30% of the nominal token budget; missing settings are left blank. The patch reference uses a $g \times g$ grid with $g = \mathrm{round}(\sqrt{K})$. These rules produce approximately, rather than exactly, matched token counts.

Interpretation. SB-F measures agreement with the selected feature reference, not with human-annotated semantic boundaries. Its value depends on the encoder, upsampler, threshold, tolerance, and token granularity. Image-guided upsampling also contributes spatial information to the reference. The metric therefore complements object-mask and attribution evaluations; it does not establish complete-object semantics or attribution faithfulness by itself.

C.4 TokenCut

TokenCut builds a similarity graph from a selected query, key, or value projection at a transformer block (Wang et al., 2022). Edge weights are 1 above the selected cosine threshold τ_{TC} and 10^{-5} otherwise. The foreground partition is obtained from the second eigenvector of the normalized graph Laplacian. Predictions are mapped through patch or region support to the annotation resolution. We report IoU, maximum F-measure with $\beta^2 = 0.3$, and accuracy, without a bilateral solver. Table C.1 lists the settings for the DINO teachers and their corresponding students.

C.5 INSID3

We evaluate one-shot part segmentation with INSID3 on four PASCAL-Part folds, each containing 959 episodes with one annotated support image (Cuttano et al., 2026). Encoders remain frozen.

Table C.2: INSID3 settings of the Base encoders: clustering threshold τ and merge threshold. For ToMATOS: region count K, compactness m, and soft-assignment temperature τ_{SNIC}. —: not applicable.

Encoder	τ	Merge	K	m	τ_{SNIC}
DINOv1	0.5	0.1	—	—	—
ToMATOS–DINOv1	0.8	0.1	1024	0.5	0.3
DINOv2	0.8	0.1	—	—	—
ToMATOS–DINOv2	0.8	0.1	1600	0.5	0.1
DINOv3	0.8	0.2	—	—	—
ToMATOS–DINOv3	0.8	0.2	1024	0.5	0.1

Table C.3: Datasets.

Dataset	Split and size	Used for
ImageNet-1K	train 1.28 M; val 50,000	classification, kNN bank, faithfulness
ImageNet-ReaL	val 50,000 (relabeled)	classification
ImageNet-v2	matched frequency, 10,000	classification
ImageNet-Segmentation (Guillaumin et al., 2014)	4,276 images, 445 synsets	attribution localization, semantic boundaries
ECSSD / DUTS-TE / DUT-OMRON	1,000 / 5,019 / 5,168	TokenCut
PASCAL-Part	4 folds × 959 episodes	INSID3
YFCC15M	unlabeled	distillation

Region encoders map their tokens to pixels and average them onto the regular grid expected by the downstream procedure; at 512 pixels the target patch-16 grid is 32×32. The selected configuration is reported in Table C.2.

C.6 Datasets

Table C.3 lists the datasets and evaluation sizes. Distillation uses unlabeled YFCC images, with the exact subset depending on the run. The annotation source used for localization and boundary evaluation is ImageNet-Segmentation.

D Additional Quantitative Results

D.1 Semantic-Boundary Alignment across Budgets and References

Table D.1 reports semantic-boundary precision, recall, and F-score using TIPSv2-L/14 features upsampled with AnyUp. At $m = 0.3$, ToMATOS achieves the highest SB-F at every listed token budget at both resolutions. Its precision is also highest throughout, while recall is competitive with the other adaptive tokenizers. This consistent advantage across token granularities supports the ability of ToMATOS to align region boundaries with coherent structures in the reference representation.

Table D.2 tests whether this advantage depends on the reference encoder and upsampler. ToMATOS achieves the highest SB-F in all six AnyUp comparisons and nine of the twelve displayed settings overall. SPiT leads in the remaining three bilinear comparisons. The agreement across the three AnyUp reference encoders supports the consistency of the boundary alignment, while the bilinear results show that the construction of the feature reference also affects the ranking.

Table D.1: Semantic-boundary precision (P), recall (R) and F-score (F) against TIPSv2-L/14 features upsampled with AnyUp, at both evaluation resolutions. ToMATOS uses $m = 0.3$. SPiT is listed at the merge level whose token count is within 30% of the budget; —: no such level. Bold: best of the four tokenizers per budget.

	ToMATOS			SuiT			SPiT			Patch grid		
Budget	P	R	F	P	R	F	P	R	F	P	R	F
224×224												
49	**.510**	.822	**.628**	.509	.741	.601	.493	**.833**	.616	.359	.316	.334
98	**.481**	**.894**	**.625**	.469	.835	.598	—	—	—	.354	.446	.392
196	**.456**	.932	**.611**	.435	.917	.588	.425	**.955**	.585	.349	.594	.437
392	**.424**	**.974**	**.589**	.410	.969	.574	—	—	—	.346	.774	.475
784	**.395**	.995	**.564**	.383	**.996**	.551	.385	**.996**	.552	.343	.932	.499
384×384												
144	**.461**	.792	**.582**	.450	.728	.554	.435	**.819**	.564	.326	.329	.325
288	**.437**	**.861**	**.578**	.420	.818	.553	—	—	—	.322	.451	.373
576	**.410**	.908	**.564**	.393	.902	.545	.386	**.946**	.545	.320	.603	.416
1152	**.384**	**.963**	**.547**	.375	.958	.537	—	—	—	.317	.774	.447

Table D.2: SB-F for different reference encoders and upsamplers, at 196 tokens (224×224) and 576 tokens (384×384). TIPSv2-L/14 and DINOv2-L/14 (with registers) see the image at 448 pixels and DINOv3-L/16 at 512 pixels, so every reference starts from a 32×32 grid. ToMATOS uses $m = 0.3$. Bold: best per row.

