CLUSTER-CONSENSUS BAYES FOR OPEN-VOCABULARY CAMOUFLAGED OBJECT SEGMENTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Open-vocabulary camouflaged object segmentation requires recognizing visually subtle objects from unseen categories. Existing methods typically process each test image independently, overlooking the latent consensus among related samples. We introduce Cluster-Consensus Bayes (CCB), a lightweight, plug-and-play empirical-Bayes module that turns this consensus into a source of semantic evidence. CCB groups unlabeled test images by visual similarity and estimates cluster-level class priors from their predicted categories. It then adaptively fuses these priors with individual vision-language predictions according to prediction confidence. This allows uncertain samples to draw on shared evidence while confident samples retain their individual information, without assigning a single label to an entire cluster. CCB supports iterative offline refinement across complementary encoder views and online inference through confidence-weighted centroid memory. Both modes refine predictions at test time without gradient updates, accommodating collection-based and streaming inference. Experiments across five base models and two additional single-target datasets demonstrate consistent recognition gains. On OVCamo, CLIP-only CCB improves OVCoser's accuracy from 68.8% to 73.2%, with a 6.2% average relative gain across six class-aware segmentation metrics. An enhanced system combining auxiliary DINO feature fusion with multi-view CCB achieves a 14.3% average relative gain over original OVCoser.

1 INTRODUCTION

Open-vocabulary camouflaged object segmentation (OV-COS) asks a model to locate visually concealed objects and recognize categories absent from its segmentation training set (Pang et al., 2024). Vision-language models provide a semantic interface to these unseen categories, but camouflage makes that interface difficult to use: the object may have few distinctive pixels, and several candidate names can plausibly match the same image. A model can therefore recover a reasonable foreground mask while assigning it the wrong category. Improving semantic inference is central to turning a detected camouflaged region into an object with a reliable semantic identity.

Most OV-COS systems construct their prediction from the current image and candidate text descriptions (Pang et al., 2024; Zhao et al., 2026). This leaves an additional source of evidence unused. Images from the same category often exhibit recurring appearance, structure, and environmental context. For camouflaged objects, such relationships are particularly relevant: the surroundings that obscure a target locally can also contribute to a recognizable pattern across related images. Related unlabeled images can thus provide semantic context for interpreting an ambiguous target.

The central question is how to convert these relationships into useful semantic evidence. Visual similarity alone does not identify a category, and assigning one label to an entire cluster can erase meaningful differences between its members. We instead represent a cluster by the distribution of its members' predicted categories. This *cluster consensus* captures both the dominant interpretation and competing alternatives. Its influence follows sample confidence: confident predictions preserve their individual evidence, while ambiguous ones draw greater support from the neighborhood.

We introduce **Cluster-Consensus Bayes (CCB)**, an empirical-Bayes approach to open-vocabulary inference that estimates semantic priors from the test population itself. CCB combines these priors

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

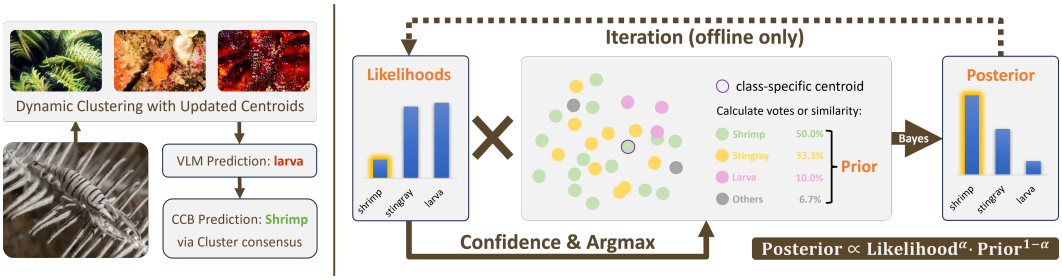

Figure 1: **From individual predictions to cluster consensus.** CCB groups visually related test images, estimates a soft class prior from their predictions, and fuses this prior with individual evidence. Offline inference repeatedly updates this consensus; streaming inference accumulates it in a compact centroid memory that supplies semantic context for every incoming prediction.

with vision-language predictions through a confidence-weighted geometric mean. The update also admits a simple variational interpretation as agreement between individual and cluster-level distributions. The update operates in label space and attaches directly to an existing model.

CCB supports two ways of organizing test-time evidence. In the **offline mode**, it alternates visual clustering and semantic refinement over an available image collection. The updated predictions supply the consensus for the next round. Importantly, different frozen encoders can define different neighborhoods for the same samples. Switching these views allows the refinement to draw on complementary relationships that repeated clustering in one feature space may miss. In the **online mode**, Prediction-Confidence Weighted Online Clustering (PC-WOC) accumulates semantic evidence in a compact centroid memory. Each incoming image can immediately use the observed context, enabling immediate inference as each new sample arrives in the stream.

Experiments show consistent gains across OVSS models and dedicated OV-COS systems. With CLIP embeddings alone, CCB improves OVCoser from 68.8% to 73.2% accuracy; its online counterpart achieves 72.4%. An enhanced system combining auxiliary feature fusion with multi-view CCB reaches 75.9%. Beyond these gains, the experiments reveal how consensus operates: geometric fusion outperforms direct prior replacement, different embedding views produce complementary refinement, and semantic correction improves class-aware segmentation while preserving the underlying masks. Evaluations on single-target VOC and COCO subsets further extend the evidence beyond camouflage.

This formulation leads to three contributions to open-vocabulary inference:

- We formulate cluster consensus as a test-time semantic prior and introduce confidence-adaptive fusion for open-vocabulary camouflaged object segmentation.
- We develop iterative multi-view refinement and confidence-weighted streaming memory as two implementations of the same consensus-based inference principle.
- We establish the effectiveness of CCB across models and datasets, with analyses of evidence fusion, complementary neighborhood views, and the available inference context.

2 Related Work

Open-vocabulary segmentation and camouflage. Vision-language pretraining (Radford et al., 2021) connects visual representations to category names beyond a segmentation model's training vocabulary. OpenSeg and GroupViT explore language-aligned visual representations (Ghiasi et al., 2022; Xu et al., 2022), while OVSeg, SAN, CAT-Seg, and SCAN develop mask adaptation, side adapters, cost aggregation, and semantic-context calibration (Liang et al., 2023; Xu et al., 2023; Cho et al., 2024; Liu et al., 2024). Camouflaged object segmentation focuses on recovering weak foreground structure (Fan et al., 2020; Mei et al., 2021; Yin et al., 2024). OVCoser brings open-vocabulary recognition to this setting (Pang et al., 2024), and cascaded VLMs combine segmentation cues with context-preserving recognition (Zhao et al., 2026). CCB operates on the semantic predictions of these systems and introduces relationships between test images into their inference.

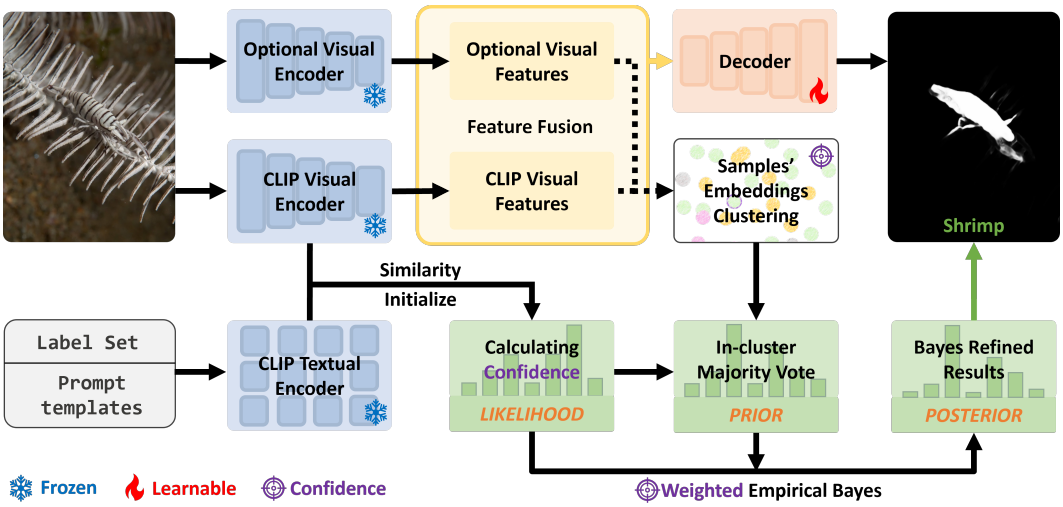

Figure 2: **CCB integrated with a segmentation model.** Frozen visual embeddings organize images into neighborhoods, whose predicted categories provide a consensus prior. Confidence controls its fusion with individual predictions. Auxiliary features can also be fused into the mask decoder.

Test-time adaptation and memory. Test-time adaptation exploits unlabeled target observations through entropy minimization, augmentation consistency, or continual parameter updates (Wang et al., 2021; Zhang et al., 2022; Wang et al., 2022). Related approaches adapt source hypotheses (Liang et al., 2020) and vision-language models (Maharana et al., 2025; Noori et al., 2025). Gradient-free methods also use target observations: T3A adjusts a classifier with test-time prototypes (Iwasawa & Matsuo, 2021), TDA maintains a dynamic cache of test features and pseudo-labels (Karmanov et al., 2024), and DMN combines dynamic and static memories (Zhang et al., 2024). PC-WOC summarizes confident observations into centroids and converts their similarities into a semantic prior. This memory supplies neighborhood evidence for the same fusion rule used in the offline setting.

Transductive inference. Transductive learning and label propagation exploit the geometry of an unlabeled population (Joachims et al., 1999; Zhu et al., 2003). LAME refines output distributions through Laplacian regularization while keeping the model fixed (Boudiaf et al., 2022). CCB develops a cluster-conditioned construction: visual neighborhoods determine which predictions share evidence, their label frequencies define a semantic prior, and sample confidence controls the update. Switching encoder views reorganizes these neighborhoods and introduces new relationships into refinement.

3 Cluster-Consensus Bayes

3.1 Semantic Evidence from Visual Neighborhoods

Let $\mathcal{X} = \{x_i\}_{i=1}^{N}$ be an inference collection and $\mathcal{C}$ a candidate vocabulary with K category names. A base model produces a distribution $p_i \in \mathbb{R}^K$ and a target mask for each image. CCB refines the category distribution using normalized visual embeddings z_i from a frozen encoder. Figures 2 and 3 illustrate how CCB connects visual grouping with semantic inference. We denote the number of clusters by N_c and obtain the candidate vocabulary from the base model's text interface.

Suppose clustering assigns image i to group G_{a_i}. The top prediction $\hat{y}_i = \arg\max_y p_i(y)$ provides one semantic vote. Aggregating votes gives a cluster-conditioned distribution

$$r_j(y) = \frac{n_j(y) + \epsilon}{|G_j| + K\epsilon}, \qquad n_j(y) = \sum_{i \in G_j} \mathbb{I}[\hat{y}_i = y], \tag{1}$$

where $\epsilon > 0$ smooths the class counts. Top-1 voting concentrates the contribution of each image on its leading interpretation, while the resulting prior remains a distribution over the entire vocabulary.

