COVSS: A Benchmark for Continual Open-Vocabulary Semantic Segmentation

Anonymous authors
Paper under double-blind review

Abstract

Semantic segmentation studies commonly evaluate continual updating and open-vocabulary recognition separately. Emerging work at their intersection leaves less understood how these capabilities develop together when supervision arrives incrementally. We introduce Continual Open-Vocabulary Semantic Segmentation (COVSS), a task and benchmark for studying this process through multiple class increments of comparable size. Built from existing COCO-Stuff171 and ADE150 annotations, COVSS specifies class partitions, stage-visible supervision, shared comparison conditions, and trajectory-based evaluation on internal and external vocabularies. A reproducible reference implementation combines anonymous scene supervision with dynamic semantic alignment. Experiments reveal weaknesses in both acquisition and retention among conventional continual segmentation methods, while open-vocabulary methods exhibit markedly different learning trajectories. Without dedicated anti-forgetting mechanisms, the COVSS baseline attains strong overall performance and positive net changes in previously learned categories, despite local declines. Pixel-level analysis attributes part of these gains to correcting earlier false positives through changes in competing category responses. These observations establish a benchmark for investigating how continual learning and open-vocabulary prediction interact, and identify challenges for developing scene understanding from successive, partially labeled experiences.

1 Introduction

Objects, concepts, and application requirements change over the lifetime of a visual system. A segmentation model may need to learn newly annotated categories while retaining useful prior capabilities and responding to requests outside its accumulated supervision. This is relevant to robots entering new environments or inspection systems receiving new annotation priorities: annotations can be extended gradually, but the scene contains more than the categories currently labeled. Continual learning concerns learning from successive experience (De Lange et al., 2022; Wang et al., 2024), while open-vocabulary segmentation (OVS) enables prediction from category descriptions. Their interaction matters when neither the learning process nor the required vocabulary ends at deployment.

Existing segmentation research provides complementary starting points. Continual semantic segmentation (CSS) studies sequentially introduced categories and background shift (Cermelli et al., 2020; Douillard et al., 2021); OVS transfers vision–language representations to dense prediction (Liang et al., 2023; Xu et al., 2023; Cho et al., 2024). Cross-over methods now incorporate language priors into CSS or adapt OVS over time (Wu et al., 2025; Gui et al., 2024; Hwang et al., 2025). However, these settings differ in what supervision arrives, how knowledge is retained, and which capabilities are measured. Good final performance on an expanding closed vocabulary does not establish continued performance on external queries, while a strong pretrained OVS model need not improve as new labels arrive.

We introduce **Continual Open-Vocabulary Semantic Segmentation (COVSS)** to study the progressive formation of scene understanding under class-incremental supervision. The benchmark distributes category acquisition across ten similarly sized stages and evaluates both learned and not-yet-supervised categories, together with external vocabularies, throughout learning. This makes each stage consequential. In contrast, the first 100 ADE150 categories used in common CSS settings

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.

account for 97.80% of valid training pixels. Such settings primarily examine extensions to an already broadly supervised system. COVSS also covers an earlier application stage, when useful scene understanding must develop from a limited initial set of labeled categories.

The contribution is a *learning and evaluation benchmark built from existing segmentation data*. Its construction turns complete annotations into reproducible stage-specific training views, controls category-identity access, and retains a common evaluation vocabulary across checkpoints. The resulting category-group matrices distinguish acquisition, interference, recovery, and prediction before direct supervision. A reference baseline makes this setting executable and provides an informative point of comparison. It builds on OVS: anonymous labels retain scene constraints outside current categories, and dynamic semantic alignment reuses spatial and query relationships across stages. The task itself does not require this architecture or prohibit replay and other continual-learning mechanisms.

The benchmark reveals more than a change in method rankings. Some CSS methods learn current categories well but lose much of that performance later; some OVS methods also deteriorate, whereas others improve earlier categories through later learning. The baseline combines strong acquisition, positive net post-supervision change, and external-vocabulary performance. Fixed-pixel analysis shows that an old category can improve because other responses correct its earlier false positives, even when its recall decreases. This connects continual improvement to the evolving organization of scene predictions. Our contributions are:

- A reproducible COVSS task and benchmark that distributes supervision across comparable class increments and follows internal and external segmentation capabilities through the complete learning process.
- A reference OVS-based implementation and an evaluation framework separating initial acquisition, subsequent change, and responses beyond direct supervision.
- Comparisons and trajectory analyses revealing distinct forms of degradation and improvement, with pixel-level evidence explaining a source of subsequent gains.