Reference	Upsampler	Resolution	ToMATOS	SuiT	SPiT	Patch grid
TIPSv2-L	AnyUp	224	**0.611**	0.588	0.585	0.437
TIPSv2-L	AnyUp	384	**0.564**	0.545	0.545	0.416
TIPSv2-L	bilinear	224	**0.395**	0.381	0.392	0.336
TIPSv2-L	bilinear	384	0.302	0.297	**0.307**	0.276
DINOv3-L	AnyUp	224	**0.600**	0.577	0.573	0.428
DINOv3-L	AnyUp	384	**0.552**	0.534	0.532	0.407
DINOv3-L	bilinear	224	**0.389**	0.370	0.374	0.305
DINOv3-L	bilinear	384	**0.304**	0.293	0.301	0.254
DINOv2-L-reg	AnyUp	224	**0.618**	0.594	0.593	0.443
DINOv2-L-reg	AnyUp	384	**0.570**	0.551	0.552	0.422
DINOv2-L-reg	bilinear	224	0.440	0.429	**0.441**	0.395
DINOv2-L-reg	bilinear	384	0.335	0.333	**0.339**	0.319

D.2 Teacher Features on Region Supports

To examine the contribution of spatial support without retraining the encoder, we pool frozen teacher features onto regions generated by the corresponding distilled student's tokenizer. Pooling uses patch–region overlap as in Eq. equation B.2. The teacher's block, feature type, and TokenCut threshold remain those of its patch-token baseline. The experiment uses ECSSD at 576 pixels, or 574 for DINOv2.

Table D.3 shows that the distilled students outperform both their patch teachers and all corresponding teacher-on-region configurations. At requested $K = 1693$, region pooling alone improves teacher IoU by 1.22, 4.53, and 0.04 points for DINOv1, DINOv2, and DINOv3, respectively. The corresponding students improve on the patch teachers by 1.61, 10.29, and 4.20 points. These results suggest that region support contributes to the observed improvement, while the learned student representations provide benefits beyond pooling the original teacher features onto regions.

The effect of region support depends on granularity: at $K = 576$, pooling reduces IoU for all three teachers. Moreover, the student rows use their own TokenCut settings, and the regions are produced

Table D.3: TokenCut IoU on ECSSD ($\times 100$) with each frozen teacher's own features pooled onto ToMATOS regions. The regions come from the corresponding distilled student's tokenizer. Block, feature and graph threshold are those of the teacher's patch row. $K = 1693$ is realized as 1681 regions. The last row is the distilled student with its own settings. Bold: highest IoU in each column.

Tokens	DINOv1-B	DINOv2-B	DINOv3-B
Teacher patches	77.05	64.55	72.90
Regions, $K = 576$	76.04	62.55	71.32
Regions, $K = 1024$	77.63	67.89	72.43
Regions, $K = 1296$	77.88	68.83	72.41
Regions, $K = 1693$	78.27	69.08	72.94
Distilled student	**78.66**	**74.84**	**77.10**

Table D.4: INSID3 one-shot foreground IoU ($\times 100$) on PASCAL-Part, per fold (959 episodes each). Only self-supervised teachers and their distilled students are included. Bold: the better mean of each teacher–student pair.

Encoder	Fold 0	Fold 1	Fold 2	Fold 3	Mean
DINOv1	39.24	58.84	35.22	57.10	47.60
ToMATOS–DINOv1	40.05	60.33	36.51	59.87	**49.19**
DINOv2	52.91	49.44	43.23	56.24	50.45
ToMATOS–DINOv2	48.82	58.14	44.91	57.94	**52.45**
DINOv3	58.57	50.10	51.21	63.31	**55.80**
ToMATOS–DINOv3	51.87	49.21	46.14	59.75	51.74
SuiT-DINO	27.65	36.30	27.55	31.50	30.75
ToMATOS–SuiT-DINO	32.15	42.57	30.82	36.14	**35.42**

by trained student tokenizers. This comparison therefore probes the contribution of region support without fully isolating it from representation learning and downstream configuration.

D.3 One-Shot Segmentation by Fold

Table D.4 reports foreground IoU for self-supervised teachers and their distilled students on all four PASCAL-Part folds. Distillation improves mean IoU for three of the four teacher–student pairs: DINOv1 from 47.60 to 49.19, DINOv2 from 50.45 to 52.45, and SuiT-DINO from 30.75 to 35.42. The DINOv1 and SuiT-DINO students improve on every fold, while the DINOv2 student improves on three folds. DINOv3 remains an exception, with student mean IoU 51.74 compared with 55.80 for its teacher. The gains therefore extend to one-shot correspondence for several teachers, but are not uniform across teacher families.

E Additional Qualitative Comparisons

The qualitative comparisons complement the boundary and segmentation results by showing region geometry, attribution maps, and dense predictions on individual images. Attribution comparisons are shown separately for the Small and Base models; the tokenizer figures use ToMATOS-S, and the teacher–student comparisons use the Base encoders reported in the main paper. Examples were selected among images on which ToMATOS or the distilled student compares favourably with the other models, as stated for each figure; they illustrate qualitative differences and are not a representative sample. Aggregate results are reported in Appendix D and the main paper.

