DINO-KNN: TRAINING-FREE MULTICLASS IN-CONTEXT SEGMENTATION VIA NEAREST NEIGHBORS

Anonymous authors
Paper under double-blind review

ABSTRACT

In-context segmentation (ICS) segments a target image using one or more annotated reference images. Existing approaches often require task-specific training, combine multiple pretrained models, or rely on multi-stage inference pipelines. Several methods predict one foreground concept at a time. Applying these methods to multiclass segmentation requires combining independently predicted masks, which may assign conflicting labels to the same pixel. We ask whether direct classification of frozen visual features is sufficient for accurate multiclass ICS. We introduce DINO-KNN, a training-free approach that applies k-NN classification to frozen DINOv3 features without a learned segmentation decoder. Given a single reference image annotated with one or more foreground concepts, DINO-KNN classifies target patches using labeled reference features. Given classes compete within a single multiclass decision, producing mutually exclusive predictions without separate per-class segmentation or post-hoc conflict resolution as in state-of-the-art ICS methods. On multiclass episodes from seven datasets spanning natural images, aerial imagery, underwater scenes, and medical images, DINO-KNN achieves the highest average mIoU among the evaluated SOTA methods, surpassing INSID3, the strongest evaluated baseline, by 6 mIoU points. DINO-KNN assigns exactly one semantic label to each pixel. For INSID3, the average percentage of predicted foreground pixels assigned to more than one class is 42%. These results establish direct k-NN classification over frozen DINOv3 features as a simple yet powerful method for training-free multiclass ICS. The code and full setup will be made public.

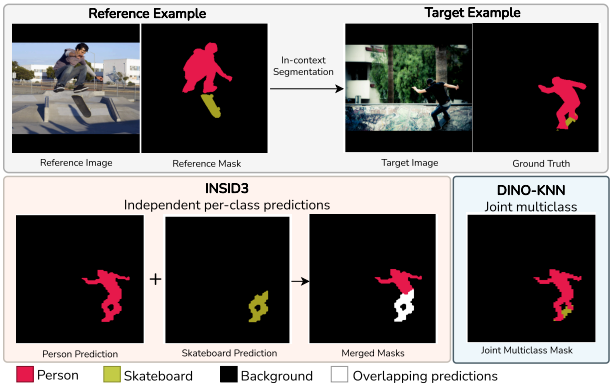


Figure 1: Multiclass in-context segmentation from one annotated reference. **(Top)** The reference image and annotation define the foreground classes; the target ground truth is shown for evaluation only. **(Bottom left)**INSID3 predicts person and skateboard independently, producing overlapping foreground assignments when the masks are combined (white). **(Bottom right)** DINO-KNN jointly assigns one class, including background, to each target patch and produces a mutually exclusive multiclass prediction.

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.

1 INTRODUCTION

Semantic segmentation assigns a semantic label to every pixel in an image. Conventional supervised models are usually trained to predict a predefined set of classes (Long et al., 2015; Cheng et al., 2022; Ronneberger et al., 2015), and obtaining dense annotations at scale is costly Cordts et al. (2016). In-context segmentation (ICS) allows the segmentation task to be specified through annotated reference images. Given one or more examples at inference time, a model segments the corresponding concepts in a target image (Wang et al., 2023b; Meng et al., 2024b). The reference annotations, therefore, provide a flexible way to define the segmentation task, which may involve a single or multiple foreground classes.

Existing ICS approaches use different mechanisms to translate annotated reference examples into target segmentation masks. SegIC Meng et al. (2024b) trains a mask decoder on top of a frozen encoder, while SegGPT Wang et al. (2023b) learns a generalist segmentation model by fine-tuning a pretrained model on diverse segmentation datasets. Its color-based formulation supports multiple semantic classes within a reference example. Both approaches use reference examples without updating model parameters at inference. However, their ability to translate reference annotations into target masks require supervised segmentation training.

Training-free approaches instead combine pretrained components without additional task-specific optimization. Matcher Liu et al. (2024b) and GF-SAM Zhang et al. (2024a) use DINOv2 Oquab et al. (2024) features to establish reference–target correspondences, generate prompts, and obtain masks from the Segment Anything Model (SAM) Kirillov et al. (2023). Evaluating both pretrained models adds inference cost, while prompt generation and mask selection or aggregation introduce additional processing stages. Their binary formulation also handles one foreground class at a time, requiring separate class predictions when applied to multiclass segmentation.

More recently, INSID3 Cuttano et al. (2026) demonstrated that frozen DINOv3 Siméoni et al. (2025) features can support training-free ICS without a learned segmentation decoder or an additional foundation model. However, converting these features into segmentation masks still involves several intermediate stages: fine-grained clustering, candidate localization through backward correspondences, seed-cluster selection, and cluster aggregation. Its binary formulation also predicts one foreground concept at a time. Applying this formulation independently to multiple foreground classes produces separate masks that may assign conflicting labels to the same pixel, requiring an additional combination rule to obtain mutually exclusive multiclass predictions.