2 Related Work

Continual segmentation. CSS inherits class-incremental learning while addressing changing pixel-label semantics (van de Ven et al., 2022; Cermelli et al., 2020). MiB and PLOP treat background shift and retention through modified losses and distillation; SDR regularizes representations (Douillard et al., 2021; Michieli & Zanuttigh, 2021). More recent approaches include decoupled distillation, classifier adaptation, and visual-query replay (Baek et al., 2022; Oh et al., 2022; Zhu et al., 2025). Existing experiments already include balanced short streams, and same-length comparisons indicate that distributing acquisition more evenly can be harder (Yin et al., 2025). COVSS does not claim the first balanced split. It makes repeated substantial acquisition the standard condition and evaluates its interaction with open-vocabulary prediction.

Open-vocabulary segmentation and its continual extensions. OVS uses vision–language models such as CLIP (Radford et al., 2021) to associate dense visual features with category descriptions. LSeg, MaskCLIP, OVSeg, and SAN establish different adaptation routes; CAT-Seg and SED develop matching aggregation and encoder–decoder designs (Li et al., 2022; Ding et al., 2022; Liang et al., 2023; Xu et al., 2023; Cho et al., 2024; Xie et al., 2024). Intersection work changes different parts of this process: LBD introduces language-derived semantics into class-incremental prediction, kNN-CLIP grows a retrieval database around frozen representations, and ConOVS incorporates stage-specific experts (Wu et al., 2025; Gui et al., 2024; Hwang et al., 2025). These are related but not interchangeable learning settings. COVSS studies one continually updated segmentation state under a shared identity boundary, measuring both its class-incremental trajectory and its external queries. Its distinctive value is the combination of a distributed learning burden, joint internal/external observation, and common comparison conditions, rather than any one of these ingredients alone.

CUE additionally considers vocabulary and referring-expression inputs (Lin et al., 2025), while OVS adaptation addresses open generalization after new supervision (Peng et al., 2025). Table 1 summarizes the research targets and cross-stage carriers.

Table 1: Research targets and cross-stage carriers in the original settings of representative OVS–CSS intersection work. ✓: fully included; △: partial or system-level observation; ×: not a primary condition.

Direction	Class-incremental	Joint in/out trajectories	One updated OVS state	Carrier
LBD	✓	×	×	model + distillation
CUE	✓	△	△	fixed experts
OVS adaptation	×	△	✓	model adaptation
kNN-CLIP	×	△	×	frozen + retrieval
ConOVS	△	△	×	growing experts
COVSS	✓	✓	✓	**shared model state**

3 Constructing the COVSS Benchmark

The benchmark is designed to separate three questions that a final score alone cannot answer: whether a model learns each new group, whether later learning improves or damages earlier predictions, and whether open queries remain useful. Table 1 locates this objective relative to established settings. The construction below keeps the data sources fixed while specifying the sequence of supervision and the observations needed to answer these questions.

3.1 Data Sources and Task-Stream Construction

We use the original training and validation splits of COCO-Stuff171 and ADE150, retaining their images and dense annotations (Caesar et al., 2018; Zhou et al., 2017). The benchmark reuses these annotations to control when category identities become available; it does not collect new images or assign new semantic labels. Source label IDs and names define each dataset's internal category vocabulary. Validation images remain separate from stage construction.

Construction has four steps. First, partition the internal vocabulary into disjoint groups. Second, select training image–mask pairs containing at least one pixel of the current group. Third, expose the identities of that group and conceal all others. Finally, evaluate each inherited checkpoint on the unchanged internal validation set and the external vocabulary datasets. The same partitions and stage views are used by all compared methods. Appendix A lists the standard category groups and implementation details.