Region alignment. Figure E.1 compares adaptive tokenizations and a regular patch grid against the same reference-feature edges on four ImageNet-Segmentation images at 224×224 pixels, with

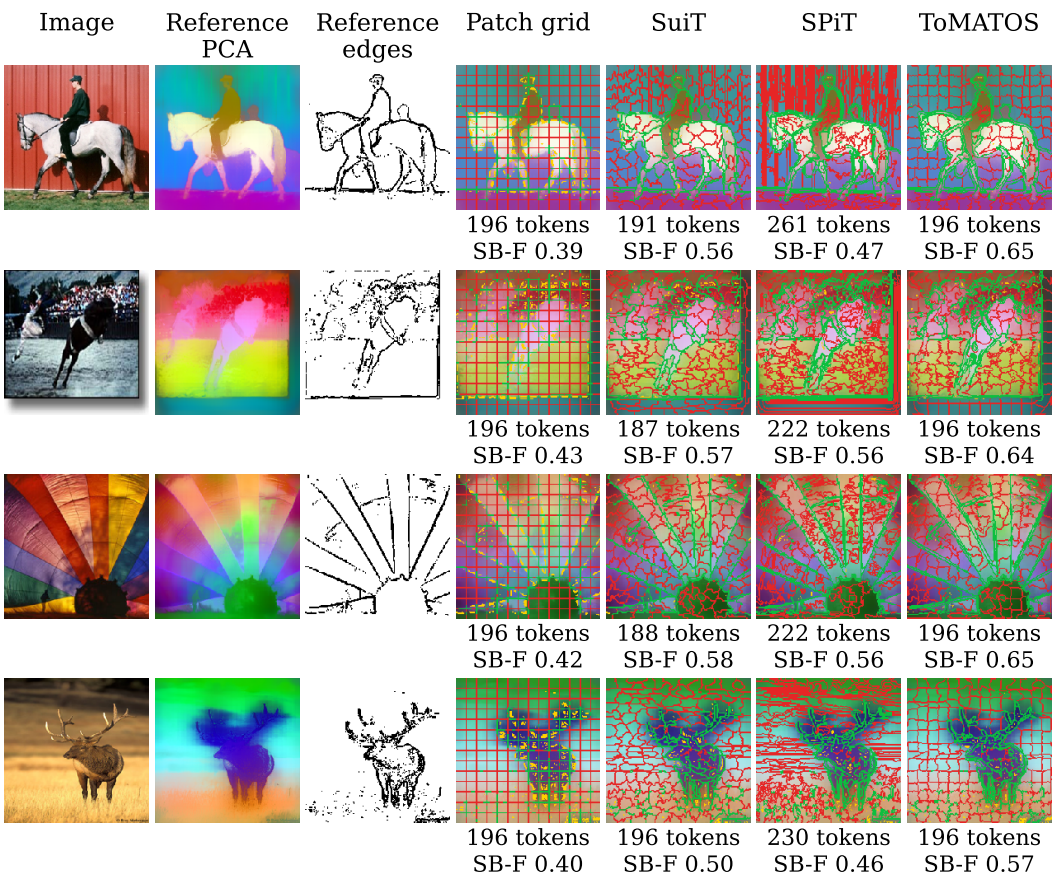

Figure E.1: Token borders compared with TIPSv2-L/14–AnyUp feature edges at approximately 196 tokens (ToMATOS-S, $m = 0.3$). Green denotes a token border within two pixels of a reference edge, red a border away from every reference edge, and yellow a reference edge without a nearby token border. Realized token counts and per-image SB-F are given below each panel.

the ToMATOS-S tokenizer at its training compactness $m = 0.3$ and $K = 196$. The overlays show where token borders follow reference edges, where they introduce additional borders, and which reference edges remain uncovered. The images were chosen where ToMATOS obtains a higher per-image SB-F than SuiT and SPiT.

Compactness and token budget. Compactness controls the relative contribution of spatial distance in Eq. equation A.1, and the requested token count sets the seed lattice. Using the ToMATOS-S tokenizer, Figure E.2 varies $m \in \{0.2, 0.5, 1.0, 1.5\}$ at a fixed budget of $K = 81$ tokens, and Figure E.3 varies $K \in \{32, 64, 81, 128\}$ at fixed $m = 0.5$. Both figures use the same four images at 224×224 pixels, and only the stated parameter changes between columns. Smaller m places more weight on feature similarity, so borders follow object contours and region sizes become more uneven; larger m approaches a regular lattice. At small m, a token can also include small detached fragments, because pixels are reassigned to a neighbouring region whenever its cost is lower. A request of $K = 128$ is realized as 126 regions on the 9×14 seed lattice.

Attribution maps. Figures E.4 and E.5 compare LeGrad maps for ViT, SuiT, SPiT, and ToMATOS at 224×224 pixels, each model using the head and start depth selected on the localization benchmark (Appendix C.2). All four models predict the annotated class on every shown image, so the explained predicted class is also the ground-truth class. The images are those on which ToMATOS attains both the highest per-image binary mIoU and the highest EPG among the four models; two images already shown in the main paper are excluded. Region tokens yield maps that are constant within a token and follow object contours, whereas the patch grid produces block-shaped evidence.