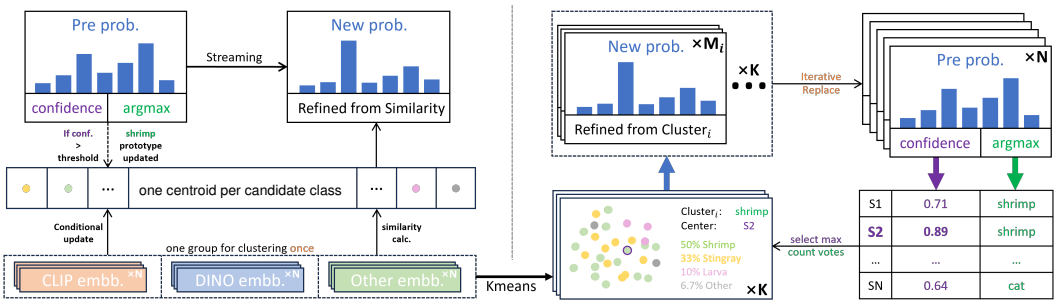

Figure 3: **Inside the CCB module.** Left: streaming PC-WOC updates predicted-class centroids with confident samples and obtains a prior from active-centroid similarities. Right: offline CCB selects confidence-guided centers, groups visual embeddings, and estimates a soft prior from category votes. Refined distributions supply the evidence for the next consensus round.

A cluster can therefore express a dominant category without requiring every member to share that label. Smoothing gives the posterior mean under a symmetric Dirichlet model for the class counts.

3.2 Confidence-Adaptive Consensus Fusion

CCB estimates a prior from the current test population and combines it with the base prediction in an empirical-Bayes formulation. After probability flooring and renormalization, let $\bar{p}_i$ denote the stabilized prediction. Confidence-adaptive fusion gives

$$\begin{aligned} q_i(y) &= \frac{\bar{p}_i(y)^{\alpha_i}\, r_{a_i}(y)^{1-\alpha_i}}{\sum_{c\in\mathcal{C}} \bar{p}_i(c)^{\alpha_i}\, r_{a_i}(c)^{1-\alpha_i}} \\ &= \mathrm{softmax}_y[\alpha_i \log \bar{p}_i + (1-\alpha_i)\log r_{a_i}]\,. \end{aligned} \tag{2}$$

Here α_i determines how strongly the sample retains its individual evidence. We combine normalized entropy and the margin between the largest two probabilities:

$$\alpha_i = \lambda\left(1 + \frac{\sum_y p_i(y)\log p_i(y)}{\log K}\right) + (1-\lambda)\left[p_{i,(1)} - p_{i,(2)}\right]. \tag{3}$$

Both terms lie in $[0,1]$ for $K > 1$. Entropy captures dispersion across the vocabulary, while the margin measures competition between the leading candidates. Consequently, a confident image contributes more of its own prediction and an ambiguous image receives more cluster guidance.

Equivalently, fusion minimizes a confidence-weighted sum of KL divergences:

$$q_i = \arg\min_{q\in\Delta^{K-1}} \alpha_i D_{\mathrm{KL}}(q\|\bar{p}_i) + (1-\alpha_i) D_{\mathrm{KL}}(q\|r_{a_i}). \tag{4}$$

It balances agreement with the individual prediction and the neighborhood distribution. Consensus therefore acts as an adaptive semantic regularizer, guiding each image while preserving differences between cluster members. The geometric update combines support shared by the two distributions and lets sample confidence determine their relative influence. Appendix A provides the derivation.

The update expresses consensus through the odds between two candidate categories a and b:

$$\log\frac{q_i(a)}{q_i(b)} = \alpha_i \log\frac{\bar{p}_i(a)}{\bar{p}_i(b)} + (1-\alpha_i)\log\frac{r_{a_i}(a)}{r_{a_i}(b)}. \tag{5}$$

A neighborhood favoring a contributes evidence against a competing category b. Its influence grows as individual confidence falls, so the same consensus can resolve an ambiguous image while preserving a confident neighbor's category. Visual grouping determines the source of evidence, and confidence determines how that evidence changes each prediction.

3.3 Iterative Multi-View Refinement

Prediction-guided grouping. Offline CCB runs K-means over an available inference collection. By default, $N_c = K$ sets the clustering resolution from the candidate vocabulary. For each candidate

class, we initialize a center with the most confident image currently predicted as that class. If no image has that top prediction, we select an unused image with the largest probability for the class. This uses lower-ranked predictions to retain initialization hypotheses that would otherwise be absent. K-means then groups images by visual distance and collects their category votes.

Alternating structure and semantics. At iteration r, the current predictions $p_i^{(r-1)}$ define the votes and confidence scores, and embedding view v_r determines the neighborhoods. CCB updates

$$\begin{aligned} \{G_j^{(r)}\} &= \mathrm{Cluster}(\{z_i^{(v_r)}\}, \{p_i^{(r-1)}\}), \\ p_i^{(r)} &= \mathcal{F}\big(p_i^{(r-1)}, r_{a_i}^{(r)}, \alpha_i^{(r)}\big), \end{aligned} \tag{6}$$

where $\mathcal{F}$ denotes Eq. (2). Refined semantics change both center selection and consensus in the next round. Our default schedule applies this update for three successive CLIP rounds.

Complementary neighborhood views. The same images can have different neighborhoods in CLIP, DINO, and ResNet feature spaces. Alternating these views lets an image receive evidence from different visual relationships across rounds. The prediction state is shared across views, so a semantic correction found in one space carries into the next. This is useful when repeating a single grouping reinforces a dominant class and progressively suppresses competing predictions. Our experiments compare repeated and mixed views using the same underlying segmentation model.

Frozen auxiliary embeddings can be used directly for clustering. We also evaluate enhanced base systems that fuse auxiliary spatial features into a trained mask decoder. At each decoder scale, a 3×3 convolution projects the auxiliary features, which are concatenated with the corresponding CLIP features and compressed by a 1×1 convolution. The encoders remain frozen during fusion-decoder training. CCB applies the same inference rule to the resulting predictions; Appendix D.1 describes the complete multi-scale architecture and the configurations of all five auxiliary encoders.

3.4 Streaming Consensus Memory

PC-WOC maintains a centroid v_y and accumulated confidence mass M_y for each active candidate class. For image x_t, the frozen model yields a normalized embedding z_t, class prediction $\hat{y}_t$, and confidence α_t. A prediction above threshold τ contributes weight $w_t = \alpha_t^{\beta}$ for updating the memory:

$$v_{\hat{y}_t} \leftarrow \mathrm{normalize}(M_{\hat{y}_t} v_{\hat{y}_t} + w_t z_t), \qquad M_{\hat{y}_t} \leftarrow M_{\hat{y}_t} + w_t. \tag{7}$$

A newly active class is initialized with $v_{\hat{y}_t} = z_t$ and $M_{\hat{y}_t} = w_t$. Thresholding selects reliable contributions, and accumulated mass stabilizes repeatedly supported directions.

The active centroids $\mathcal{A}_t$ yield a similarity-based prior,

$$r_t(y) = \frac{\exp(v_y^\top z_t / \kappa)}{\sum_{c \in \mathcal{A}_t} \exp(v_c^\top z_t / \kappa)}, \qquad y \in \mathcal{A}_t. \tag{8}$$

We scatter this distribution into the full vocabulary, stabilize it, and apply Eq. (2). Before any centroid is active, the prior equals the original prediction. Each output uses only the current image and accumulated observations. For embeddings of dimension d, the memory cost is $O(Kd)$ and each similarity computation costs $O(|\mathcal{A}_t|d)$, independently of the length of the observed stream.

4 Experiments

4.1 Experimental Setup

Datasets and metrics. Our main benchmark is OVCamo (Pang et al., 2024), with 7,713 training images from 14 categories and 3,770 test images from 61 disjoint categories. The test vocabulary contains these 61 category names. We also evaluate single-target subsets of Pascal VOC 2010 (Everingham et al., 2010) and COCO-Stuff (Caesar et al., 2018), containing 1,878 images from 20 categories and 4,452 images from 57 categories, respectively. These subsets retain one target mask per image. We report classification accuracy and six class-aware segmentation metrics: cS_m, cF_β^ω, cMAE, cF_β, cE_m, and cIoU. The latter three use averaged variants. Class-aware scores evaluate the predicted mask when its category is correct and assign zero otherwise, except cMAE, which assigns unit error. Class-agnostic metrics, denoted ag, evaluate mask geometry (Appendix C).

Table 1: **CCB across base models on OVCamo.** Each +CCB row uses the same base model with CLIP-only inference refinement. OVSS models use oracle mask selection. Δ averages relative improvement across six class-aware metrics, with accuracy reported separately.

Model	Acc (%) ↑	cS_m ↑	cF_β^ω ↑	cMAE ↓	cF_β ↑	cE_m ↑	cIoU ↑	$\mathbf{\Delta}(\%)$
OVSS methods' results on OVCamo								
SAN	66.6	0.386	0.318	0.518	0.331	0.426	0.293	0.0
+CCB	69.3	0.400	0.328	0.501	0.342	0.441	0.301	3.3
CAT-Seg	70.3	0.428	0.332	0.437	0.350	0.466	0.301	0.0
+CCB	73.3	0.444	0.342	0.414	0.363	0.485	0.309	3.7
SCAN	53.4	0.340	0.294	0.593	0.305	0.366	0.269	0.0
+CCB	57.4	0.363	0.312	0.564	0.324	0.391	0.284	6.1
OVCOS methods' results on OVCamo								
OVCoser	68.8	0.579	0.490	0.336	0.520	0.616	0.443	0.0
+CCB	73.2	0.611	0.513	0.294	0.545	0.651	0.462	6.2
CVLM	75.4	0.670	0.617	0.264	0.633	0.700	0.571	0.0
+CCB	76.7	0.681	0.626	0.251	0.642	0.713	0.579	2.1

Table 2: **Transfer to single-target VOC and COCO subsets.** The frozen OVCoser model is evaluated with and without CLIP-only CCB. All six class-aware metrics are shown, with the number of images and categories listed below each dataset name to describe the evaluation set.

Data	Method	Acc (%) ↑	cS_m ↑	cF_β^ω ↑	cMAE ↓	cF_β ↑	cE_m ↑	cIoU ↑	$\mathbf{\Delta}(\%)$
VOC-S	OVCoser	79.6	0.470	0.317	0.340	0.361	0.474	0.248	0.0
(1878/20)	+CCB	83.7	0.486	0.321	0.308	0.366	0.493	0.250	+3.4
COCO-S	OVCoser	27.6	0.131	0.076	0.799	0.092	0.132	0.056	0.0
(4452/57)	+CCB	33.0	0.154	0.086	0.760	0.103	0.155	0.063	+12.9

Models and implementation. We integrate CCB into SAN, CAT-Seg, SCAN, OVCoser, and a cascaded-VLM model (CVLM). OVSS models use their COCO-Stuff pretrained weights; their multiple outputs are evaluated with the highest-IoU mask, following the OV-COS oracle-mask protocol. OVCoser-based systems follow the original training setup. Main plug-in results use frozen CLIP embeddings, with three offline rounds by default. PC-WOC uses $\lambda = 0.5$, $\beta = 2$, $\tau = 0.85$, and $\kappa = 0.07$; online results are averaged over five input orders. Enhanced systems fuse DINOv3 or ResNet features into the decoder before applying CCB. CVLM uses two refinement rounds, with its configuration detailed in Appendix D alongside the other model settings.

4.2 Open-Vocabulary Segmentation Results

Table 1 shows that CCB improves every evaluated base model. OVCoser gains 4.4 accuracy points, from 68.8% to 73.2%, together with a 6.2% average relative improvement in class-aware segmentation. SAN, CAT-Seg, and SCAN gain 2.7, 3.0, and 4.0 accuracy points, respectively. The stronger CVLM also benefits, improving from 75.4% to 76.7%. These paired comparisons demonstrate that consensus can extract additional semantic information from the outputs of distinct segmentation architectures.