Let $\mathcal{C}_t$ be the categories introduced at stage t, with T stages and $\mathcal{C}_{\text{all}} = \bigcup_{t=0}^{T-1} \mathcal{C}_t$. Groups are disjoint and satisfy

$$\max_{i,j} \big| \, |\mathcal{C}_i| - |\mathcal{C}_j| \, \big| \leq 1. \tag{1}$$

The standard $T = 10$ stream has 18 categories followed by nine groups of 17 on COCO, and ten groups of 15 on ADE. COCO categories are shuffled and partitioned; ADE uses frequency-ranked, back-and-forth allocation to avoid concentrating frequent categories in one stage. Seeds specify the partition and order; seeds 0, 1, and 2 are used for order diagnostics.

Why distribute acquisition across stages? Similar category counts keep later learning demands substantial, but do not imply equal pixel frequencies. For the complete source training annotations $\mathcal{D}_{\text{train}}$, let $y_x(p)$ denote the label at pixel p in image x, whose pixel grid is Ω_x. Define the global pixel mass of a category set by

$$N_{\text{pix}}(\mathcal{C}) = \sum_{(x,y_x)\in\mathcal{D}_{\text{train}}} \sum_{p\in\Omega_x} \mathbf{1}[y_x(p) \in \mathcal{C}], \qquad \rho_t = \frac{N_{\text{pix}}(\mathcal{C}_t)}{N_{\text{pix}}(\mathcal{C}_{\text{all}})}. \tag{2}$$

In ADE's standard category order, the first 100 categories account for 97.80% of valid pixels; subsequent groups in 100–10 contribute only 0.30–0.60% each. Even the three-stage 50–50 setting allocates 91.81% to its first group. Under COVSS, group shares range from 3.95–21.26% on COCO and 4.93–20.40% on ADE. These are statistics over the full source training set, not the current-label fraction within filtered stage images. They show that later identity supervision concerns substantial scene content while preserving natural frequency imbalance (Fig. 1).

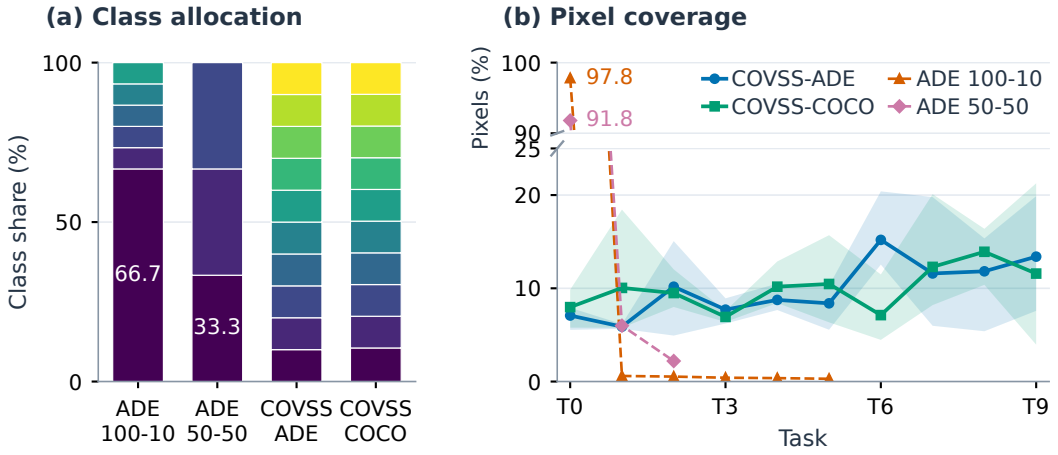


Figure 1: Category allocation and global training-pixel mass across stages. COVSS lines show means over three orders; shading shows their ranges. Panel (b) breaks the vertical axis between 25% and 90%. Pixel statistics use complete source annotations before stage-image selection. Near-equal category counts distribute acquisition without implying equal pixel mass.

This setting models systems that cannot begin with identity labels covering most scene content. Dense annotation is costly and can follow changing application priorities (Shin et al., 2021; Das et al., 2023). COVSS provides a controlled instance of such gradual supervision, rather than equating either source dataset with a fully realistic open environment. The benchmark's advantage is that acquisition and later interaction can be observed repeatedly, instead of being dominated by one initial training phase.