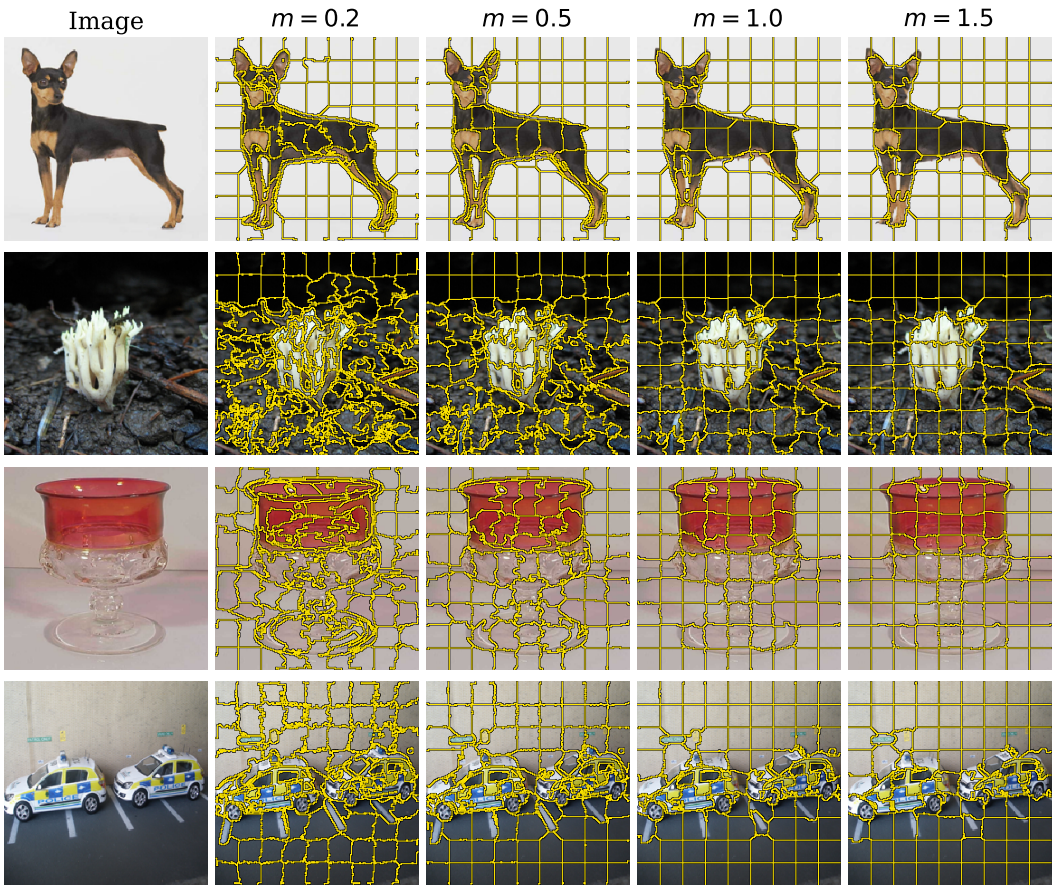

Figure E.2: Effect of compactness on the ToMATOS-S tokenizer at $K = 81$. Only m changes between columns; the checkpoint, image resolution, and token budget remain fixed.

Teacher and student representations. Figures E.6–E.9 compare the Base DINO teachers with their corresponding region-token students. The feature visualizations project the last-layer features onto their first three principal components at 448×448 pixels, with one PCA fitted jointly on the tokens of each teacher and its student. Because each student is distilled onto its teacher's block outputs, the two share a feature space, and matching colours indicate matching feature directions. The images are those on which each student separates the annotated object along its first principal component better than its teacher, for all three teacher–student pairs. The TokenCut examples use the settings of Table C.1; for each teacher, two saliency images with the largest IoU gain of the student are shown, restricted to images on which the teacher attains at least 0.3 IoU.

One-shot segmentation. Figures E.10 and E.11 show INSID3 support–query pairs from PASCAL-Part with the predictions of a Base teacher and its distilled student. The episodes are those with the largest gain of the student over its teacher, among episodes with images of at least 150 pixels on the short side; only one DINOv3 episode meets these conditions, consistent with the lower mean IoU of the DINOv3 student.

F Limitations

ToMATOS constructs content-adaptive regions that provide a more natural spatial support for visual representations. The resulting granularity depends on the token budget and compactness: a region may represent an object part, a larger coherent structure, or a background area. Consequently, region alignment should not be interpreted as a one-to-one correspondence between tokens and complete semantic objects. Our semantic-boundary evaluation measures agreement with dense reference fea-

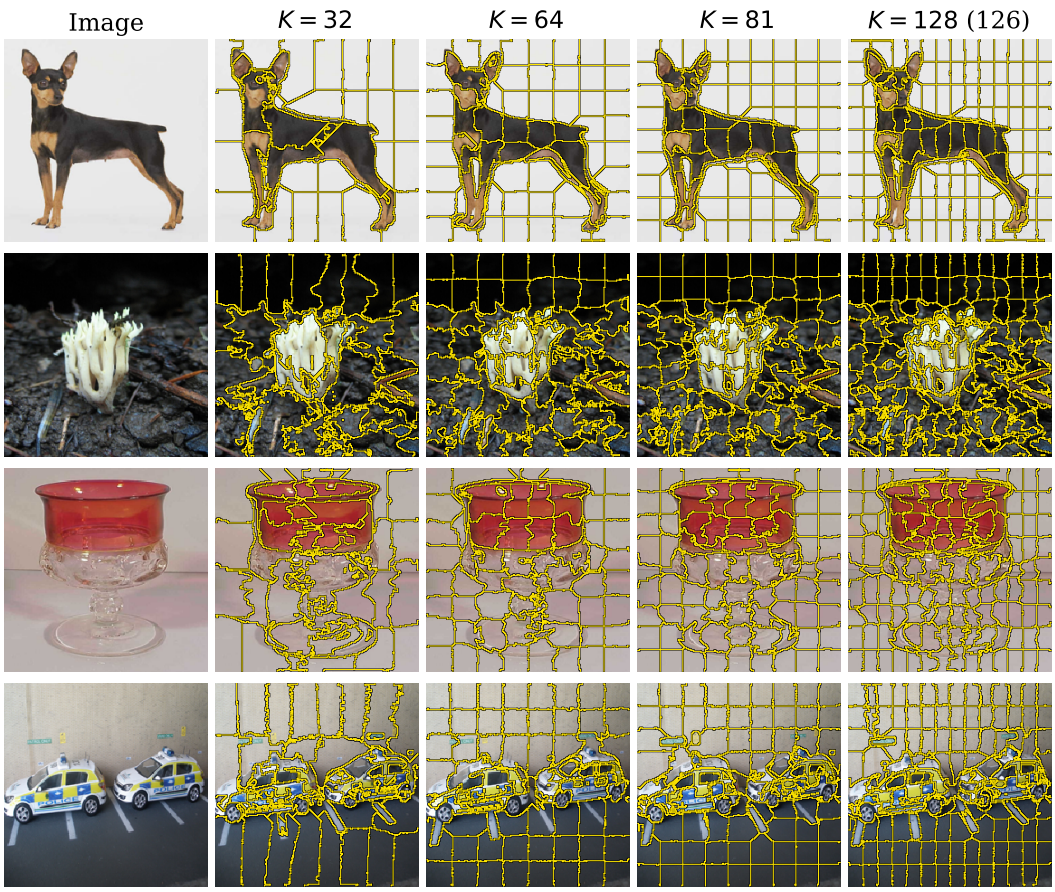

Figure E.3: Effect of the token budget on the ToMATOS-S tokenizer at $m = 0.5$, on the images of Figure E.2. Only K changes between columns; the realized count is given in parentheses where it differs from the request.