Table 2 extends the evaluation to two additional visual domains. CCB raises accuracy from 79.6% to 83.7% on VOC-S and from 27.6% to 33.0% on COCO-S. The gains occur without retraining OVCoser for these datasets. Their single-target protocol preserves the prediction unit used by CCB while changing both the visual domain and the candidate category vocabulary.

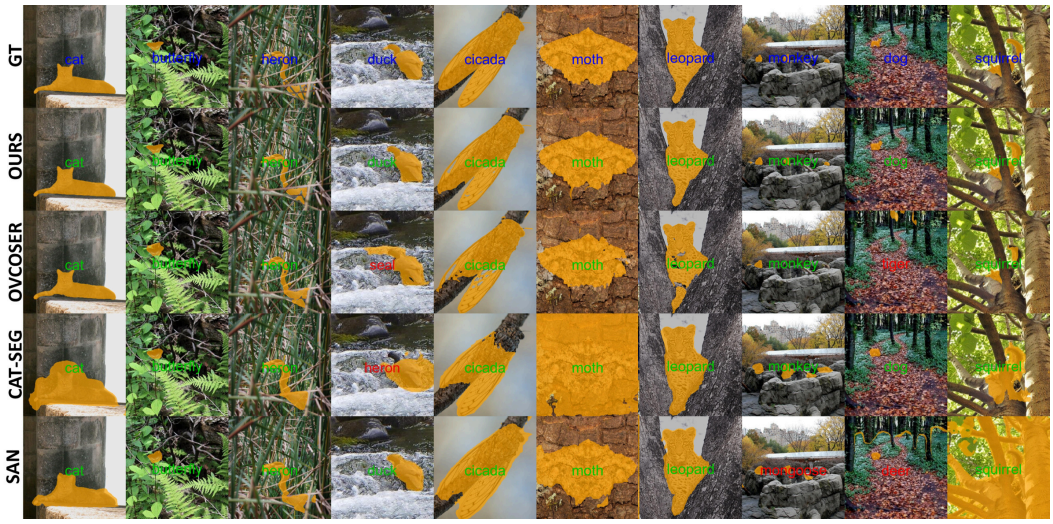

Figure 4: **Qualitative comparison on OVCamo.** Ten examples show ground truth, the CCB-enhanced system, OVCoser, CAT-Seg, and SAN from top to bottom. Green/red category names mark correct/incorrect recognition. CCB uses the auxiliary-feature decoder; Appendix G provides the complete twelve-scene comparison alongside additional qualitative examples. The duck and dog examples contrast correct recognition with plausible but mislabeled foreground masks; the monkey scene further shows a category error in SAN despite a reasonable predicted target region.

Table 3: **Refinement across embedding views.** C, D, and R denote CLIP, DINOv3, and ResNet clustering views. Fusion systems use a trained auxiliary-feature decoder; sequences within a block share the same base model. The complete configuration table appears in Appendix E.

Clustering sequence	Acc (%) ↑	cS_m ↑	cF_β^ω ↑	cMAE ↓	cF_β ↑	cE_m ↑	cIoU ↑	Δ(%)
CLIP features								
Baseline	68.8	0.579	0.490	0.336	0.520	0.616	0.443	0.0
+$\boldsymbol{C}$	71.8	0.601	0.505	0.306	0.536	0.640	0.455	4.2
+$\boldsymbol{C}, \boldsymbol{C}$	72.6	0.606	0.509	0.299	0.541	0.647	0.459	5.4
+$\boldsymbol{C}, \boldsymbol{C}, \boldsymbol{C}$	73.2	0.611	0.513	0.294	0.545	0.651	0.462	6.2
+$\boldsymbol{C}, \boldsymbol{C}, \boldsymbol{C}, \boldsymbol{C}$	73.1	0.609	0.511	0.295	0.543	0.650	0.461	5.9
DINO feature fusion								
Baseline	70.3	0.605	0.521	0.317	0.544	0.637	0.483	0.0
+$\boldsymbol{D}, \boldsymbol{D}, \boldsymbol{D}$	73.5	0.630	0.542	0.285	0.565	0.664	0.502	5.0
+$\boldsymbol{C}, \boldsymbol{C}, \boldsymbol{C}$	73.7	0.629	0.539	0.285	0.563	0.665	0.499	4.8
+$\boldsymbol{D}, \boldsymbol{D}, \boldsymbol{C}$	75.9	0.649	0.555	0.263	0.580	0.684	0.514	8.5
ResNet feature fusion								
Baseline	70.1	0.590	0.510	0.322	0.535	0.630	0.465	0.0
+$\boldsymbol{R}, \boldsymbol{C}, \boldsymbol{C}, \boldsymbol{C}$	75.4	0.626	0.533	0.272	0.561	0.671	0.486	7.0

4.3 Complementary Views for Iterative Consensus

Table 3 reveals two effects. First, semantic refinement accumulates over a small number of rounds: CLIP-only accuracy rises from 71.8% after one round to 73.2% after three. A fourth round provides no further benefit. Second, neighborhood diversity improves refinement beyond repeatedly using a strong representation. Within the same DINO-fusion system, DDD reaches 73.5%, while changing the final

Table 4: **Consensus construction and inference context.** CLIP-only OVCoser with alternative evidence rules, components, and inference contexts. Default CCB uses three rounds; prior-only accuracy is from round one. Online random-order results average five runs. Accuracy is in percent.

Setting	Acc (%) ↑	cS_m ↑	cF_β^ω ↑	cMAE ↓	cF_β ↑	cE_m ↑	cIoU ↑
Evidence fusion and component analysis							
No CCB	68.8	0.579	0.490	0.336	0.520	0.616	0.443
Prior only	62.6	0.421	0.353	0.515	0.373	0.450	0.322
Centroid correction	71.6	0.599	0.504	0.308	0.535	0.639	0.454
Average distributions	71.7	0.600	0.505	0.307	0.536	0.640	0.455
Arithmetic fusion	70.3	0.586	0.492	0.322	0.522	0.626	0.443
Fixed confidence	72.0	0.601	0.506	0.306	0.537	0.640	0.456
Random centers	71.5	0.596	0.500	0.310	0.530	0.637	0.451
CCB (default)	**73.2**	**0.611**	**0.513**	**0.294**	**0.545**	**0.651**	**0.462**
Offline inference context							
Batch 512	71.3	0.596	0.501	0.312	0.532	0.636	0.452
Batch 1024	72.5	0.604	0.506	0.300	0.537	0.645	0.456
$N_c = 30$	72.2	0.601	0.501	0.303	0.532	0.640	0.453
$N_c = 122$	72.1	0.604	0.508	0.304	0.539	0.642	0.457
Streaming input order							
Random (five-run mean)	72.4	0.605	0.506	0.301	0.537	0.643	0.457
Hard-to-easy	68.8	0.579	0.490	0.335	0.520	0.616	0.443
Class-grouped	73.6	0.615	0.514	0.288	0.546	0.654	0.464

view to CLIP yields 75.9%. Even the ResNet-fusion system reaches 75.4% with a mixed sequence. Neighborhood diversity thus complements representation strength during iterative refinement.

The three-round sweep in Appendix E.1 extends this observation to all five auxiliary encoders. Within each encoder block, the highest-accuracy mixed sequence exceeds both repeated-view alternatives. The preferred sequence varies across encoders, indicating that complementary views offer useful semantic evidence through different neighborhood structures across successive rounds of inference. DINOv1-ResNet-50 reaches 76.1% accuracy with RRC, while DINOv3 reaches 75.9% with DDC and obtains higher cIoU (.514 versus .462). These complementary strengths distinguish the quality of semantic neighborhoods from the spatial representation used by the decoder.

The enhanced DINO system achieves a 14.3% average relative gain across six class-aware mask metrics over original OVCoser, combining auxiliary feature fusion with semantic consensus. Relative to its 70.3% fusion baseline, CCB adds 5.6 accuracy points and improves cIoU from .483 to .514. Figure 4 shows the resulting predictions across diverse camouflaged scenes. Competing methods often localize plausible foreground regions but assign incorrect categories.

4.4 How Consensus Changes Predictions

Individual evidence and cluster evidence. Table 4 shows that replacing predictions with the cluster prior reduces first-round accuracy to 62.6%, below the unrefined model's 68.8%. A similarity-based centroid correction reaches 71.6%, while full CCB reaches 73.2%. The result favors preserving individual evidence while drawing on the group's class distribution. Geometric fusion outperforms arithmetic weighting (70.3%), and adaptive confidence improves on a fixed coefficient (72.0%). Appendix E.7 follows these updates across rounds, illustrating the role of adaptive fusion.

Two further comparisons examine how the prior is constructed. Averaging the full class distributions within each cluster yields 71.7%, compared with 73.2% for top-1 voting, supporting the use of each

Table 5: **Semantic and spatial effects of inference refinement.** CCB refines category probabilities; TENT-lite updates decoder normalization affine parameters by entropy minimization with frozen CLIP. The ag prefix denotes mask evaluation independent of the predicted category.

Method	Acc (%) ↑	cS_m ↑	cF_β^ω ↑	cIoU ↑	agS_m ↑	agF_β^ω ↑	agMAE ↓	agIoU ↑
OVCoser	68.8	0.579	0.490	0.443	0.809	0.647	0.036	0.583
+CCB	**73.2**	**0.611**	**0.513**	**0.462**	0.809	0.647	0.036	0.583
+TENT-lite	68.6	0.580	0.509	0.456	**0.813**	**0.674**	**0.032**	**0.602**

sample's leading interpretation to form consensus. Replacing prediction-guided centers with random initialization gives 71.5%. K-means still assigns images by visual distance, but the starting centers are selected using the current predictions. These comparisons support using semantic guidance in both center selection and prior estimation to obtain effective consensus evidence.

Semantic correction and mask refinement. Table 5 contrasts changes in category recognition and mask geometry. CCB improves recognition and class-aware metrics with unchanged class-agnostic masks. A TENT-lite adaptation of the decoder's normalization affine parameters (Wang et al., 2021) improves mask geometry, reaching .602 agIoU, while its recognition accuracy is 68.6%. This exposes a useful distinction: mask refinement and semantic refinement exploit different opportunities in OV-COS. CCB supplies the latter through test-population structure.

4.5 Consensus under Different Inference Contexts

Available population structure. Offline gains persist with smaller collections: batches of 512 and 1,024 yield 71.3% and 72.5% accuracy, respectively (Table 4). Changing the clustering resolution to 30 or 122 groups with random initialization also improves on the base model. The full 3,770-image collection supplies the strongest CLIP-only refinement, while smaller collections retain useful semantic structure. Its first-round clustering has NMI .754, ARI .466, and purity .723; final purity is approximately .728. These neighborhoods provide useful priors despite mixing categories.

Streaming observations. PC-WOC reaches 72.4% accuracy with a standard deviation of .2 points across five random input orders. When images arrive grouped by their ground-truth class, accuracy reaches 73.6%; this diagnostic ordering concentrates same-class evidence in the memory. A hard-to-easy stream gives 68.8%, showing that early access to reliable examples influences the benefit of consensus. These results connect online performance to the evidence available when each prediction is made. Figure 5 in Appendix F extends this comparison across five encoder configurations. Recognition varies by less than 1% relative to its mean in every configuration, and the segmentation metrics show similarly small variations. Table 18 reports the corresponding five-run statistics.