3.2 Training Information and Evaluation Conditions

For $(x, y_x) \in \mathcal{D}_{\text{train}}$, the stage-visible mask and training view are

$$y_t^{\text{vis}}(p) = \begin{cases} y_x(p), & y_x(p) \in \mathcal{C}_t, \\ \perp, & \text{otherwise}, \end{cases} \qquad \mathcal{D}_t = \{(x, y_t^{\text{vis}}) : (x, y_x) \in \mathcal{D}_{\text{train}},\ \exists p,\ y_x(p) \in \mathcal{C}_t\}. \tag{3}$$

The symbol $\perp$ means that the specific identity is unavailable, not that the underlying content is absent. This follows the stage-wise identity restriction of CSS (Cermelli et al., 2020; Douillard et al., 2021). The model starts from parameters θ_{-1} and updates sequentially to $\theta_0, \ldots, \theta_{T-1}$. It can use pretrained visual or vision–language weights and historical state formed from permitted information. Future category identities and external evaluation vocabularies are not training inputs. Replay, distillation, and regularization are allowed when their information and resource use are reported.

At checkpoint j, categories are old ($i < j$), current ($i = j$), or future ($i > j$) according to their introduction group $\mathcal{C}_i$. Future means not directly supervised in this stream; it does not exclude pretraining exposure. Internal evaluation uses the same full vocabulary $\mathcal{C}_{\text{all}}$ at every checkpoint. External evaluation uses VOC20, PC59, and PC459 on their respective validation sets (Everingham et al., 2010; Mottaghi et al., 2014), following OVS practice. Their semantic overlap with internal categories is retained: these are out-of-stream vocabulary tests, not uniformly tests of unseen concepts.

All stages are evaluated through one inherited model state, without selecting independently trained heads or experts by task identity. Fixed pretrained systems are separate references. These conditions isolate the behavior of continually updated segmentation systems without requiring a particular network architecture.

3.3 Trajectory-Based Evaluation

Let $\text{mIoU}(f; \mathcal{D}, \mathcal{C}, \mathcal{V})$ average IoU over scored categories $\mathcal{C}$ on dataset $\mathcal{D}$ with prediction vocabulary $\mathcal{V}$. For model $f_j = f_{\theta_j}$ and internal validation set $\mathcal{D}_{\text{val}}$, the category-group matrix is

$$M_{i,j} = \text{mIoU}(f_j; \mathcal{D}_{\text{val}}, \mathcal{C}_i, \mathcal{C}_{\text{all}}), \qquad 0 \le i, j < T. \tag{4}$$

Following task-response analyses in continual learning (Lopez-Paz & Ranzato, 2017; De Lange et al., 2022), rows follow category groups and columns follow checkpoints. Diagonals measure first acquisition; cells before and after the diagonal measure pre-supervision response and subsequent change. We summarize

$$\text{Final All} = \text{mIoU}(f_{T-1}; \mathcal{D}_{\text{val}}, \mathcal{C}_{\text{all}}, \mathcal{C}_{\text{all}}), \qquad \text{Diag} = \frac{1}{T}\sum_{i=0}^{T-1} M_{i,i},$$
$$\text{BWT} = \frac{1}{T-1}\sum_{i=0}^{T-2}(M_{i,T-1} - M_{i,i}), \qquad \text{Pre} = \frac{2}{T(T-1)}\sum_{i=1}^{T-1}\sum_{j=0}^{i-1} M_{i,j}. \tag{5}$$

Positive BWT indicates a net post-acquisition gain, not monotonically improving performance. Pre averages all pre-supervision cells and consequently weights later groups more heavily. External mIoU is recorded at every checkpoint with a fixed vocabulary. These measures distinguish what is initially learned, what changes later, and how learning affects queries outside direct supervision; full matrices retain variation lost in the summaries.

4 A Reference Implementation

The COVSS baseline builds on query-conditioned OVS and matching aggregation (Cho et al., 2024). Its role is to demonstrate a workable approach to the benchmark's partial supervision and changing semantic demands, rather than to define which algorithms the benchmark permits. Anonymous scene supervision preserves constraints outside the currently named categories; the Dynamic Alignment Field Aggregator (DAFA) organizes spatial and query relations; a shared decoder produces dense predictions (Fig. 2). The implementation uses no historical-sample memory, distillation, continual-learning-specific regularization, or pseudo-label updating.