tures, whose spatial structure also depends on the reference encoder and upsampling procedure. We therefore examine multiple reference configurations and complement these measurements with downstream evaluations.

The transfer of foundation-model representations also depends on the teacher and downstream task. The observed improvements are not uniform across all teacher–student pairs, suggesting that the teacher's representation and the correspondence requirements of the task influence the benefits of region tokenization. Our evaluation focuses on classification, zero-shot and one-shot segmentation, and attribution; extending these findings to other tasks remains an opportunity for further study.

Finally, reducing the number of transformer tokens does not remove the cost of constructing dense image features and regions. The computational benefit therefore depends on input resolution, token budget, and implementation. Our efficiency results characterize the evaluated configurations rather than implying identical gains across all operating settings.

G Future Work

An important direction is to further adapt region granularity to the input and task. While ToMATOS provides direct control over token count, selecting the budget according to image complexity could allocate additional capacity to detailed scenes while using fewer tokens for simpler inputs. Extending training across a broader range of budgets and resolutions could further support this flexibility.

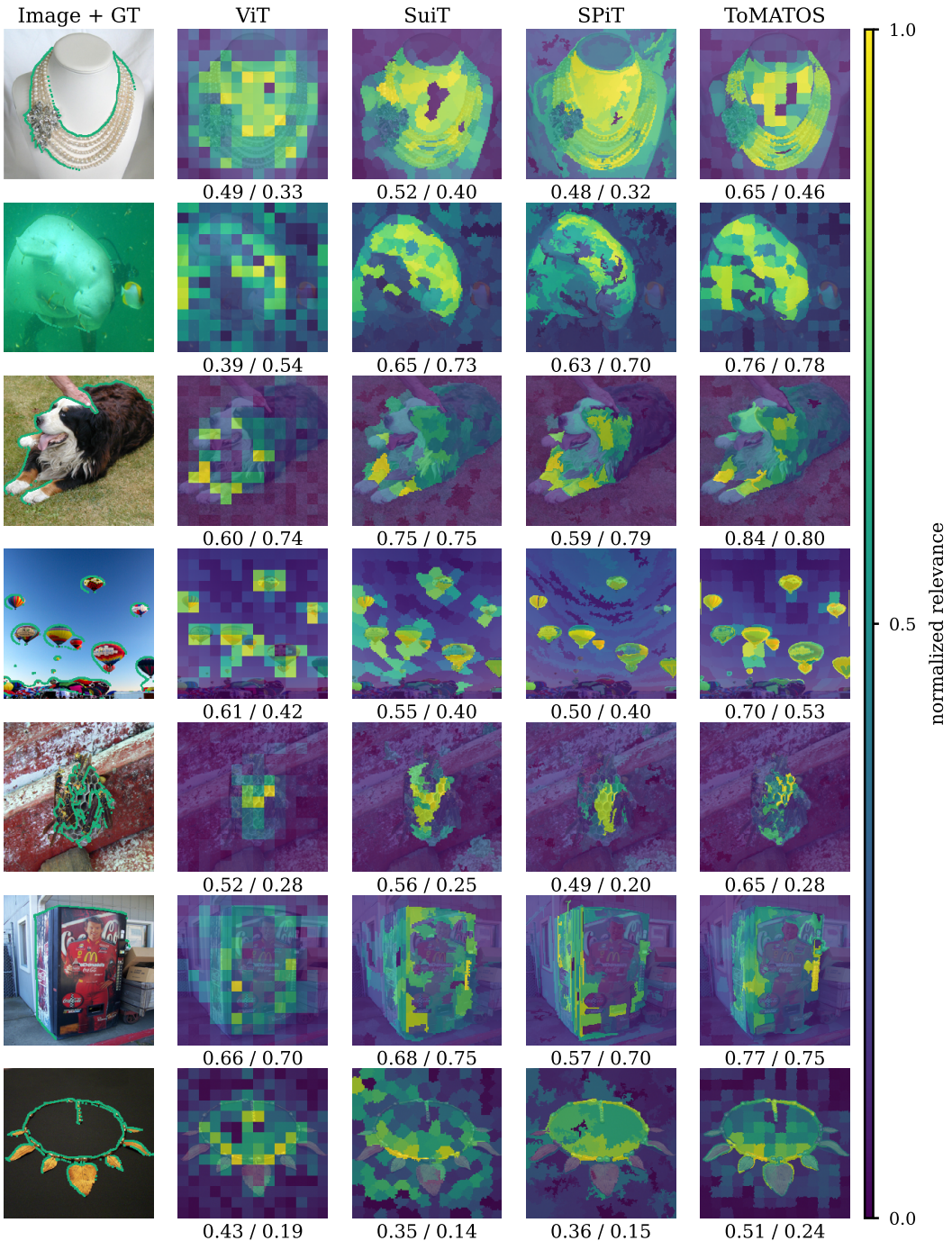

Figure E.4: LeGrad attribution maps of the Small models. The green contour marks the annotated foreground. All maps are min–max normalized per image and shown with the same colour scale. Numbers below each map give the per-image binary mIoU and EPG.