This motivates our central question: *Can direct non-parametric classification of frozen DINOv3 features provide accurate, training-free multiclass ICS without clustering or a learned segmentation decoder?* We focus on the one-shot setting, where a single annotated reference image specifies the foreground classes and background. We evaluate episodes with multiple foreground classes to assess multiclass segmentation accuracy and cross-class conflicts.

To this end, we introduce **DINO-KNN**, a training-free framework for joint multiclass ICS based on frozen DINOv3 features. We adopt the positional debiasing procedure introduced by INSID3 Cuttano et al. (2026). Given a single annotated reference image, DINO-KNN uses k-nearest-neighbor (k-NN) classification to predict target patch labels from labeled reference features. Retaining individual reference features preserves the within-class appearance variation captured by the encoder, while aggregating several neighbors may reduce sensitivity to errors in an individual correspondence. All represented classes, including background, compete within a single multiclass decision that assigns each target patch one semantic label. This produces mutually exclusive predictions without generating and merging separate binary masks. Our novelty lies in a simple design for training-free multiclass ICS that replaces the complex multi-stage, class-wise binary pipelines of leading training-free methods.

Our contributions are summarized as follows:

- We introduce **DINO-KNN**, a training-free and decoder-free framework for joint multiclass ICS based on direct k-NN classification of frozen DINOv3 patch features. Its joint classification of classes produces mutually exclusive predictions without cross-class conflicts.

- Across seven datasets spanning natural images, aerial imagery, underwater scenes, and medical images, DINO-KNN outperforms the state-of-the-art training-free methods, surpassing the strongest competitor by 6 mIoU points.
- We demonstrate the practical efficiency of joint multiclass inference. In the evaluated three-class setting, DINO-KNN is approximately $2.8\times$ faster than INSID3 while using 13.7% less GPU memory, avoiding the repeated inference required by class-wise binary methods.

2 Related Work

2.1 Learning-Based In-Context Segmentation

Learning-based ICS methods learn to interpret annotated reference examples through segmentation supervision (Bar et al., 2022; Wang et al., 2023a; Meng et al., 2024a). SegGPT Wang et al. (2023b) formulates segmentation as image inpainting and is trained on diverse segmentation datasets. Its color-based formulation supports multiple semantic classes within a reference example. SegIC Meng et al. (2024b) uses correspondences from a frozen visual encoder to construct instructions for a learned mask decoder. SINE Liu et al. (2024a) combines a frozen image encoder with learned interaction and decoding modules to accommodate different segmentation tasks, including semantic and instance segmentation. These approaches can use reference examples at inference without updating model parameters, but their ability to produce segmentation masks is learned using segmentation annotations. DINO-KNN uses a frozen self-supervised encoder and a non-parametric classifier, without additional segmentation training.

2.2 Training-Free In-Context Segmentation

Training-free methods perform ICS without additional task-specific optimization of pretrained models. PerSAM Zhang et al. (2024b) reuses the image encoder of the Segment Anything Model (SAM) Kirillov et al. (2023) and guides its mask decoder through target-guided attention and semantic prompting. Matcher Liu et al. (2024b) and GF-SAM Zhang et al. (2024a) instead combine a separate DINOv2 Oquab et al. (2024) correspondence encoder with SAM. Matcher generates prompts from dense feature matches and selects the resulting masks, while GF-SAM uses foreground and background correspondences for prompt selection and graph-based grouping for mask aggregation. Evaluating both DINOv2 and SAM adds inference cost to these two methods.

INSID3 Cuttano et al. (2026) obtains segmentation masks from frozen DINOv3 features without an additional segmentation model or a learned mask decoder. Its pipeline combines positional debiasing, fine-grained clustering, candidate localization, seed-cluster selection, and cluster aggregation. Candidate localization uses 1-nearest-neighbor (1-NN) backward matching to restrict the candidate clusters, after which prototype similarity guides seed selection and cluster aggregation.

These methods formulate each prediction as foreground–background segmentation of one reference concept. Applying them independently to multiple classes requires separate class-specific predictions, which can produce overlapping masks that assign conflicting labels to the same pixel. An additional combination rule is therefore needed to obtain mutually exclusive multiclass predictions.

DINO-KNN adopts INSID3's positional debiasing procedure and replaces the remaining segmentation stages with direct k-NN classification of target patches. By aggregating reference labels across foreground classes and background within a single multiclass decision, it produces mutually exclusive predictions without a learned segmentation decoder, clustering, or separate binary-mask merging.

2.3 Non-Parametric Segmentation

Non-parametric segmentation has been explored through prototype matching and nearest-neighbor label transfer. PANet Wang et al. (2019) constructs class prototypes through masked average pooling and classifies query pixels by prototype matching. Its encoder is trained using segmentation supervision and prototype alignment regularization. DINO-KNN uses a frozen encoder and retains individual reference patch features for k-NN classification.

Hummingbird Balazevic et al. (2023) studies dense prediction through nearest-neighbor label aggregation and introduces contextual pretraining to improve representations for retrieval. Its downstream inference keeps the encoder frozen and transfers labels from a reference bank of sampled patch features and annotations. In its reported semantic segmentation experiments, this bank is constructed from augmented images drawn from the full downstream training set or reduced subsets of that set. DINO-KNN focuses on a different reference setting: each episode uses a single annotated image to construct the reference feature bank. We retain all reference patch features and use an existing frozen DINOv3 encoder without additional pretraining or fine-tuning.