Anonymous scene supervision. Non-current regions still provide evidence that their pixels do not belong to current categories. The baseline maps $\perp$ to a training-only complement query c_{cmp} named *unseen*. Its training vocabulary is $\mathcal{V}_t^{\text{tr}} = \mathcal{C}_t \cup \{c_{\text{cmp}}\}$, with one-hot targets $Y_{p,q}$ for each pixel p and query q. Given logit $z_{p,q}$ and sigmoid σ, the per-image loss is

$$\mathcal{L}_t = -\frac{\sum_{p\in\Omega_x}\sum_{q\in\mathcal{V}_t^{\text{tr}}}[Y_{p,q}\log\sigma(z_{p,q}) + (1-Y_{p,q})\log(1-\sigma(z_{p,q}))]}{|\Omega_x|\,|\mathcal{V}_t^{\text{tr}}|}. \tag{6}$$

Complement pixels supply negative-region constraints for current categories without identifying old or future content. This does not distinguish individual non-current classes. The complement query is removed at evaluation.

Reusable alignment and dense readout. A category name q is inserted into a text prompt $\tau(q)$ and encoded as a normalized prototype $\mathbf{e}_{t,q} = \text{norm}(g_{\text{text}}(\tau(q); \theta_t^{\text{text}}))$. The image encoder supplies token vectors $\mathbf{p}_{t,u} \in \mathbb{R}^d$ at token location u and multiscale guidance $\mathbf{G}_t$. Here d is the shared embedding dimension. DAFA forms the Raw Alignment Field and aggregates its learned projection:

$$A_{t,u,q}^{(0)} = \langle \text{norm}(\mathbf{p}_{t,u}), \mathbf{e}_{t,q}\rangle, \qquad \mathbf{H}_t^{(0)} = \text{Proj}(\mathbf{A}_t^{(0)}),$$
$$\overline{\mathbf{H}}_t^{(\ell)} = \text{SSE}_t^{(\ell)}(\mathbf{H}_t^{(\ell)}, \mathbf{G}_t), \qquad \mathbf{H}_t^{(\ell+1)} = \text{SQE}_t^{(\ell)}(\overline{\mathbf{H}}_t^{(\ell)}, \mathbf{E}_t). \tag{7}$$

For layers $\ell = 0, \ldots, L-1$, $\mathbf{E}_t$ stacks the query prototypes. Projection maps each scalar correlation to an h-dimensional hidden vector. Hidden states retain spatial, query, and hidden-channel axes. SSE propagates spatial evidence under visual guidance; SQE exchanges query information under text guidance. A decoder combines $\mathbf{H}_t^{(L)}$ with $\mathbf{G}_t$ and upsamples to pixel logits. These operations share parameters across queries instead of attaching parameters to a fixed category list. Temporal accumulation occurs through parameter inheritance, not attention over stored historical alignment fields. Appendix D gives the full analysis and cross-checkpoint decoder/DAFA combinations.

5 Experiments

5.1 Setup and Method Selection

Methods share the category order, stage views, 80k training iterations per task, and evaluation conditions. Released methods retain their official architecture, pretrained weights, and core training