Another direction is to develop distillation objectives that more explicitly preserve cross-image correspondence alongside spatial coherence. This could improve the transfer of region representations across teacher families and dense prediction tasks. Extending region tokenization to video is also promising, where temporal consistency would allow tokens to follow coherent structures across frames.



Figure E.5: LeGrad attribution maps of the Base models, shown as in Figure E.4. Numbers below each map give the per-image binary mIoU and EPG.

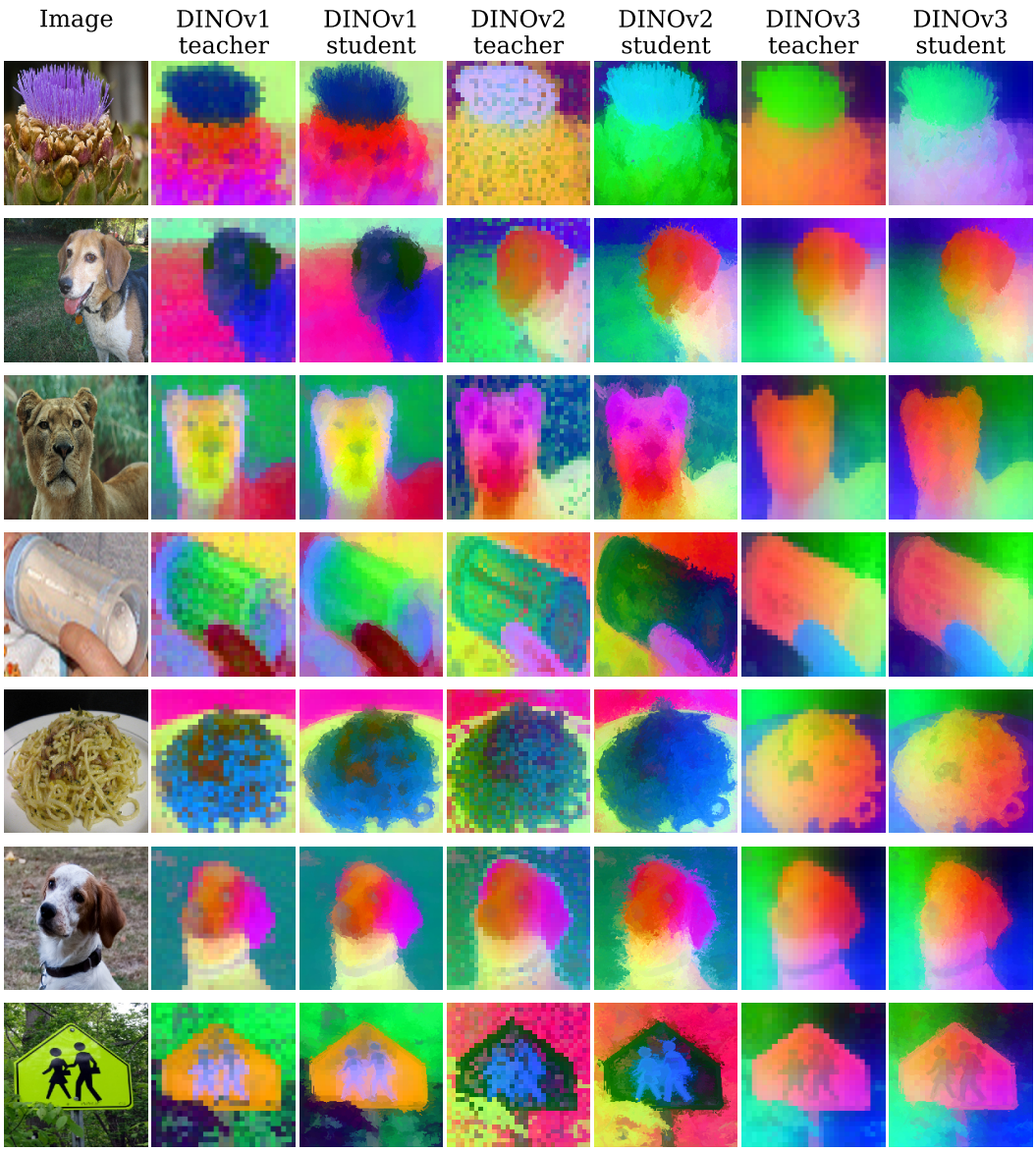

Figure E.6: Teacher and region-student feature visualizations (Base). Teacher features are shown on their native patch grids without interpolation, and student features are mapped to pixels through their region labels. Each teacher and its student share one PCA basis and colour scaling, fitted jointly on both token sets; colours are comparable within a pair but not across pairs.