Computation. The CLIP-only module uses existing 768-dimensional image features. On OVCamo, their storage is approximately 11 MB, with about 30 MB peak memory during CPU clustering and a measured clustering overhead of .19 ms per image. Auxiliary-view systems additionally require frozen encoder feature extraction. PC-WOC replaces collection-level clustering with a centroid bank and one similarity computation per incoming image against the currently active class centroids. Its storage therefore depends on the candidate vocabulary size rather than the stream length.

5 Conclusion

CCB makes cross-sample consensus an explicit component of open-vocabulary inference. It estimates semantic priors from visual neighborhoods and combines them with individual predictions through confidence-adaptive fusion. Iterative multi-view clustering and streaming centroid memory offer two ways to obtain this evidence, improving existing OV-COS and OVSS models without gradient-based adaptation. The results highlight both the value of test-population structure and the complementary role of different neighborhood views. Extending this image-level formulation to region-level consensus offers a path toward consensus among multiple objects in complex visual scenes.

AI USE STATEMENT

Generative AI tools were used to assist with manuscript restructuring, language editing, mathematical exposition, presentation of existing experimental results, translation for internal author review, and LaTeX preparation. All reported empirical measurements originate from the authors' existing experiments, and all AI-assisted content was reviewed and verified by the authors. The authors take full responsibility for the accuracy and integrity of the final manuscript.

ETHICS STATEMENT

This work studies semantic refinement on existing image segmentation benchmarks, without collecting new data or conducting new human-subject studies. The method uses pretrained visual representations and predicted categories, so biases in those representations or in the observed image collection can influence the consensus prior. Camouflaged object recognition can support applications such as ecological monitoring, while deployment for surveillance can raise privacy concerns. Any downstream use should respect the source datasets' terms and assess privacy risks before deployment.

REPRODUCIBILITY STATEMENT

Section 3 specifies the consensus update and the offline and streaming implementations. Appendix A derives the fusion objective, Appendix B provides pseudocode, and Appendices C–D describe evaluation and model configurations. Appendices E–F report the complete quantitative comparisons and five-run statistics for all evaluated online encoder and feature-fusion configurations.

REFERENCES

Malik Boudiaf, Romain Mueller, Ismail Ben Ayed, and Luca Bertinetto. Parameter-free online test-time adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Coco-stuff: Thing and stuff classes in context. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1209–1218, 2018.

Seokju Cho, Heeseong Shin, Sunghwan Hong, Anurag Arnab, Paul Hongsuck Seo, and Seungryong Kim. Cat-seg: Cost aggregation for open-vocabulary semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4113–4123, 2024.

Mark Everingham, Luc Van Gool, Christopher K. I. Williams, John Winn, and Andrew Zisserman. The PASCAL visual object classes (VOC) challenge. *International Journal of Computer Vision*, 88(2):303–338, 2010. doi: 10.1007/s11263-009-0275-4.

Deng-Ping Fan, Ge-Peng Ji, Guolei Sun, Ming-Ming Cheng, Jianbing Shen, and Ling Shao. Camouflaged object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2777–2787, 2020.

Golnaz Ghiasi, Xiuye Gu, Yin Cui, and Tsung-Yi Lin. Scaling open-vocabulary image segmentation with image-level labels. In *European conference on computer vision*, pp. 540–557. Springer, 2022.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Yusuke Iwasawa and Yutaka Matsuo. Test-time classifier adjustment module for model-agnostic domain generalization. *Advances in Neural Information Processing Systems*, 34:2427–2440, 2021.

Thorsten Joachims et al. Transductive inference for text classification using support vector machines. In *Icml*, volume 99, pp. 200–209, 1999.

Adilbek Karmanov, Dayan Guan, Shijian Lu, Abdulmotaleb El Saddik, and Eric Xing. Efficient test-time adaptation of vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Feng Liang, Bichen Wu, Xiaoliang Dai, Kunpeng Li, Yinan Zhao, Hang Zhang, Peizhao Zhang, Peter Vajda, and Diana Marculescu. Open-vocabulary semantic segmentation with mask-adapted clip. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 7061–7070, 2023.

Jian Liang, Dapeng Hu, and Jiashi Feng. Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation. In *International conference on machine learning*, pp. 6028–6039. PMLR, 2020.

Yong Liu, Sule Bai, Guanbin Li, Yitong Wang, and Yansong Tang. Open-vocabulary segmentation with semantic-assisted calibration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3491–3500, 2024.

Sarthak Maharana, Baoming Zhang, Leonid Karlinsky, Rogerio Feris, and Yunhui Guo. Batclip: Bimodal online test-time adaptation for clip. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1569–1579, 2025.

Haiyang Mei, Ge-Peng Ji, Ziqi Wei, Xin Yang, Xiaopeng Wei, and Deng-Ping Fan. Camouflaged object segmentation with distraction mining. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8772–8781, 2021.

Mehrdad Noori, David Osowiechi, Gustavo Adolfo Vargas Hakim, Ali Bahri, Moslem Yazdanpanah, Sahar Dastani, Farzad Beizaee, Ismail Ben Ayed, and Christian Desrosiers. Test-time adaptation of vision-language models for open-vocabulary semantic segmentation. *arXiv preprint arXiv:2505.21844*, 2025.

Youwei Pang, Xiaoqi Zhao, Jiaming Zuo, Lihe Zhang, and Huchuan Lu. Open-vocabulary camouflaged object segmentation. In *European Conference on Computer Vision*, pp. 476–495. Springer, 2024.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. DINOv3, 2025. URL https://arxiv.org/abs/2508.10104.

Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. In *International Conference on Learning Representations*, 2021.

Qin Wang, Olga Fink, Luc Van Gool, and Dengxin Dai. Continual test-time domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7201–7211, 2022.

Jiarui Xu, Shalini De Mello, Sifei Liu, Wonmin Byeon, Thomas Breuel, Jan Kautz, and Xiaolong Wang. Groupvit: Semantic segmentation emerges from text supervision. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 18134–18144, 2022.

Mengde Xu, Zheng Zhang, Fangyun Wei, Han Hu, and Xiang Bai. Side adapter network for open-vocabulary semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2945–2954, 2023.

Bowen Yin, Xuying Zhang, Deng-Ping Fan, Shaohui Jiao, Ming-Ming Cheng, Luc Van Gool, and Qibin Hou. Camoformer: Masked separable attention for camouflaged object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

Marvin Zhang, Sergey Levine, and Chelsea Finn. Memo: Test time robustness via adaptation and augmentation. *Advances in neural information processing systems*, 35:38629–38642, 2022.

Yabin Zhang, Wenjie Zhu, Hui Tang, Zhiyuan Ma, Kaiyang Zhou, and Lei Zhang. Dual memory networks: A versatile adaptation approach for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Kai Zhao, Wubang Yuan, Zheng Wang, Guanyi Li, Xiaoqiang Zhu, Deng-Ping Fan, and Dan Zeng. Open-vocabulary camouflaged object segmentation with cascaded vision language models. *Computational Visual Media*, 2026. Accepted. arXiv:2506.19300.

Xiaojin Zhu, Zoubin Ghahramani, and John D Lafferty. Semi-supervised learning using gaussian fields and harmonic functions. In *Proceedings of the 20th International conference on Machine learning (ICML-03)*, pp. 912–919, 2003.

A Variational Interpretation of Consensus Fusion

For fixed positive distributions $p, r \in \Delta^{K-1}$ and a weight $\alpha \in [0, 1]$, consider

$$J(q) = \alpha D_{\mathrm{KL}}(q\|p) + (1-\alpha) D_{\mathrm{KL}}(q\|r). \tag{9}$$

Expanding the KL divergences and collecting their shared entropy term gives

$$J(q) = \sum_y q(y) \log q(y) - \sum_y q(y) \big[\alpha \log p(y) + (1-\alpha) \log r(y)\big]. \tag{10}$$

Define $s(y) = p(y)^\alpha r(y)^{1-\alpha}$ and $Z = \sum_y s(y)$. Then

$$J(q) = D_{\mathrm{KL}}\Big(q \,\Big\|\, \frac{s}{Z}\Big) - \log Z. \tag{11}$$

The second term is independent of q, and the divergence is minimized uniquely at $q = s/Z$. This recovers Eq. (2). Probability stabilization ensures that the logarithms are defined. The result characterizes each fusion step for its current evidence distributions and confidence weight.

In CCB, the cluster distribution is estimated from predicted categories. Under the auxiliary categorical count model $\theta_j \sim \mathrm{Dir}(\epsilon \mathbf{1})$, observing the counts n_j gives the posterior mean $(n_j + \epsilon \mathbf{1})/(|G_j| + K\epsilon)$. This provides the smoothed prior in Eq. (1). The empirical-Bayes interpretation refers to estimating this semantic prior from the unlabeled test population and then using it as a neighborhood-specific source of evidence for the confidence-adaptive update of individual predictions.

B Algorithms

Algorithm 1 describes offline CCB. The candidate vocabulary determines the default number of centers. Its class associations guide initialization; the K-means assignments themselves are based on visual embeddings. The update recomputes confidence and pseudo-labels after each round, and can switch the embedding view to obtain a complementary set of visual neighborhoods.

Algorithm 2 describes streaming PC-WOC. The current image can update its predicted-class centroid before obtaining the memory prior. Uninitialized centroids are excluded from the similarity softmax. The output requires neither future observations nor delayed batch formation.

Algorithm 1 Offline CCB with iterative multi-view consensus refinement

Require: Initial distributions $P^{(0)}$, embedding views $\{Z^{(r)}\}_{r=1}^{R}$, vocabulary $\mathcal{C}$, smoothing $\epsilon > 0$
Ensure: Refined distributions $P^{(R)}$

for $r = 1, \ldots, R$ **do**
  Compute $\hat{y}_i = \arg\max_y P_{iy}^{(r-1)}$ and α_i by Eq. (3)
  Initialize selected-index set $U \leftarrow \varnothing$
  for each class $c \in \mathcal{C}$ **do**
    $I_c \leftarrow \{i : \hat{y}_i = c\}$
    if $I_c \neq \varnothing$ **then**
      $i_c \leftarrow \arg\max_{i \in I_c} \alpha_i$
    else
      $i_c \leftarrow \arg\max_{i \notin U} P_{ic}^{(r-1)}$
    end if
    $\mu_c \leftarrow Z_{i_c:}^{(r)}$; $U \leftarrow U \cup \{i_c\}$
  end for
  $a \leftarrow \mathrm{KMeans}(Z^{(r)}; \{\mu_c\}_{c \in \mathcal{C}})$
  for each cluster j **do**
    $n_j(y) \leftarrow \sum_{i:a_i=j} \mathbb{I}[\hat{y}_i = y]$
    $r_j(y) \leftarrow (n_j(y) + \epsilon)/(\sum_y n_j(y) + K\epsilon)$
  end for
  for each image i **do**
    $\bar{p}_i \leftarrow \max(P_{i:}^{(r-1)}, \epsilon)$; normalize $\bar{p}_i$ to sum to one
    $P_{i:}^{(r)} \leftarrow \mathrm{softmax}(\alpha_i \log \bar{p}_i + (1 - \alpha_i) \log r_{a_i})$
  end for
end for
return $P^{(R)}$

Algorithm 2 Streaming PC-WOC with confidence-weighted centroid memory

Require: Stream $\{x_t\}$, frozen base model and encoder, vocabulary $\mathcal{C}$, threshold τ, exponent β, temperature κ

Initialize centroid bank V, mass vector M, and active set $\mathcal{A} \leftarrow \varnothing$
for each incoming image x_t **do**
  Obtain class probabilities p_t and normalized embedding z_t
  Compute confidence α_t and predicted class $\hat{y}_t$
  if $\alpha_t > \tau$ **then**
    $w_t \leftarrow \alpha_t^{\beta}$
    if $\hat{y}_t \notin \mathcal{A}$ **then**
      $v_{\hat{y}_t} \leftarrow z_t$; $M_{\hat{y}_t} \leftarrow w_t$
      $\mathcal{A} \leftarrow \mathcal{A} \cup \{\hat{y}_t\}$
    else
      Update $v_{\hat{y}_t}$ and $M_{\hat{y}_t}$ by Eq. (7)
    end if
  end if
  if $\mathcal{A} = \varnothing$ **then**
    $r_t \leftarrow p_t$
  else
    Compute active-centroid similarities and prior by Eq. (8)
    Scatter the prior into the K-class vocabulary
  end if
  Stabilize probabilities and output the fused prediction by Eq. (2)
end for

C Evaluation Protocol

OVCamo. OVCamo contains 11,483 images from 75 categories, partitioned into 7,713 training images from 14 categories and 3,770 test images from 61 other categories. We use the test category names as the candidate vocabulary. Offline CCB groups the available unlabeled inference collection, while PC-WOC processes the image sequence causally. Mini-batch offline evaluations apply clustering separately to batches of 512 or 1,024 images. The clustering-granularity experiments use 30 and 122 centers with random initialization; the default uses 61 prediction-guided centers.