FROST Park (2026) performs training-free few-shot segmentation using non-parametric density-ratio estimation over foreground and background features from a frozen DINOv3 encoder. Its binary formulation predicts one foreground concept at a time. DINO-KNN uses k-NN label aggregation to predict all represented classes jointly.

3 DINO-KNN: Multiclass In-Context Segmentation

Given a single annotated reference image and an unlabeled target image, our goal is to segment the target using the semantic classes specified by the reference annotation. We introduce **DINO-KNN** (Figure 2), a training-free and non-parametric framework that performs this task using frozen DINOv3 patch features and k-nearest-neighbor (k-NN) classification. DINO-KNN extracts patch features from both images and applies the positional debiasing procedure adopted from INSID3 Cuttano et al. (2026) (Sec. 3.1). It then associates each reference patch feature with its semantic label and organizes these labeled features into a class-wise reference feature bank (Sec. 3.2). For each target patch, DINO-KNN computes class scores through k-NN voting across the represented foreground classes and background. Assigning each patch the class with the highest score produces a mutually exclusive multiclass mask without a learned segmentation decoder.

Problem formulation. Let $\mathcal{R} = (I_r, \mathbf{Y}_r)$ denote an annotated reference example, where $I_r \in \mathbb{R}^{H\times W\times 3}$ is the reference image and $\mathbf{Y}_r \in \mathcal{C}^{H\times W}$ is its semantic annotation. The episode-specific label set is $\mathcal{C} = \{0, 1, \ldots, C\}$, where 0 denotes background and the remaining labels identify the C foreground classes represented in the reference annotation. Given an unlabeled target image $I_t \in \mathbb{R}^{H\times W\times 3}$, the objective is to predict

$$\widehat{\mathbf{Y}}_t = f(I_t \mid I_r, \mathbf{Y}_r) \in \mathcal{C}^{H\times W}, \tag{1}$$

where f denotes the DINO-KNN inference procedure. Target categories outside the foreground label set specified by the reference are treated as background. The number and identity of foreground classes may vary between episodes.

3.1 Feature Extraction and Positional Debiasing

We extract patch features from the reference and target images using the final output of the frozen DINOv3 encoder Φ Siméoni et al. (2025):

$$\mathbf{F}_r = \Phi(I_r) \in \mathbb{R}^{P\times D}, \qquad \mathbf{F}_t = \Phi(I_t) \in \mathbb{R}^{P\times D}, \tag{2}$$

where D is the feature dimension and $P = h \times w$ is the number of patches.

Reliable correspondence matching requires reference and target patches to be compared according to their visual and semantic content. However, DINOv3 features can also contain structured positional information, causing patches at similar locations to match even when their content differs Cuttano et al. (2026). We therefore adopt the positional debiasing procedure introduced by INSID3 before k-NN classification.

The procedure estimates a positional subspace from features extracted from a noisy image. Specifically, we sample $I^{\text{noise}} \sim \mathcal{N}(0, 1) \in \mathbb{R}^{H\times W\times 3}$ and pass it through the frozen encoder:

$$\mathbf{F}^{\text{noise}} = \Phi\big(I^{\text{noise}}\big) \in \mathbb{R}^{P\times D}. \tag{3}$$

We apply singular value decomposition to these features and retain the top s right singular vectors:

$$\mathbf{F}^{\text{noise}} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^{\top}, \qquad \mathbf{B} = \mathbf{V}_{[:,1:s]} \in \mathbb{R}^{D\times s}. \tag{4}$$

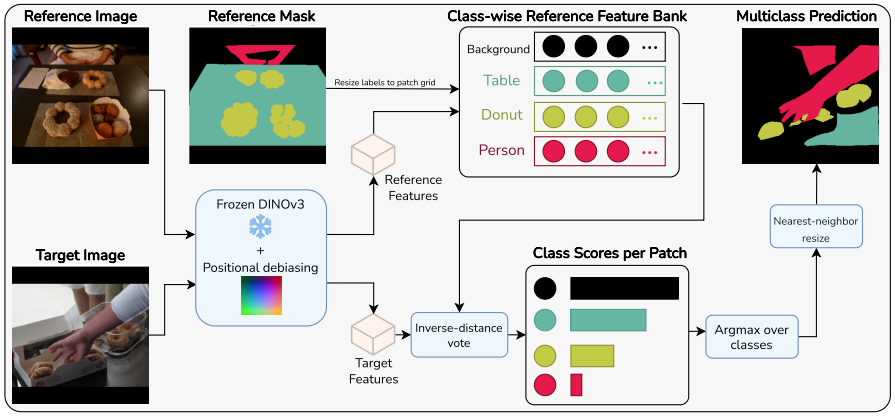


Figure 2: **Overview of DINO-KNN.** DINO-KNN inference. A shared frozen DINOv3 encoder extracts reference and target patch features, which undergo the positional debiasing adopted from INSID3. Nearest-neighbor resizing aligns the reference annotation with the patch grid. All labeled reference patches, including background, form the reference bank. For each target patch, the method retrieves k neighbors in total across all classes using cosine distance, aggregates inverse-distance class votes, and selects the highest-scoring class. If any retrieved distances are zero, only those neighbors vote, with unit weight. Nearest-neighbor resizing maps the predicted patch labels to the original image resolution.