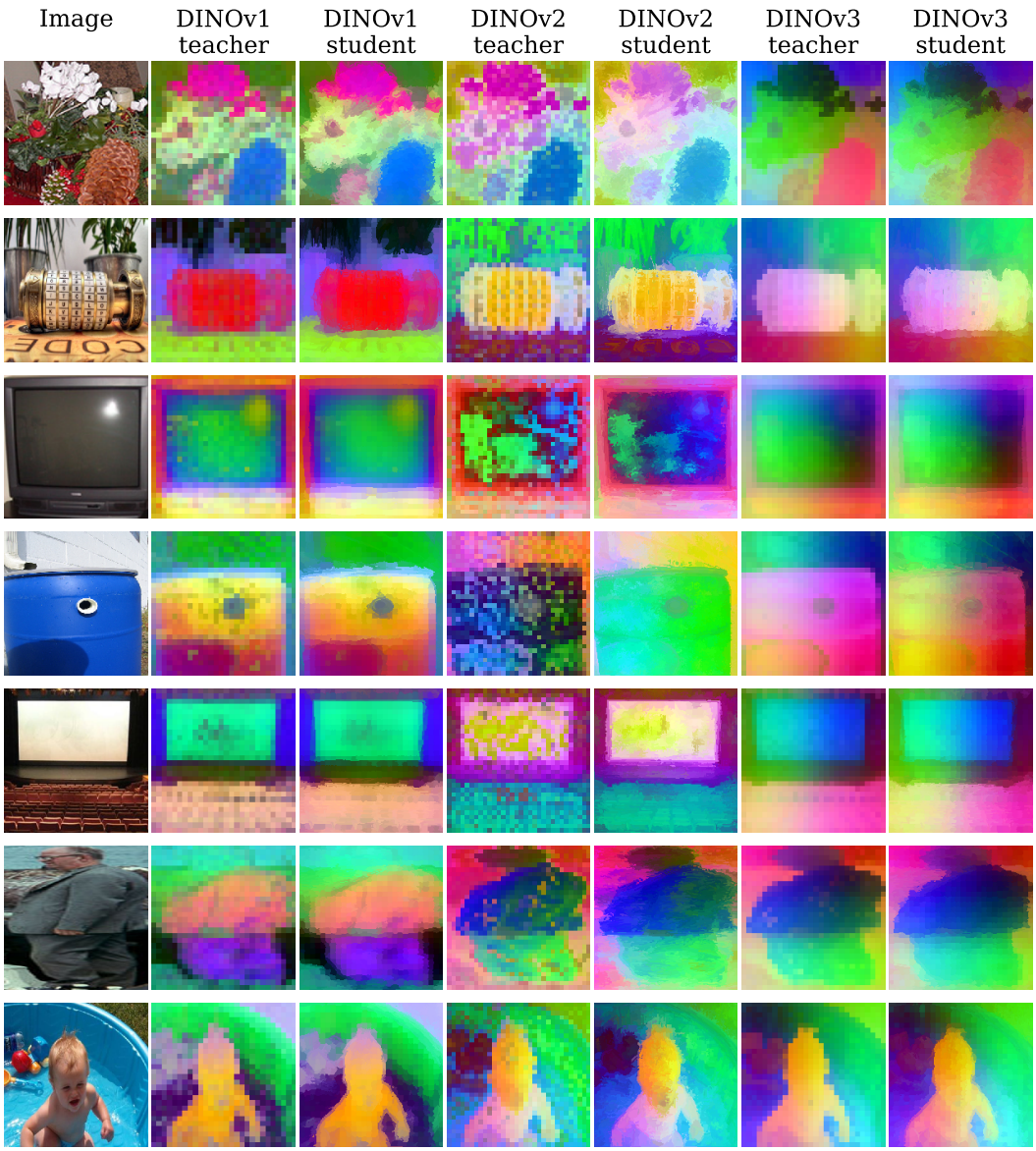

Figure E.7: Further teacher and region-student feature visualizations (Base), shown as in Figure E.6. These images satisfy a relaxed selection: the student separates the annotated object better than its teacher on average and for at least two of the three pairs.



Figure E.8: TokenCut predictions of the Base DINO teachers and their region-token students. Green: annotated foreground; blue: predicted foreground. Per-image IoU is given below each prediction.