VOC-S. The source pool is the 1,928-image Pascal VOC 2010 trainval segmentation split. We retain the largest valid foreground object per image as a binary target and discard images without a usable target. This yields 1,878 images spanning the 20 VOC categories. OVCoser is evaluated without retraining on this subset, with and without CLIP-only CCB.

COCO-S. The source pool consists of 5,000 images from COCO-Stuff 164K val2017. We exclude semantically ambiguous or non-object-like categories, retain the largest eligible foreground region in each remaining image, and convert it to a binary mask. The resulting evaluation contains 4,452 images and 57 categories. Global CLIP image embeddings are used for clustering. Both transfer subsets evaluate one target per image; other objects can remain in the image background.

Metrics. For a mask score $\mathcal{M}(\hat{m}, m)$, the class-aware counterpart is

$$c\mathcal{M} = \mathbb{I}[\hat{y} = y]\, \mathcal{M}(\hat{m}, m). \tag{12}$$

For MAE, an incorrect category receives unit error, giving the corresponding rule

$$c\mathrm{MAE} = \mathbb{I}[\hat{y} = y]\, \mathrm{MAE}(\hat{m}, m) + \mathbb{I}[\hat{y} \neq y]. \tag{13}$$

The prefix ag removes this category condition and evaluates the same predicted mask directly. cF_β, cE_m, and cIoU refer to averaged evaluation variants throughout this manuscript. Accuracy is reported as a percentage. For six class-aware metrics s_m and baseline values b_m, the aggregate relative gain is

$$\Delta = \frac{100}{6} \sum_{m=1}^{6} d_m \frac{s_m - b_m}{b_m}, \qquad d_m = \begin{cases} -1, & m \text{ is cMAE}, \\ +1, & \text{otherwise}. \end{cases} \tag{14}$$

Each table identifies the baseline used to compute relative gains. Measurements are displayed at the reported precision, and aggregate gains retain the values recorded in the source experiments.

Structure-measure combines object-aware and region-aware similarity with equal weights, $S_m = .5S_o + .5S_r$. Weighted F-measure uses $\beta^2 = 1$, whereas the ordinary F-measure uses $\beta^2 = .3$. MAE is the mean absolute difference between the predicted probability map and the binary target over all image pixels. E-measure combines pixel-level alignment with image-level statistics, and IoU measures the intersection of predicted and target foreground regions divided by their union. The class-aware rules above are applied to these mask scores before averaging over the evaluation images.

Mask selection for OVSS models. SAN, CAT-Seg, and SCAN produce multiple output masks. Their OVCamo evaluation selects the mask with the highest IoU with the target, following the oracle-mask protocol used for adapting these models to OV-COS evaluation. Each corresponding +CCB row uses exactly the same mask-selection protocol and pretrained model checkpoint.

D Model Configurations

CLIP-only inference. CCB uses the original model's class probabilities and frozen CLIP image embeddings. Its default offline sequence is CCC. Center selection uses the current entropy-margin confidence; the next round uses the preceding round's refined probabilities. The random-center ablation replaces prediction-guided centers with K-means using ten random initializations.

CVLM. The cascaded-VLM result uses raw CLIP features and two CCB rounds. The configuration sets `confidence_alpha` to 1.0, `prior_weight` to 1.0, and `prior_smoothing` to 0.05. This configuration produces the CVLM results in Table 1, including 76.7% recognition accuracy.

Streaming settings. PC-WOC uses $\lambda = .5$, $\beta = 2$, $\tau = .85$, and $\kappa = .07$. Random-order experiments use seeds 42, 1024, 2048, 3407, and 9999. The hard-to-easy diagnostic sorts images by ascending base-prediction confidence. The class-grouped diagnostic uses ground-truth category grouping to analyze the effect of consecutively observed same-class samples; the centroid updates continue to use predicted categories and their associated confidence scores.

TENT-lite. The comparison uses a frozen CLIP encoder and updates decoder-side normalization affine parameters with an entropy objective. The reported variant uses batch size two and one adaptation step per batch. Its class-aware and class-agnostic results are given in Table 9.

D.1 Auxiliary Feature Fusion

The enhanced systems use auxiliary features for mask decoding as well as clustering. Multi-scale features are extracted from five backbone stages. Intermediate ViT features are reshaped to spatial maps and resized to the decoder scales; CNN encoders provide their native stage features. At each scale, a 3×3 convolution projects the auxiliary features, which are concatenated with CLIP features and compressed by a 1×1 convolution. The fused features are then passed to the mask decoder. The encoders remain frozen while the fusion decoder is trained. The representative auxiliary encoders are DINOv3-ViT-L/16 (Siméoni et al., 2025) and ResNet-50-IN1K-v2 (He et al., 2016).

The complete offline sweep also includes ResNet-50-IN1K-v1, DINOv1-ResNet-50, and DINOv2-ViT-L/14, giving five auxiliary encoders at an input resolution of 448×448. In the sequence tables, F denotes decoder feature fusion; C denotes a CLIP clustering round; R denotes a round using the auxiliary ResNet features, including DINOv1; and D denotes a round using the auxiliary ViT features. Thus F + DDC means feature fusion followed by three consensus rounds, first using DINO twice and then CLIP once. Fusion is fixed within each encoder block throughout the sequence comparison.

E Extended Quantitative Results

Tables 7 and 8 report all six class-aware segmentation metrics for the representative configurations and ablations. The CLIP-only block of Table 7 uses the original base system. Each auxiliary-encoder block uses its own feature-fusion baseline, and the block's Δ measures the additional gain from CCB. Table 9 provides the full semantic/spatial comparison; Table 2 provides the full transfer results.

Table 6 expands the CVLM recognition results beyond top-1 accuracy. With two rounds of raw-CLIP consensus, macro-F1 rises from 69.91% to 71.40%, and top-5 accuracy rises from 90.66% to 92.31%. These gains show improved recognition when averaging equally across categories and when considering the five highest-ranked candidate categories for each image.

Table 6: **CVLM recognition on OVCamo.** The +CCB row uses raw CLIP features and two rounds, matching Table 1. All three metrics are percentages; higher is better.

Method	Top-1 accuracy	Macro-F1	Top-5 accuracy
CVLM	75.36	69.91	90.66
+CCB	**76.68**	**71.40**	**92.31**

Table 7: **Encoder integration and clustering sequences.** S denotes online inference; C, D, and R denote offline CLIP, DINOv3, and ResNet views. Each block has its own base system.

Method	Acc (%) ↑	cS_m ↑	cF_β^ω ↑	cMAE ↓	cF_β ↑	cE_m ↑	cIoU ↑	Δ(%)
Only CLIP								
Baseline	68.8	0.579	0.490	0.336	0.520	0.616	0.443	0.0
+$\boldsymbol{S_C}$	72.4	0.605	0.506	0.301	0.537	0.643	0.457	4.8
+$\boldsymbol{C}$	71.8	0.601	0.505	0.306	0.536	0.640	0.455	4.2
+$\boldsymbol{C}, \boldsymbol{C}$	72.6	0.606	0.509	0.299	0.541	0.647	0.459	5.4
+$\boldsymbol{C}, \boldsymbol{C}, \boldsymbol{C}$	73.2	0.611	0.513	0.294	0.545	0.651	0.462	6.2
+$\boldsymbol{C}, \boldsymbol{C}, \boldsymbol{C}, \boldsymbol{C}$	73.1	0.609	0.511	0.295	0.543	0.650	0.461	5.9
Fuse DINO Encoder								
Baseline	70.3	0.605	0.521	0.317	0.544	0.637	0.483	0.0
+$\boldsymbol{S_D}$	75.0	0.640	0.547	0.271	0.571	0.676	0.507	6.9
+$\boldsymbol{S_C}$	74.2	0.632	0.537	0.278	0.561	0.667	0.497	5.1
+$\boldsymbol{D}, \boldsymbol{D}, \boldsymbol{D}$	73.5	0.630	0.542	0.285	0.565	0.664	0.502	5.0
+$\boldsymbol{C}, \boldsymbol{C}, \boldsymbol{C}$	73.7	0.629	0.539	0.285	0.563	0.665	0.499	4.8
+$\boldsymbol{D}, \boldsymbol{D}, \boldsymbol{C}$	75.9	0.649	0.555	0.263	0.580	0.684	0.514	8.5
Fuse ResNet Encoder								
Baseline	70.1	0.590	0.510	0.322	0.535	0.630	0.465	0.0
+$\boldsymbol{S_R}$	73.2	0.612	0.523	0.292	0.550	0.654	0.478	4.2
+$\boldsymbol{S_C}$	74.1	0.617	0.525	0.283	0.552	0.660	0.479	5.1
+$\boldsymbol{R}, \boldsymbol{C}, \boldsymbol{C}, \boldsymbol{C}$	75.4	0.626	0.533	0.272	0.561	0.671	0.486	7.0

Table 8: **Complete component and context analysis.** Relative gains use each group's baseline and all six class-aware metrics. Prior-only accuracy is from round one; Table 16 traces its accuracy across rounds alongside confidence-adaptive CCB with the same CLIP embedding view.

Ablation Setting	Acc (%) ↑	cS_m ↑	cF_β^ω ↑	cMAE ↓	cF_β ↑	cE_m ↑	cIoU ↑	Δ(%)
Offline Clustering Components								
Baseline (CLIP, 3x clustering)	73.2	0.611	0.513	0.294	0.545	0.651	0.462	0.0
(1) Avg. Cluster Distribution	71.7	0.600	0.505	0.307	0.536	0.640	0.455	-2.1
(2) Simple Prob. Weighting	70.3	0.586	0.492	0.322	0.522	0.626	0.443	-5.0
(3) Fixed Confidence	72.0	0.601	0.506	0.306	0.537	0.640	0.456	-1.9
(4) No Cluster Center	71.5	0.596	0.500	0.310	0.530	0.637	0.451	-3.0
Online Clustering Input Order								
Baseline (Random Avg. of 5)	72.4	0.605	0.506	0.301	0.537	0.643	0.457	0.0
(1) Hard-to-Easy	68.8	0.579	0.490	0.335	0.520	0.616	0.443	-4.9
(2) GT Class Grouped	73.6	0.615	0.514	0.288	0.546	0.654	0.464	+2.1
Protocol Robustness and Relabeling Baselines								
Baseline (No CCB)	68.8	0.579	0.490	0.336	0.520	0.616	0.443	0.0
Prior-only	62.6	0.421	0.353	0.515	0.373	0.450	0.322	-31.8
Centroid correction	71.6	0.599	0.504	0.308	0.535	0.639	0.454	+4.0
Batch 512	71.3	0.596	0.501	0.312	0.532	0.636	0.452	+3.3
Batch 1024	72.5	0.604	0.506	0.300	0.537	0.645	0.456	+4.9
$N_c = 0.5K$	72.2	0.601	0.501	0.303	0.532	0.640	0.453	+4.1
$N_c = 2K$	72.1	0.604	0.508	0.304	0.539	0.642	0.457	+4.8

Table 9: **Complete semantic and spatial results.** Upper block: class-aware scores. Lower block: class-agnostic scores. Each Δ is relative to the block's OVCoser row.