The columns of $\mathbf{B}$ form an orthonormal basis for the estimated positional subspace. We project the reference and target features onto its orthogonal complement:

$$\widetilde{\mathbf{F}}_r = \mathbf{F}_r \left(\mathbf{I}_D - \mathbf{B}\mathbf{B}^\top\right), \qquad \widetilde{\mathbf{F}}_t = \mathbf{F}_t \left(\mathbf{I}_D - \mathbf{B}\mathbf{B}^\top\right), \tag{5}$$

where $\mathbf{I}_D$ is the D-dimensional identity matrix, and $\widetilde{\mathbf{F}}_t, \widetilde{\mathbf{F}}_r$ are debiased features.

3.2 Multiclass k-NN Classification

To align the reference annotation with the feature grid, we resize it to $h \times w$ using nearest-neighbor interpolation. We flatten the resized annotation and associate each debiased reference feature $\widetilde{\mathbf{f}}_{r,i} \in \mathbb{R}^D$ with the semantic label $y_{r,i}$ at the corresponding grid position.

We organize these labeled features into a class-wise reference feature bank:

$$\mathcal{B}_c = \left\{ \left(\widetilde{\mathbf{f}}_{r,i}, y_{r,i} \right) \right\}_{i:\, y_{r,i}=c}, \qquad c \in \mathcal{C}. \tag{6}$$

The bank includes features from both foreground and background classes. Retaining individual patch features allows each target patch to draw on the most similar reference regions. For each target patch j, we compute its cosine distance to every reference patch i:

$$d_{ji} = 1 - \frac{\widetilde{\mathbf{f}}_{t,j}^\top \widetilde{\mathbf{f}}_{r,i}}{\left\| \widetilde{\mathbf{f}}_{t,j} \right\|_2 \left\| \widetilde{\mathbf{f}}_{r,i} \right\|_2}. \tag{7}$$

We retrieve the k reference patches with the smallest distances:

$$\mathcal{N}_k(j) = \mathrm{BottomK}_{i \in \{1,\dots,P\}} \left(d_{ji}\right), \tag{8}$$

where $\mathcal{N}_k(j)$ contains the indices of the k nearest reference patches in total across all classes.

The retrieved neighbors may carry different semantic labels. We combine their evidence through inverse-distance voting, so that closer reference patches contribute more strongly to the prediction. When all retrieved distances are positive, the score for class $c \in \mathcal{C}$ is

$$S_{j,c} = \sum_{\substack{i \in \mathcal{N}_k(j) \\ y_{r,i}=c}} \frac{1}{d_{ji}}. \tag{9}$$

If any retrieved neighbors have zero distance, only those neighbors contribute to the class scores, each with a unit weight. Accumulating these contributions by class allows foreground classes and background to compete for each target patch. We assign the patch to the class with the highest score:

$$\widehat{y}_{t,j} = \arg\max_{c \in \mathcal{C}} S_{j,c}. \tag{10}$$

After classifying all target patches, we reshape their predicted labels into an $h \times w$ mask and upsample it to the original target image resolution $H \times W$ using nearest-neighbor interpolation. The resulting prediction $\widehat{\mathbf{Y}}_t$ assigns one semantic label to each pixel without post-hoc conflict resolution.

4 Experiments

4.1 Experimental Setup

Datasets. We evaluate DINO-KNN on one-shot multiclass semantic segmentation using seven datasets spanning natural images, aerial imagery, underwater scenes, and medical ultrasound images. For natural image segmentation, we use COCO-20^i Lin et al. (2014), containing 80 object categories, and LVIS-92^i Gupta et al. (2019), containing 920 categories. We additionally evaluate on iSAID-5^i Waqas Zamir et al. (2019), an aerial imagery dataset with 15 categories, and SUIM Islam et al. (2020), an underwater dataset with eight categories.

Ultrasound datasets; one of the most challenging medical imaging data due to operator dependence, different artifacts, and low quality; span distinct anatomical and clinical domains, with differences in the structures depicted and acquisition settings Meyer et al. (2026). To evaluate joint multiclass segmentation across these domains, we select three 2D ultrasound datasets with annotations for multiple anatomical structures within each image: CAMUS Leclerc et al. (2019) for adult cardiac structures in echocardiography, FASS Correggio et al. (2023) for fetal abdominal structures, and PSFHS Chen et al. (2024) for the pubic symphysis and fetal head in intrapartum pelvic ultrasound.

Reference–target construction. Each evaluation episode consists of one annotated reference image and one target image, with at least two foreground classes shared between them. For COCO-20^i, we use $1{,}000$ evaluation episodes, following the episode count in the official INSID3 implementation Cuttano et al. (2026). For LVIS-92^i, iSAID-5^i, and SUIM, we evaluate every eligible target image using one fixed valid reference for that target. For each ultrasound dataset, we select one annotated reference image containing the evaluated anatomical classes and use every remaining image as a target. For each evaluation episode, the foreground label set consists of the semantic classes shared by the reference and target images. All other categories, including those present in only one of the two images, are treated as background. All evaluated methods use exactly the same reference images, target images, and semantic classes.