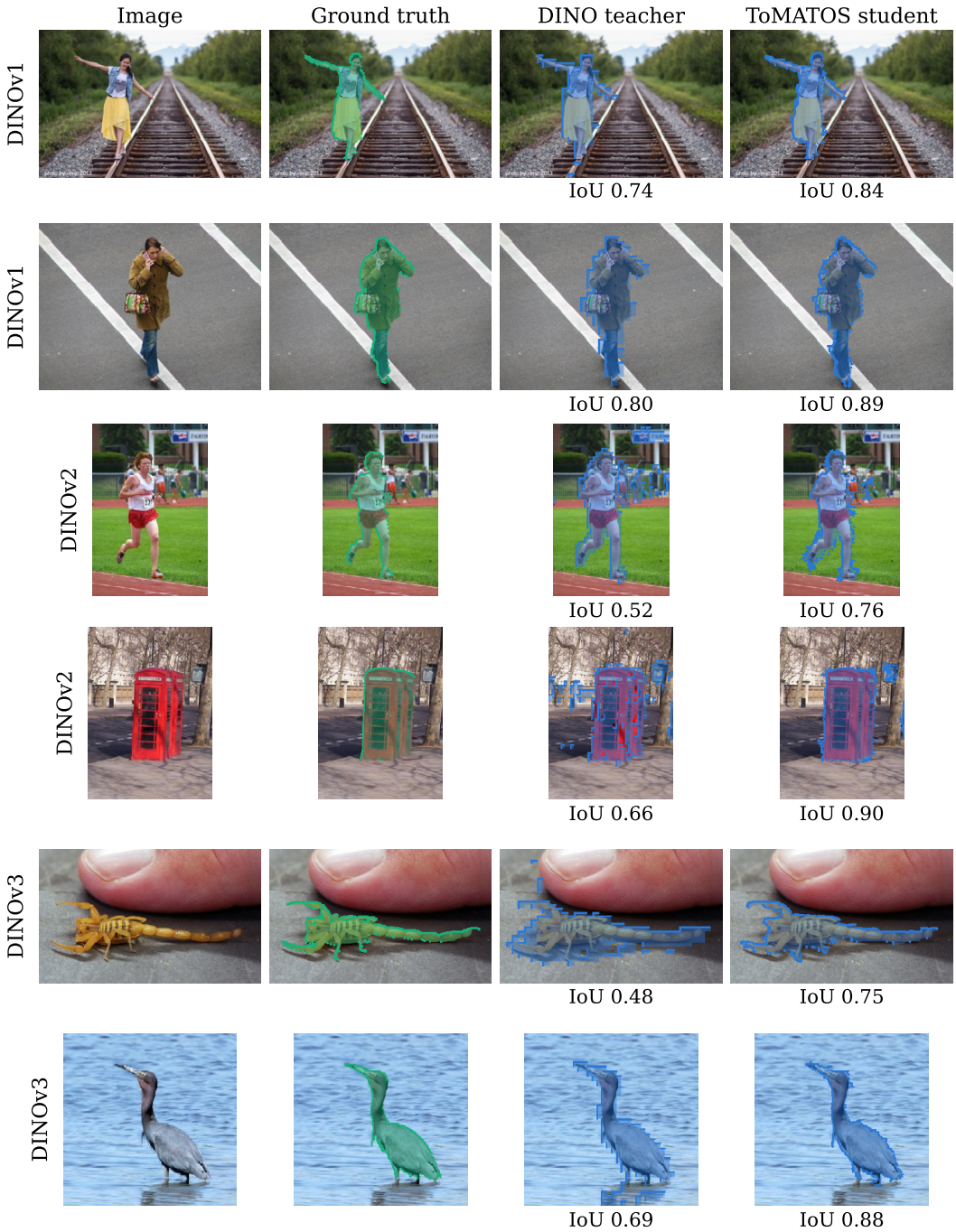

Figure E.9: Further TokenCut predictions of the Base DINO teachers and their region-token students: the next two images per teacher under the selection of Figure E.8. Green: annotated foreground; blue: predicted foreground. Per-image IoU is given below each prediction.



Figure E.10: One-shot PASCAL-Part segmentation with INSID3 (Base encoders). Each row uses one support annotation and query image for both models. Green: support annotation and query ground truth; blue: predicted part. The part class is given below the support image and the per-episode IoU below each prediction.

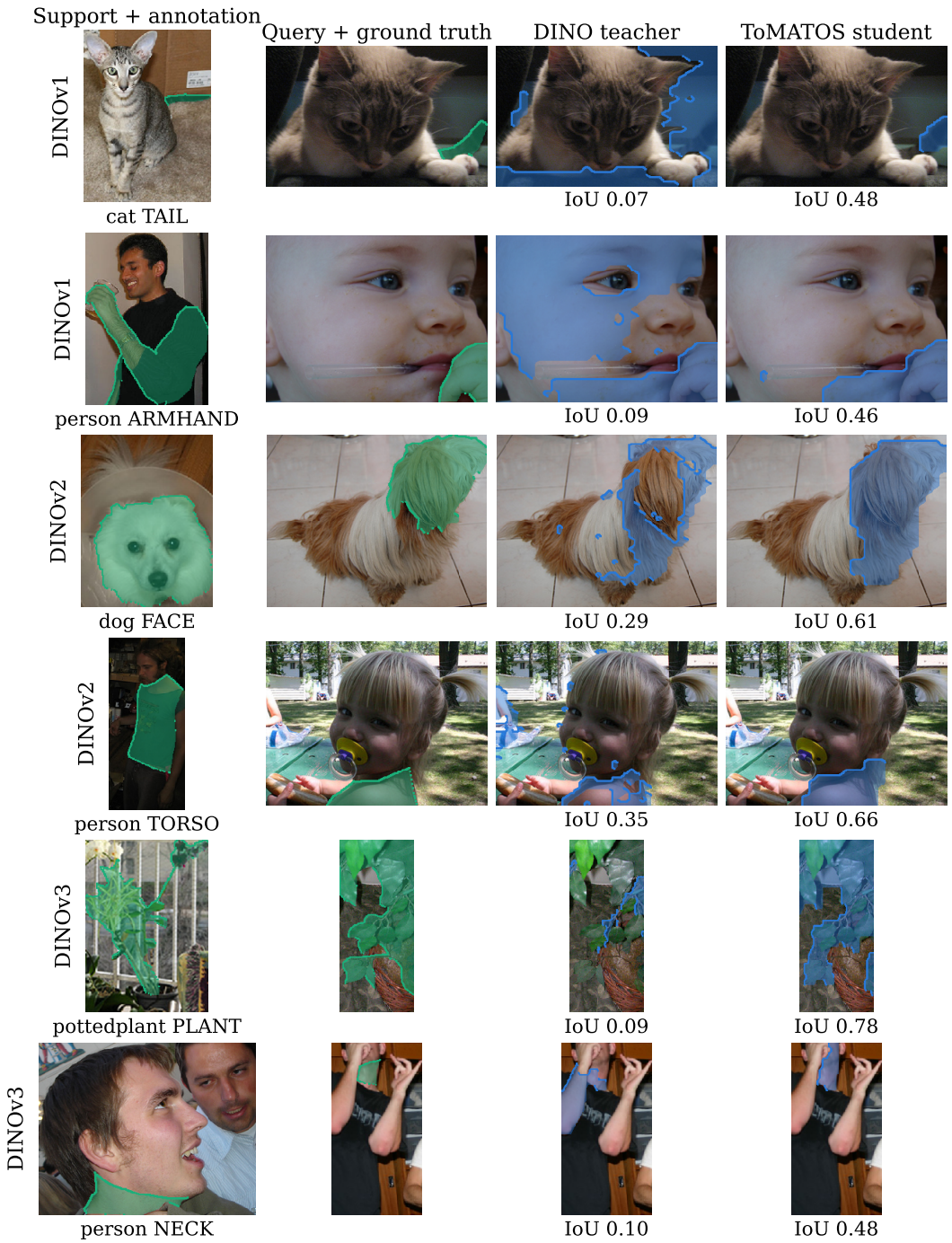

Figure E.11: Further one-shot PASCAL-Part segmentations with INSID3 (Base encoders), two episodes per teacher with the student above its teacher, under relaxed image-size and IoU thresholds. Colours and annotations as in Figure E.10.