Method	$S_m \uparrow$	$F_\beta^\omega \uparrow$	MAE $\downarrow$	$F_\beta \uparrow$	$E_m \uparrow$	IoU $\uparrow$	$\Delta(\%)$
Class-aware Metrics							
OVCoser	0.579	0.490	0.335	0.520	0.616	0.443	0.0
+CCB	**0.611**	**0.513**	**0.294**	**0.545**	**0.651**	**0.462**	**+6.2**
+TENT-lite	0.580	0.509	0.334	0.535	0.626	0.456	+2.0
Class-agnostic Metrics							
OVCoser	0.809	0.647	0.036	0.691	0.862	0.583	0.0
+CCB	0.809	0.647	0.036	0.691	0.862	0.583	0.0
+TENT-lite	**0.813**	**0.674**	**0.032**	**0.713**	**0.881**	**0.602**	**+4.1**

E.1 Encoder and View-Sequence Overview

Table 10 compares all five auxiliary encoders with the same three-round consensus budget. For each encoder, it contrasts repeated CLIP views, repeated auxiliary views, and the mixed sequence with the highest recognition accuracy; ties are resolved by cIoU. The best mixed sequence exceeds both repeated-view alternatives for every encoder in this sweep. This pattern connects the gains to complementary visual neighborhoods across successive refinement rounds.

Table 10: **Three-round comparison across auxiliary encoders.** F is the fusion-only accuracy; CCC repeats CLIP, and AAA repeats the corresponding auxiliary view. The mixed sequence is selected by accuracy, with cIoU as the tie-breaker. Accuracy is reported as a percentage.

		Repeated views		Best mixed sequence		
Auxiliary encoder	F	CCC	AAA	Views	Acc (%)	cIoU
ResNet-50-IN1K-v1	69.4	73.3	72.3	CRC	**74.1**	0.489
ResNet-50-IN1K-v2	70.1	73.0	72.0	RCC	**75.0**	0.484
DINOv1-ResNet-50	69.1	73.9	72.9	RRC	**76.1**	0.462
DINOv2-ViT-L/14	70.2	74.3	71.1	DCC	**75.2**	0.497
DINOv3-ViT-L/16	70.3	73.7	73.5	DDC	**75.9**	0.514

The recognition and segmentation rankings reveal complementary strengths. DINOv1-ResNet-50 reaches 76.1% accuracy with RRC, while DINOv3 reaches 75.9% with DDC and obtains the higher cIoU, .514 versus .462. ResNet-50-IN1K-v2 reaches 75.0% in three rounds and 75.4% with RCCC in four rounds. These results show that a compact auxiliary encoder can provide useful neighborhoods for semantic refinement, while the decoder's spatial representation shapes localization.

Tables 11–15 retain the complete one- to four-round sweeps: all 30 binary view sequences, a fusion-only row, and the recorded reference row for each encoder. The reference accuracy in this sweep is 69.3%; each encoder's F row is the starting point for comparing its consensus sequences. The tables report accuracy and all six class-aware mask scores directly, so that recognition, localization, and the effect of additional rounds can be inspected together across all view sequences.

E.2 RESNET-50-IN1K-V1

With ResNet-50-IN1K-v1, CRC reaches 74.1% accuracy in three rounds, compared with 73.3% for CCC and 72.3% for RRR. Adding a final CLIP round yields 74.3% with CRCC, while cIoU remains .489. Repeated ResNet refinement reaches 71.0% at four rounds. The ordering of complementary views therefore matters alongside the number of consensus updates in this encoder configuration.

Table 11: **Complete ResNet-50-IN1K-v1 sweep.** F denotes decoder feature fusion; each subsequent C or R denotes one consensus round. Bold entries mark the best value in each metric column, including ties; shading marks the sequence selected by accuracy and then cIoU.

Sequence	Acc (%) ↑	cS_m ↑	cF_β^ω ↑	cMAE ↓	cF_β ↑	cE_m ↑	cIoU ↑
Reference and fusion-only configurations							
Reference	69.3	0.581	0.498	0.330	0.526	0.622	0.449
F	69.4	0.589	0.514	0.327	0.537	0.630	0.468
1 consensus round							
F + C	72.4	0.610	0.527	0.298	0.551	0.653	0.480
F + R	72.0	0.608	0.527	0.302	0.551	0.650	0.480
2 consensus rounds							
F + CC	73.4	0.617	0.531	0.289	0.556	0.661	0.484
F + CR	73.9	0.620	0.534	0.285	0.559	0.664	0.486
F + RC	73.7	0.619	0.534	0.286	0.559	0.664	0.487
F + RR	73.0	0.614	0.531	0.293	0.555	0.657	0.484
3 consensus rounds							
F + CCC	73.3	0.616	0.531	0.290	0.556	0.661	0.484
F + CCR	74.0	0.621	0.534	0.284	0.559	0.666	0.487
F + CRC	74.1	0.622	0.536	0.282	0.561	0.667	**0.489**
F + CRR	73.4	0.617	0.532	0.289	0.557	0.662	0.485
F + RCC	73.7	0.619	0.534	0.286	0.559	0.664	0.487
F + RCR	73.2	0.615	0.530	0.291	0.555	0.659	0.483
F + RRC	73.9	0.618	0.531	0.284	0.555	0.662	0.484
F + RRR	72.3	0.608	0.524	0.300	0.549	0.650	0.478
4 consensus rounds							
F + CCCC	73.2	0.614	0.529	0.291	0.554	0.659	0.482
F + CCCR	73.1	0.614	0.529	0.292	0.553	0.659	0.482
F + CCRC	74.0	0.621	0.535	0.283	0.560	0.666	0.488
F + CCRR	73.8	0.619	0.533	0.286	0.558	0.664	0.486
F + CRCC	**74.3**	**0.623**	**0.537**	**0.281**	**0.562**	**0.669**	**0.489**
F + CRCR	74.0	0.621	0.535	0.283	0.560	0.666	0.488
F + CRRC	73.5	0.617	0.532	0.288	0.556	0.662	0.485
F + CRRR	72.3	0.608	0.526	0.300	0.550	0.652	0.479
F + RCCC	73.2	0.615	0.531	0.291	0.555	0.660	0.484
F + RCCR	73.3	0.616	0.531	0.290	0.556	0.660	0.484
F + RCRC	73.7	0.619	0.534	0.286	0.558	0.663	0.487
F + RCRR	72.3	0.608	0.524	0.299	0.549	0.652	0.478
F + RRCC	74.2	0.621	0.532	**0.281**	0.557	0.665	0.485
F + RRCR	73.4	0.615	0.527	0.289	0.551	0.658	0.481
F + RRRC	73.0	0.611	0.525	0.293	0.549	0.654	0.478
F + RRRR	71.0	0.597	0.516	0.312	0.539	0.638	0.470

E.3 RESNET-50-IN1K-V2

The ResNet-50-IN1K-v2 sweep begins at 70.1% accuracy with fusion alone. RCC raises accuracy to 75.0% in three rounds, and RCCC reaches 75.4% with cIoU .486 in four rounds. In contrast, RRRR gives 71.2%. An initial ResNet neighborhood followed by CLIP refinement provides the strongest recognition and class-aware overlap within this encoder's evaluated view sequences.

Table 12: **Complete ResNet-50-IN1K-v2 sweep.** C uses CLIP neighborhoods and R uses auxiliary ResNet neighborhoods. The fusion decoder is fixed across sequences. Bold entries mark column-wise optima; the shaded row is selected by accuracy and then cIoU over all evaluated round counts.

Sequence	Acc (%) ↑	cS_m ↑	cF_β^ω ↑	cMAE ↓	cF_β ↑	cE_m ↑	cIoU ↑
Reference and fusion-only configurations							
Reference	69.3	0.581	0.498	0.330	0.526	0.622	0.449
F	70.1	0.590	0.510	0.322	0.535	0.630	0.465
1 consensus round							
F + C	73.5	0.613	0.524	0.290	0.552	0.656	0.478
F + R	72.5	0.607	0.520	0.299	0.547	0.649	0.475
2 consensus rounds							
F + CC	73.7	0.615	0.526	0.289	0.554	0.658	0.480
F + CR	74.7	0.623	0.531	0.279	0.559	0.667	0.485
F + RC	74.5	0.620	0.529	0.280	0.558	0.663	0.482
F + RR	72.5	0.608	0.522	0.298	0.549	0.650	0.476
3 consensus rounds							
F + CCC	73.0	0.612	0.526	0.295	0.553	0.654	0.480
F + CCR	73.7	0.616	0.527	0.288	0.554	0.659	0.481
F + CRC	74.0	0.617	0.527	0.285	0.555	0.661	0.481
F + CRR	74.0	0.617	0.527	0.285	0.554	0.662	0.481
F + RCC	75.0	0.623	0.531	0.276	0.559	0.668	0.484
F + RCR	74.1	0.619	0.530	0.284	0.557	0.663	0.483
F + RRC	73.9	0.617	0.528	0.286	0.555	0.661	0.482
F + RRR	72.0	0.605	0.519	0.302	0.546	0.647	0.474
4 consensus rounds							
F + CCCC	72.7	0.609	0.523	0.298	0.550	0.651	0.478
F + CCCR	72.9	0.611	0.525	0.296	0.552	0.653	0.479
F + CCRC	73.7	0.615	0.527	0.289	0.554	0.659	0.481
F + CCRR	72.9	0.610	0.522	0.296	0.549	0.652	0.477
F + CRCC	73.9	0.615	0.525	0.287	0.552	0.660	0.479
F + CRCR	73.3	0.611	0.521	0.292	0.548	0.655	0.475
F + CRRC	73.9	0.614	0.523	0.286	0.550	0.658	0.476
F + CRRR	72.7	0.607	0.518	0.297	0.545	0.651	0.473
F + RCCC	**75.4**	**0.626**	**0.533**	**0.272**	**0.561**	**0.671**	**0.486**
F + RCCR	74.5	0.619	0.527	0.280	0.555	0.664	0.480
F + RCRC	73.3	0.612	0.524	0.292	0.551	0.655	0.477
F + RCRR	73.1	0.611	0.523	0.293	0.550	0.653	0.477
F + RRCC	73.9	0.616	0.526	0.286	0.553	0.660	0.479
F + RRCR	73.5	0.614	0.524	0.289	0.551	0.657	0.478
F + RRRC	73.6	0.616	0.526	0.288	0.554	0.658	0.480
F + RRRR	71.2	0.598	0.512	0.310	0.539	0.639	0.468

E.4 DINOv1-ResNet-50

DINOv1-ResNet-50 gives 69.1% accuracy with fusion alone and reaches 76.1% with RRC. This is the highest recognition accuracy in the five-encoder sweep. Its best cIoU, .465, instead occurs with CCCC, whose accuracy is 74.4%. The complete metric columns capture this distinction between selecting a sequence for category recognition and selecting it for class-aware segmentation.