Implementation details. We use the frozen DINOv3 ViT-L/16 encoder and extract the spatial patch tokens from its final transformer block. Each image is resized to 1024×1024, producing a 64×64 patch-feature grid with a patch size of 16×16. We set the positional-subspace dimension to $s = 500$. A single Gaussian noise image is generated using a fixed random seed, and the resulting positional basis is computed once and reused for all evaluation episodes. All reference patches, including those labeled as background, are retained in the class-wise reference feature bank. We set $k = 15$ based on the neighborhood-size analysis in Sec. 4.3.

Evaluation metrics. For each evaluation episode, we compute mean intersection over union (mIoU) as the mean IoU over foreground classes. Binary class masks are evaluated independently, without resolving overlaps. For each episode, the cross-class conflict rate (CCR) measures the percentage of predicted foreground pixels assigned to more than one class:

$$\mathrm{CCR} = 100 \times \frac{N_{\mathrm{overlap}}}{N_{\mathrm{foreground}}}, \tag{11}$$

where N_{overlap} counts pixels predicted as foreground by two or more class masks, and $N_{\mathrm{foreground}}$ counts pixels predicted as foreground by at least one class mask. If $N_{\mathrm{foreground}} = 0$, we set $\mathrm{CCR} = 0$. We report both mIoU and CCR averaged across evaluation episodes.

Table 1: **Comparison of training-free segmentation methods.** We report mIoU (%, ↑) and CCR (%, ↓) across seven datasets. DINO-KNN performs multi-class inference using frozen DINOv3 features. The best mIoU is shown in **bold**, and the second-best is underlined.

Method	Encoder	#Params	COCO		LVIS		SUIM		iSAID		CAMUS		FASS		PSFHS		Avg.	
			mIoU	CCR	mIoU	CCR	mIoU	CCR	mIoU	CCR	mIoU	CCR	mIoU	CCR	mIoU	CCR	mIoU	CCR
Mask-supervised pre-training + binary prediction																		
PerSAM	SAM	640M	8	28	4	24	20	45	12	17	14	100	3	92	13	99	11	58
Matcher	DINOv2 + SAM	945M	24	20	19	16	42	29	28	43	16	95	4	91	14	61	21	51
GF-SAM	DINOv2 + SAM	945M	37	8	33	6	56	7	41	32	15	96	4	82	14	44	29	39
Unsupervised pre-training + binary inference																		
INSID3 (Base)	DINOv3-Base	86M	34	8	26	8	53	12	27	42	14	100	3	97	12	100	24	52
INSID3	DINOv3-Large	303M	**42**	2	**37**	4	53	4	46	6	14	100	3	83	14	96	30	42
DINO-KNN-Binary	DINOv3-Large	303M	36	0	29	0	**57**	1	**48**	1	39	3	13	0	**31**	0	**36**	1
Ours: *Unsupervised pre-training + multi-class inference*																		
DINO-KNN (Base)	DINOv3-Base	86M	34	0	27	0	55	0	42	0	**41**	0	**16**	0	26	0	34	0
DINO-KNN	DINOv3-Large	303M	36	0	29	0	**57**	0	**48**	0	40	0	14	0	**31**	0	**36**	0

4.2 Main Results

Baselines. Table 1 compares DINO-KNN with the state-of-the-art training-free ICS methods PerSAM Zhang et al. (2024b), Matcher Liu et al. (2024b), GF-SAM Zhang et al. (2024a), and INSID3 Cuttano et al. (2026). Because these baselines predict a binary mask for one foreground concept at a time, we run each method separately for every foreground class present in the reference annotation. An episode containing C foreground classes, therefore, requires C inference runs.

Why not resolve conflicts post hoc? Binary ICS methods can enforce mutually exclusive predictions by applying a post-hoc rule such as selecting the class with the highest foreground confidence. However, this does not change the fact that the model performs C independent binary segmentation problems and only converts their outputs into a multiclass prediction afterwards. DINO-KNN instead formulates the task as multiclass segmentation from the outset: all classes compete within the same k-NN decision, producing one semantic label per patch in a single inference. We therefore view post-hoc conflict resolution as a valid adaptation of binary ICS, but not as equivalent to direct joint multiclass inference.

We additionally evaluate **DINO-KNN-Binary**, a controlled variant that runs DINO-KNN separately for each foreground class, following the same concept-wise protocol as the existing methods. Because DINO-KNN-Binary and DINO-KNN use the same framework; their comparison isolates the effect of joint multiclass prediction on cross-class conflicts.

To control for encoder capacity, we also report DINOv3 ViT-B/16 variants of DINO-KNN and INSID3. We denote these as DINO-KNN (Base) and INSID3 (Base), while the unqualified method names use DINOv3 ViT-L/16.