Table 13: **Complete DINOv1-ResNet-50 sweep.** R denotes DINOv1 ResNet features and C denotes CLIP features. Bold entries mark the best value for each metric, including ties. The shaded RRC sequence maximizes recognition accuracy, whereas CCCC maximizes class-aware IoU.

Sequence	Acc (%) ↑	cS_m ↑	cF_β^ω ↑	cMAE ↓	cF_β ↑	cE_m ↑	cIoU ↑
Reference and fusion-only configurations							
Reference	69.3	0.581	0.498	0.330	0.526	0.622	0.449
F	69.1	0.573	0.497	0.332	0.524	0.610	0.445
1 consensus round							
F + C	72.0	0.592	0.508	0.304	0.536	0.631	0.454
F + R	71.9	0.593	0.509	0.304	0.538	0.631	0.455
2 consensus rounds							
F + CC	73.3	0.600	0.513	0.293	0.542	0.640	0.459
F + CR	74.7	0.609	0.518	0.279	0.548	0.650	0.463
F + RC	74.0	0.605	0.516	0.286	0.546	0.645	0.461
F + RR	73.7	0.602	0.510	0.287	0.539	0.643	0.456
3 consensus rounds							
F + CCC	73.9	0.604	0.517	0.288	0.546	0.645	0.462
F + CCR	73.8	0.603	0.515	0.288	0.544	0.643	0.460
F + CRC	74.1	0.605	0.515	0.284	0.544	0.645	0.460
F + CRR	74.8	0.608	0.513	0.277	0.543	0.650	0.459
F + RCC	74.1	0.605	0.516	0.285	0.545	0.645	0.461
F + RCR	74.3	0.605	0.514	0.283	0.543	0.645	0.459
F + RRC	**76.1**	**0.615**	0.516	**0.266**	0.546	**0.657**	0.462
F + RRR	72.9	0.594	0.501	0.295	0.529	0.635	0.449
4 consensus rounds							
F + CCCC	74.4	0.608	**0.520**	0.283	**0.549**	0.649	**0.465**
F + CCCR	73.7	0.602	0.515	0.289	0.544	0.643	0.461
F + CCRC	73.4	0.600	0.512	0.292	0.541	0.641	0.458
F + CCRR	73.0	0.596	0.509	0.296	0.538	0.636	0.455
F + CRCC	73.8	0.602	0.513	0.288	0.543	0.643	0.459
F + CRCR	73.5	0.599	0.510	0.291	0.539	0.639	0.456
F + CRRC	74.6	0.605	0.510	0.280	0.540	0.647	0.456
F + CRRR	74.6	0.603	0.506	0.279	0.535	0.645	0.452
F + RCCC	74.0	0.604	0.515	0.286	0.545	0.645	0.460
F + RCCR	74.1	0.604	0.515	0.284	0.544	0.645	0.460
F + RCRC	73.7	0.600	0.511	0.289	0.540	0.641	0.456
F + RCRR	73.7	0.599	0.507	0.288	0.536	0.639	0.453
F + RRCC	74.0	0.602	0.510	0.286	0.539	0.642	0.457
F + RRCR	75.0	0.606	0.507	0.275	0.537	0.648	0.454
F + RRRC	74.5	0.602	0.505	0.280	0.534	0.645	0.452
F + RRRR	71.1	0.578	0.487	0.312	0.515	0.619	0.437

E.5 DINOv2-ViT-L/14

For DINOv2, DCC reaches 75.2% accuracy and .497 cIoU, improving on both CCC at 74.3% and DDD at 71.1%. A fourth repeated DINO round reduces accuracy to 69.2%, whereas CCCC retains 73.5%. The strongest three-round sequence combines an initial DINO neighborhood with two CLIP refinements, showing how subsequent CLIP views refine the initial DINO consensus.

Table 14: **Complete DINOv2-ViT-L/14 sweep.** D denotes the auxiliary DINOv2 view and C denotes CLIP. F is feature fusion and does not contribute to the round count. Bold entries mark column-wise optima; shading identifies the sequence selected by accuracy and then cIoU from the complete sweep.

Sequence	Acc (%) ↑	cS_m ↑	cF_β^ω ↑	cMAE ↓	cF_β ↑	cE_m ↑	cIoU ↑
Reference and fusion-only configurations							
Reference	69.3	0.581	0.498	0.330	0.526	0.622	0.449
F	70.2	0.598	0.525	0.318	0.548	0.639	0.476
1 consensus round							
F + C	73.9	0.626	0.543	0.280	0.569	0.671	0.492
F + D	72.1	0.610	0.532	0.301	0.557	0.654	0.483
2 consensus rounds							
F + CC	74.8	0.629	0.546	0.275	0.572	0.674	0.494
F + CD	74.2	0.625	0.542	0.280	0.568	0.670	0.491
F + DC	74.8	0.629	0.545	0.275	0.571	0.675	0.494
F + DD	72.5	0.613	0.534	0.297	0.559	0.657	0.485
3 consensus rounds							
F + CCC	74.3	0.625	0.542	0.280	0.568	0.670	0.491
F + CCD	74.6	0.627	0.543	0.277	0.569	0.672	0.492
F + CDC	74.9	0.629	0.545	0.275	0.571	0.675	0.494
F + CDD	73.4	0.618	0.536	0.289	0.562	0.662	0.486
F + DCC	**75.2**	**0.633**	**0.549**	**0.271**	**0.575**	**0.679**	**0.497**
F + DCD	74.4	0.626	0.542	0.279	0.568	0.671	0.491
F + DDC	74.5	0.627	0.543	0.278	0.569	0.672	0.492
F + DDD	71.1	0.602	0.525	0.310	0.549	0.645	0.476
4 consensus rounds							
F + CCCC	73.5	0.619	0.538	0.288	0.563	0.664	0.488
F + CCCD	74.6	0.627	0.544	0.277	0.570	0.672	0.493
F + CCDC	74.6	0.627	0.544	0.278	0.570	0.672	0.492
F + CCDD	74.1	0.623	0.540	0.282	0.566	0.668	0.489
F + CDCC	74.9	0.630	0.546	0.274	0.572	0.675	0.495
F + CDCD	74.5	0.626	0.543	0.278	0.570	0.671	0.492
F + CDDC	74.4	0.626	0.543	0.279	0.570	0.671	0.492
F + CDDD	71.4	0.603	0.525	0.308	0.550	0.646	0.476
F + DCCC	74.5	0.628	0.545	0.278	0.570	0.673	0.494
F + DCCD	74.8	0.629	0.546	0.275	0.572	0.675	0.494
F + DCDC	74.9	0.630	0.546	0.274	0.572	0.675	0.494
F + DCDD	72.9	0.613	0.532	0.293	0.557	0.658	0.482
F + DDCC	74.7	0.628	0.544	0.276	0.570	0.674	0.493
F + DDCD	73.6	0.620	0.538	0.286	0.563	0.664	0.488
F + DDDC	73.3	0.617	0.536	0.289	0.561	0.662	0.485
F + DDDD	69.2	0.586	0.511	0.328	0.535	0.628	0.464

E.6 DINOv3-ViT-L/16

DINOv3 reaches 75.9% accuracy with either DCD or DDC in three rounds; DDC gives the higher cIoU, .514 versus .512. DDCC preserves .514 cIoU in four rounds with 75.8% accuracy. Repeating only DINO views gives 73.5% with DDD and 72.9% with DDDD. Together, these sequences show the benefit of introducing a complementary CLIP neighborhood after the initial DINO updates.

Table 15: **Complete DINOv3-ViT-L/16 sweep.** D and C denote DINOv3 and CLIP consensus rounds following the same fusion decoder. Bold entries mark column-wise optima, including ties; the shaded DDC sequence is selected by recognition accuracy with cIoU used to resolve the tie with DCD.

Sequence	Acc (%) ↑	cS_m ↑	cF_β^ω ↑	cMAE ↓	cF_β ↑	cE_m ↑	cIoU ↑
Reference and fusion-only configurations							
Reference	69.3	0.581	0.498	0.330	0.526	0.622	0.449
F	70.3	0.605	0.521	0.317	0.544	0.637	0.483
1 consensus round							
F + C	73.1	0.626	0.537	0.290	0.561	0.661	0.497
F + D	73.3	0.627	0.537	0.288	0.561	0.661	0.498
2 consensus rounds							
F + CC	74.0	0.633	0.541	0.281	0.566	0.668	0.501
F + CD	74.6	0.637	0.546	0.276	0.571	0.673	0.505
F + DC	75.6	0.645	0.550	0.266	0.575	0.681	0.509
F + DD	74.4	0.637	0.546	0.277	0.570	0.672	0.506
3 consensus rounds							
F + CCC	73.7	0.629	0.539	0.285	0.563	0.665	0.499
F + CCD	74.3	0.636	0.545	0.278	0.569	0.672	0.505
F + CDC	74.4	0.637	0.547	0.278	0.571	0.674	0.507
F + CDD	74.7	0.640	0.549	0.274	0.573	0.675	0.508
F + DCC	75.1	0.641	0.549	0.271	0.573	0.678	0.508
F + DCD	**75.9**	0.647	0.553	**0.263**	0.577	**0.684**	0.512
F + DDC	**75.9**	**0.649**	**0.555**	**0.263**	**0.580**	**0.684**	**0.514**
F + DDD	73.5	0.630	0.542	0.285	0.565	0.664	0.502
4 consensus rounds							
F + CCCC	73.4	0.627	0.537	0.287	0.561	0.663	0.498
F + CCCD	73.6	0.630	0.540	0.285	0.563	0.665	0.500
F + CCDC	73.8	0.631	0.541	0.284	0.565	0.667	0.502
F + CCDD	73.9	0.632	0.542	0.283	0.566	0.668	0.502
F + CDCC	74.0	0.634	0.545	0.282	0.569	0.670	0.505
F + CDCD	73.9	0.633	0.544	0.282	0.568	0.669	0.504
F + CDDC	75.7	0.645	0.551	0.265	0.575	0.681	0.510
F + CDDD	74.4	0.636	0.546	0.277	0.569	0.672	0.505
F + DCCC	74.2	0.635	0.546	0.280	0.570	0.671	0.506
F + DCCD	75.5	0.643	0.548	0.267	0.573	0.679	0.508
F + DCDC	75.7	0.645	0.550	0.265	0.575	0.681	0.510
F + DCDD	75.2	0.641	0.547	0.270	0.571	0.677	0.507
F + DDCC	75.8	0.648	**0.555**	0.264	0.579	0.683	**0.514**
F + DDCD	75.1	0.643	0.551	0.271	0.575	0.677	0.510
F + DDDC	75.1	0.641	0.547	0.271	0.571	0.676	0.507
F + DDDD	72.9	0.624	0.535	0.291	0.558	0.658	0.496

E.7 Consensus Dynamics across Rounds

Table 16 follows recognition over three updates. Prior-only relabeling progressively loses accuracy, while confidence-adaptive CCB improves from 71.83% to 73.18%. Preserving individual evidence thus supports repeated use of the cluster prior. With 30 or 122 randomly initialized centers, accuracy peaks at round two; the default prediction-guided setting improves through round three.

Table 16: **Recognition trajectories on OVCamo.** Accuracy (%) by round, starting from 68.81%. Default CCB has 61 prediction-guided centers; other counts use random initialization.

Setting	Round 1	Round 2	Round 3
Prior-only	62.57	59.26	50.11
CCB (default)	71.83	72.57	73.18
CCB, $N_c = 30$	72.47	73.42	72.20
CCB, $N_c = 122$	71.54	72.23	72.12

Table 17 connects refinement to the visual neighborhoods available at each collection size. The full collection has higher NMI and ARI than either mini-batch setting in every round. Recognition improves across all three rounds at each size, while agreement with the preceding predictions increases. The smaller batches make fewer label changes but reach lower final accuracy, linking the available population structure to the scope of semantic correction across images.

Table 17: **Per-round clustering and recognition.** NMI, ARI, and purity describe each round's clusters; accuracy is measured after fusion. Agreement compares labels before and after that round's update. Accuracy and agreement are percentages; shaded rows mark the final round.

Collection size	Round	NMI	ARI	Purity	Acc (%)	Agreement (%)
Full (3,770)	1	.7538	.4655	.7228	71.83	88.83
	2	.7538	.4655	.7228	72.57	91.51
	3	.7635	.5090	.7284	73.18	93.40
Batch 1,024	1	.6523	.1649	.6971	70.80	90.64
	2	.6613	.1787	.7164	71.96	92.52
	3	.6588	.1750	.7069	72.49	94.11
Batch 512	1	.6310	.0962	.7225	70.13	92.49
	2	.6258	.0948	.7183	71.09	94.75
	3	.6279	.0948	.7228	71.30	95.84

F Online Stability

Table 18 reports means and standard deviations across five random input orders. The CLIP-only system has an accuracy standard deviation of .2 percentage points, while the enhanced-system configurations range from .2 to .6 points. These results complement the structured-order diagnostics in Table 4: random permutations give stable aggregate performance, whereas concentrating low-confidence images at the beginning changes the information available for early predictions.

Table 18: **Five-run online statistics.** Each mean row is followed by its standard deviation; accuracy is in percentage points. Encoders used for the centroid memory are indicated by subscripts.

Method	Acc (%) ↑	cS_m ↑	cF_β^ω ↑	cMAE ↓	cF_β ↑	cE_m ↑	cIoU ↑
Only CLIP							
$+\boldsymbol{S_C}$	72.4	0.605	0.506	0.301	0.537	0.643	0.457
std	± 0.2	± 0.001	± 0.001	± 0.002	± 0.001	± 0.002	± 0.001
Fuse DINO Encoder							
$+\boldsymbol{S_D}$	75.0	0.640	0.547	0.271	0.571	0.676	0.507
std	± 0.4	± 0.002	± 0.001	± 0.004	± 0.001	± 0.003	± 0.001
$+\boldsymbol{S_C}$	74.2	0.632	0.537	0.278	0.561	0.667	0.497
std	± 0.6	± 0.003	± 0.001	± 0.006	± 0.001	± 0.004	± 0.001
Fuse ResNet Encoder							
$+\boldsymbol{S_R}$	73.2	0.612	0.523	0.292	0.550	0.654	0.478
std	± 0.2	± 0.002	± 0.001	± 0.002	± 0.002	± 0.002	± 0.001
$+\boldsymbol{S_C}$	74.1	0.617	0.525	0.283	0.552	0.660	0.479
std	± 0.3	± 0.002	± 0.001	± 0.003	± 0.001	± 0.002	± 0.001

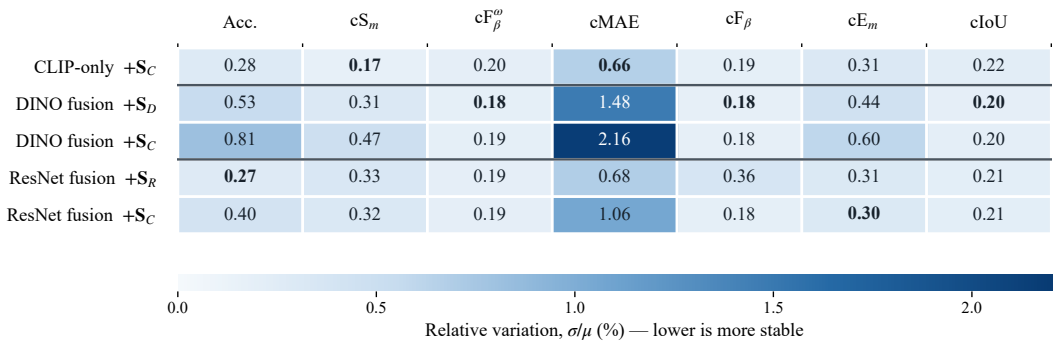

Figure 5: **Streaming stability across encoder configurations.** Relative variation $100\sigma/\mu$ over five random input orders, computed from the reported means and standard deviations. C and D/R denote CLIP and auxiliary clustering views. Darker cells indicate larger variation across orders, and bold entries mark the lowest relative variation in each reported metric column.

G Qualitative Results

Figures 6 and 7 present the enhanced DINO feature-fusion system with CCB. The category annotations highlight the semantic contribution of consensus, while the masks show the complete system's localization. The examples span varied target sizes and challenging forms of visual camouflage.

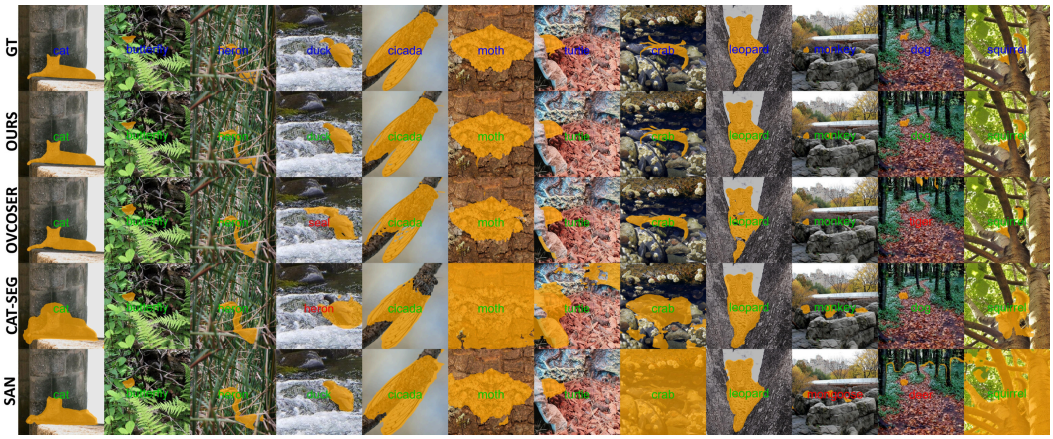

Figure 6: **Complete qualitative comparison on OVCamo.** Ground truth and the predictions of CCB, OVCoser, CAT-Seg, and SAN are arranged from top to bottom. Figure 4 enlarges ten of these examples in the main text, retaining all five comparison rows and their category predictions.

H Discussion and Extensions

The image-level formulation is well matched to one-target-category evaluation. A region-level extension could organize mask embeddings across images, allowing different objects within an image to draw on distinct neighborhoods. This would extend consensus to multi-category segmentation with each region drawing its semantic prior from visually related regions in the collection.

Consensus quality depends on both the embedding geometry and the base predictions. A confidently wrong prediction may resist correction, while an ambiguous sample can inherit errors from a biased cluster. The current confidence score and smoothing moderate this interaction. The finite-round results also motivate studying adaptive iteration schedules and cluster-quality-aware weighting. Streaming settings introduce an additional variable: the order in which useful evidence becomes available. PC-WOC's accumulating memory provides stable estimates, while handling persistent distribution drift could be supported by selectively renewing centroids as the observed visual distribution changes.

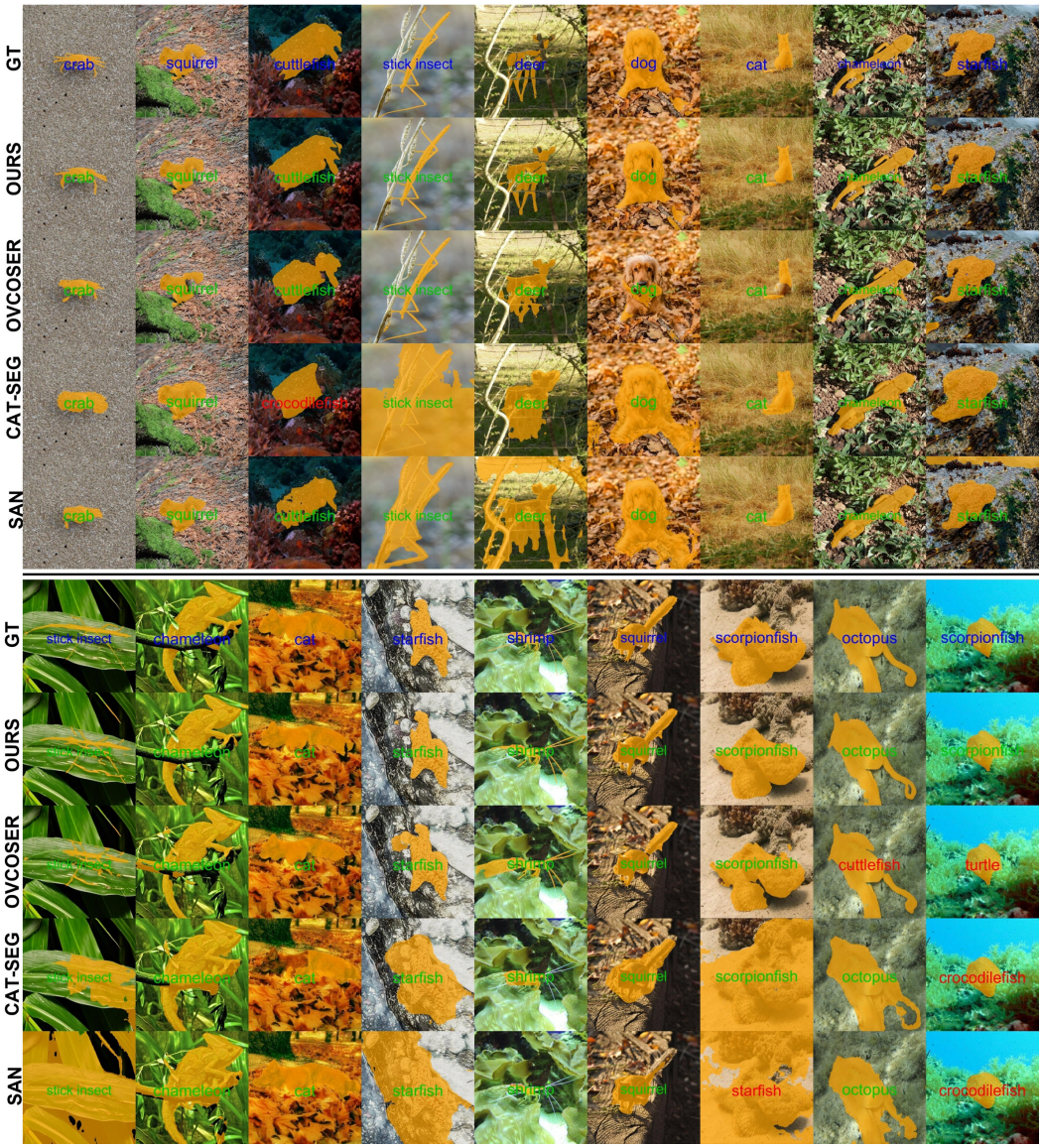

Figure 7: **Additional enhanced-system examples.** Category annotations and foreground masks show how consensus-based recognition combines with the visual predictions of the feature-fusion model.