Segmentation accuracy. DINO-KNN achieves the highest average mIoU among the evaluated methods, reaching 36 across seven datasets, compared with 30 for INSID3. It also outperforms GF-SAM by 7 mIoU points on average. DINO-KNN exceeds INSID3 on five datasets, with the largest gains on ultrasound: 26, 11, and 17 mIoU points on CAMUS, FASS, and PSFHS, respectively. It also improves performance by 4 and 2 mIoU points on SUIM and iSAID, respectively. INSID3 performs better on COCO and LVIS by 6 and 8 mIoU points, respectively, indicating that the advantage varies across domains. These results support direct k-NN classification as an effective alternative to INSID3's multi-stage inference pipeline, with particularly strong performance on the evaluated datasets outside the natural image domain.

Cross-class conflicts. DINO-KNN achieves a CCR of 0 through joint multiclass prediction, which assigns exactly one semantic label to each target patch. DINO-KNN-Binary, which predicts each foreground class independently, achieves a mean CCR of 1, the lowest among the evaluated binary methods, while maintaining mIoU comparable to DINO-KNN. This result shows that cross-class conflicts are already rare with the k-NN framework under independent binary inference. Joint multiclass prediction eliminates the remaining conflicts and guarantees mutually exclusive predictions without mask merging or post-hoc conflict resolution. In comparison, INSID3, the strongest prior

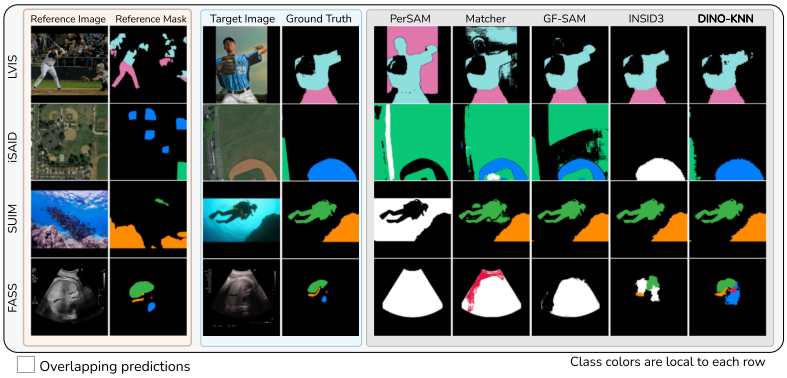


Figure 3: **Qualitative comparison of multiclass in-context segmentation.** Binary baselines are applied independently to each foreground class, and their masks are combined for visualization. White regions indicate pixels assigned to multiple foreground classes. DINO-KNN predicts all represented classes jointly and produces mutually exclusive labels.

baseline by average mIoU, has a mean CCR of 42, driven primarily by its high conflict rates on the ultrasound datasets. Together, these results show that DINO-KNN combines strong segmentation accuracy with overlap-free multiclass predictions, avoiding the cross-class conflicts observed in prior binary ICS methods. Figure 3 presents qualitative comparisons across four datasets, illustrating differences in segmentation quality and highlighting cross-class conflicts where they occur.

Backbone size. Replacing DINOv3 ViT-L/16 with ViT-B/16 reduces average mIoU for both INSID3 and DINO-KNN, with a larger decrease for INSID3. INSID3 exhibits its largest accuracy drops on iSAID, LVIS, and COCO, decreasing by 19, 11, and 8 mIoU points, respectively. DINO-KNN achieves lower mIoU on five datasets but improves on two ultrasound datasets, increasing from 40 to 41 on CAMUS and from 14 to 16 on FASS. INSID3's mean CCR increases from 42 to 52, while DINO-KNN maintains zero CCR with both backbones by construction. The smaller decrease in average mIoU indicates that DINO-KNN's segmentation accuracy is less sensitive to this reduction in backbone size.

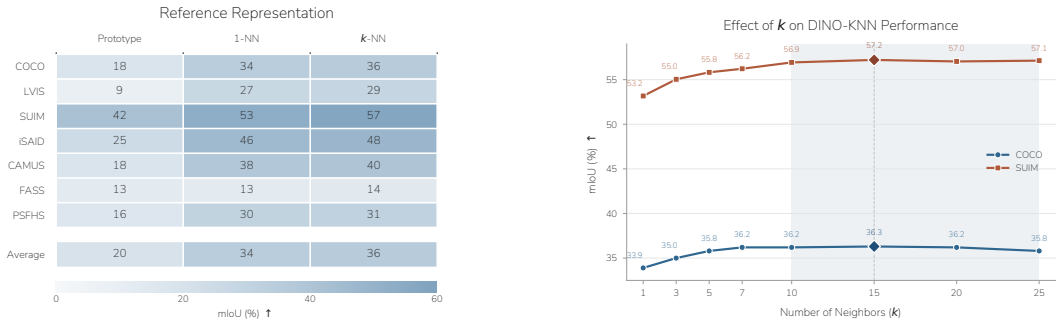
4.3 Ablation Studies

We examine the choice of reference representation and the sensitivity of DINO-KNN to the number of retrieved neighbors.

Reference representation. We compare DINO-KNN with two simpler alternatives. DINO-Prototype represents each class by means of its reference features, while DINO-1NN transfers the label of the closest reference patch. DINO-KNN instead uses the k nearest reference patches for classification. As shown in Fig. 4a, DINO-Prototype achieves only 20 average mIoU, compared with 34 for DINO-1NN and 36 for DINO-KNN. Moreover, DINO-KNN outperforms DINO-1NN across all seven datasets by 1 to 4 mIoU points. These results indicate that preserving individual reference features captures intra-class variation more effectively than a single class prototype, while combining multiple neighbors improves segmentation accuracy over relying on a single correspondence.

Effect of neighborhood size. We analyze the effect of k on COCO and SUIM. As shown in Fig. 4b, performance improves as k increases from 1 and then stabilizes, reaching the highest mIoU at $k = 15$ on both COCO (36.3) and SUIM (57.2). This consistent plateau indicates that DINO-KNN benefits from aggregating multiple neighbors while remaining relatively insensitive to the precise neighborhood size. We therefore use $k = 15$ for all experiments.

Computational cost. We evaluate the runtime and GPU memory usage on 30 CAMUS images using a single NVIDIA RTX A5000. As shown in Table 2, binary baselines perform a separate



(a) Comparison of feature-matching strategies.

(b) Effect of the number of neighbors k.

Figure 4: **Ablation studies of DINO-KNN. (a)** Comparison of prototype-based, 1-nearest-neighbor, and kNN inference. DINO-KNN consistently outperforms the alternative reference representation strategies across all evaluated datasets. All values are reported as mIoU (%). **(b)** mIoU on COCO & SUIM datasets as the number of neighbors k varies. Performance improves from $k = 1$ to $k = 7$, remains stable for $k \in [7, 20]$, and reaches its maximum at $k = 15$.

Table 2: **Computational cost.** Results are measured on 30 CAMUS images using a single NVIDIA RTX A5000. DINO-KNN predicts all requested classes jointly, whereas binary baselines require a separate inference call for each class. Runtime is reported as the mean in seconds. GPU memory is measured for the three-class setting. The lowest mean runtime for each class count and the lowest GPU memory usage are highlighted in **bold**.

Method	Input resolution	Runtime (s) ↓			GPU memory (MB) ↓
		1 class	2 classes	3 classes	Peak
PerSAM	1024×1024	1.160	2.394	3.661	5,743
Matcher	518×518	3.411	6.822	9.267	6,934
GF-SAM	1024×1024	**1.018**	**2.071**	3.122	6,929
INSID3	1024×1024	1.066	2.127	3.188	1,573
DINO-KNN	1024×1024	–	–	**1.129**	**1,358**

inference call for each foreground class, whereas DINO-KNN predicts all three foreground classes jointly. For three-class prediction, DINO-KNN achieves the lowest runtime at 1.129 seconds, compared with 3.188 seconds for INSID3 and 3.122 seconds for GF-SAM, providing approximately $2.8\times$ faster inference than both methods under the evaluated protocol. DINO-KNN also achieves the lowest peak GPU memory usage at 1,358 MB, approximately 13.7% below INSID3 and 76–80% below the SAM-based methods. These results demonstrate that DINO-KNN has faster prediction and lower peak GPU memory usage through joint multiclass inference.

5 Conclusion

We introduced **DINO-KNN**, a simple training-free and decoder-free approach to one-shot multi-class in-context segmentation using frozen DINOv3 patch features and k-NN classification. By jointly predicting foreground and background classes, DINO-KNN produces mutually exclusive masks without binary per-class inference or post hoc conflict resolution. Across seven datasets spanning natural images, aerial imagery, underwater scenes, and ultrasound, DINO-KNN achieves the highest average mIoU among the evaluated methods, outperforming INSID3 by 6 mIOU points. Our ablations demonstrate the benefits of retaining individual reference features rather than using a single class prototype and combining multiple neighbors for classification. These findings establish direct classification of frozen patch features as a strong baseline for training-free multiclass ICS. The patch-level prediction scheme can miss small structures and fine boundaries, motivating future work to improve spatial detail while preserving the approach's simplicity.

THE USE OF LARGE LANGUAGE MODELS. In this work, we used generative AI tools to assist in coding, and enhance writing. After using these tools, the authors review and edit the content as needed and take full responsibility for the publication's content.

REFERENCES

Ivana Balazevic, David Steiner, Nikhil Parthasarathy, Relja Arandjelović, and Olivier Henaff. Towards in-context scene understanding. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 63758–63778. Curran Associates, Inc., 2023. doi: 10.52202/075280-2785. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/c94a632545000531f0b47000e9caa5b6-Paper-Conference.pdf.

Amir Bar, Yossi Gandelsman, Trevor Darrell, Amir Globerson, and Alexei Efros. Visual prompting via image inpainting. *Advances in neural information processing systems*, 35:25005–25017, 2022.

Gaowen Chen, Jieyun Bai, Zhanhong Ou, Yaosheng Lu, and Huijin Wang. PSFHS: Intrapartum ultrasound image dataset for AI-based segmentation of pubic symphysis and fetal head. *Scientific Data*, 11:436, 2024. doi: 10.1038/s41597-024-03266-4.

Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1280–1289. IEEE, 2022.

Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3213–3223, 2016.

K. S. D. Correggio, R. N. Galluzzo, L. O. Santos, F. S. M. Barroso, T. Z. L. Chaves, A. S. C. Onofre, and A. von Wangenheim. Fetal abdominal structures segmentation dataset using ultrasonic images, October 2023. URL https://data.mendeley.com/datasets/4gcpm9dsc3/1.

Claudia Cuttano, Gabriele Trivigno, Christoph Reich, Daniel Cremers, Carlo Masone, and Stefan Roth. INSID3: Training-free in-context segmentation with DINOv3. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21638–21648, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Cuttano_INSID3_Training-Free_In-Context_Segmentation_with_DINOv3_CVPR_2026_paper.html.

Agrim Gupta, Piotr Dollar, and Ross Girshick. LVIS: A dataset for large vocabulary instance segmentation. pp. 5356–5364, 2019.

Md Jahidul Islam, Chelsey Edge, Yuyang Xiao, Peigen Luo, Muntaqim Mehtaz, Christopher Morse, Sadman Sakib Enan, and Junaed Sattar. Semantic segmentation of underwater imagery: Dataset and benchmark. In *IROS*, pp. 1769–1776, 2020.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollar, and Ross Girshick. Segment Anything. pp. 4015–4026, 2023.

Sarah Leclerc, Erik Smistad, Joao Pedrosa, Andreas Ostvik, Frederic Cervenansky, Florian Espinosa, Torvald Espeland, Erik Andreas Rye Berg, Pierre-Marc Jodoin, Thomas Grenier, Carole Lartizien, Jan Dhooge, Lasse Lovstakken, and Olivier Bernard. Deep learning for segmentation using an open large-scale dataset in 2d echocardiography. *IEEE Transactions on Medical Imaging*, 38(9):2198–2210, 2019. ISSN 1558-254X. doi: 10.1109/tmi.2019.2900516. URL http://dx.doi.org/10.1109/TMI.2019.2900516.

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

Yang Liu, Chenchen Jing, Hengtao Li, Muzhi Zhu, Hao Chen, Xinlong Wang, and Chunhua Shen. A simple image segmentation framework via in-context examples. *Advances in Neural Information Processing Systems*, 37:25095–25119, 2024a.

Yang Liu, Muzhi Zhu, Hengtao Li, Hao Chen, Xinlong Wang, and Chunhua Shen. Matcher: Segment anything with one shot using all-purpose feature matching. In *The Twelfth International Conference on Learning Representations*, 2024b. URL https://openreview.net/forum?id=yzRXdhk2he.

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

Lingchen Meng, Shiyi Lan, Hengduo Li, Jose M Alvarez, Zuxuan Wu, and Yu-Gang Jiang. Segic: Unleashing the emergent correspondence for in-context segmentation. In *European Conference on Computer Vision*, pp. 203–220. Springer, 2024a.

Lingchen Meng, Shiyi Lan, Hengduo Li, Jose M. Alvarez, Zuxuan Wu, and Yu-Gang Jiang. SegIC: Unleashing the emergent correspondence for in-context segmentation. volume 38, pp. 203–220, 2024b.

Adrien Meyer, Aditya Murali, Farahdiba Zarin, Didier Mutter, and Nicolas Padoy. Ultrasam: a foundation model for ultrasound using large open-access segmentation datasets. *International Journal of Computer Assisted Radiology and Surgery*, 21(1):93–102, 2026.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. DINOv2: Learning robust visual features without supervision. 2024.

Junghwan Park. Frost: Training-free few-shot segmentation with frozen features and nonparametric statistics. *arXiv preprint arXiv:2606.31136*, 2026.

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

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. DINOv3. *arXiv:2508.10104 [cs.CV]*, 2025.

Kaixin Wang, Jun Hao Liew, Yingtian Zou, Daquan Zhou, and Jiashi Feng. Panet: Few-shot image semantic segmentation with prototype alignment. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9196–9205. IEEE, 2019.

Xinlong Wang, Wen Wang, Yue Cao, Chunhua Shen, and Tiejun Huang. Images speak in images: A generalist painter for in-context visual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6830–6839, June 2023a.

Xinlong Wang, Xiaosong Zhang, Yue Cao, Wen Wang, Chunhua Shen, and Tiejun Huang. SegGPT: Towards segmenting everything in context. pp. 1130–1140, 2023b.

Syed Waqas Zamir, Aditya Arora, Akshita Gupta, Salman Khan, Guolei Sun, Fahad Shahbaz Khan, Fan Zhu, Ling Shao, Gui-Song Xia, and Xiang Bai. isaid: A large-scale dataset for instance segmentation in aerial images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2019.

Anqi Zhang, Guangyu Gao, Jianbo Jiao, Chi Harold Liu, and Yunchao Wei. Bridge the points: Graph-based few-shot segment anything semantically. pp. 33232–33261, 2024a.

Renrui Zhang, Zhengkai Jiang, Ziyu Guo, Shilin Yan, Junting Pan, Xianzheng Ma, Hao Dong, Peng Gao, and Hongsheng Li. Personalize segment anything model with one shot. 2024b.