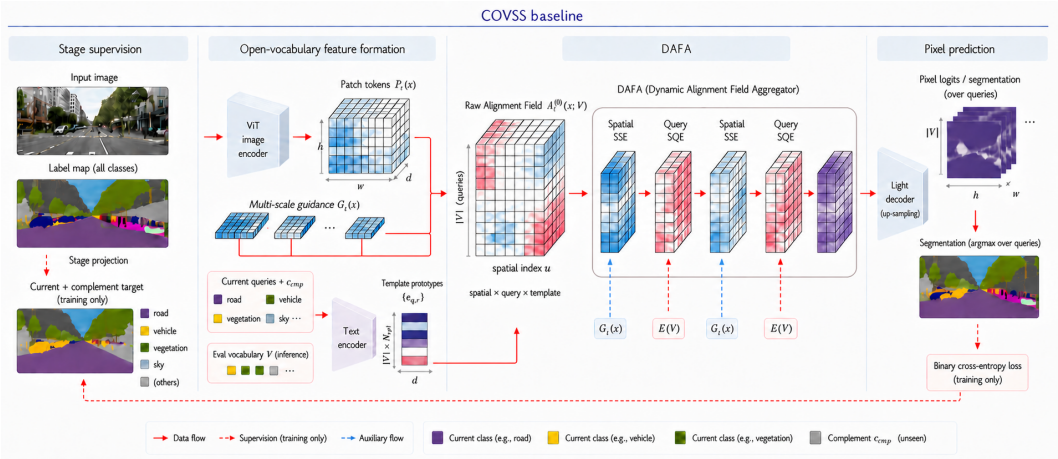


Figure 2: COVSS baseline. Current identities and an anonymous complement supervise one query-conditioned model. DAFA includes raw alignment-field construction and spatial/query aggregation; the decoder restores pixel predictions. Parameters are inherited across stages.

procedure; LBD is reproduced from its paper. CSS methods test incremental acquisition and retention, OVS methods test the adaptation of open prediction, and LBD tests language-assisted incremental learning. Methods whose adapters provide only pretrained evaluation are reported separately in Appendix C. This reflects compatibility with stage-wise mask supervision, not a claim that those methods cannot be trained. Expanding retrieval stores or task-specific experts represent different update settings.

We use the final checkpoint of each stage without selecting on evaluation results. Baseline main-table scores average ten stochastic training runs with category-order seed 0. Cohort matrices and pixel diagnostics use single reference runs under that order. Order experiments use three partitions/orders and 10k iterations per stage. Joint training uses the baseline architecture with complete labels for 80k iterations as an offline reference. Appendix A provides configurations and run-level statistics; Appendix B provides detailed method trajectories.

5.2 Acquisition, Subsequent Change, and Open Prediction

Table 2 separates first-supervised performance from later change. MiB, PLOP, and SDR combine low Diag with negative BWT, while SimCIS reaches only 0.20/0.54 Final All on COCO/ADE. These outcomes extend earlier observations of weak new-category learning (Cermelli et al., 2020; Douillard et al., 2021; Michieli & Zanuttigh, 2021): when acquisition remains substantial at every stage, preserving old predictions alone is insufficient. Yet CSS does not fail uniformly. ALIFE attains the highest ADE Diag (36.25) but ends at 31.94; its COCO BWT is -10.88. Both acquisition and later retention shape the endpoint.

Open-vocabulary structure also does not guarantee improvement. CAT-Seg and SAN deteriorate; SED has local recoveries but slightly negative BWT. LBD shows limited recovery with language priors, yet low Diag and non-positive BWT constrain its final performance. HyperCLIP achieves $+3.04/+2.87$ BWT together with external querying. Thus improvement through later tasks is not unique to the proposed reference implementation, nor a universal consequence of adding semantic priors.

The baseline achieves the highest Final All and external scores among continually trained comparators. On COCO, its Diag exceeds SED by only 0.80 points, but Final All by 7.09 points; much of the advantage develops after acquisition. On ADE, ALIFE acquires more strongly, while the baseline's subsequent gains yield a higher endpoint. Baseline BWT is $+6.17 \pm 1.10$ and $+5.48 \pm 1.05$ across ten runs (sample SD), positive in every run; a two-sided exact sign test against zero with Holm correction across datasets gives $p_{\text{adj}} = 0.0039$. This tests within-method net change, not pairwise significance. Joint training remains 1.90/4.84 points higher in Final All.

Table 2: Continually trained methods under COVSS (80k iterations per stage). Scores are mIoU (%); BWT is in points. Bold: best continual result; †: baseline BWT differs from zero (Sec. 5.2). N/A denotes an inapplicable measure. Baseline entries average ten runs; joint training is an offline reference.

		COCO-Stuff171						
Family	Method	Final All	Diag	BWT	Pre	VOC20	PC59	PC459
CSS	MiB	7.46	16.00	−9.46	N/A	N/A	N/A	N/A
CSS	PLOP	3.82	6.51	−2.97	N/A	N/A	N/A	N/A
CSS	SDR	9.45	13.66	−4.63	N/A	N/A	N/A	N/A
CSS	SimCIS	0.20	1.90	−1.88	N/A	N/A	N/A	N/A
CSS	DKD	8.90	19.17	−11.38	0.00	N/A	N/A	N/A
CSS	ALIFE	23.65	33.51	−10.88	0.00	N/A	N/A	N/A
Cross-over	LBD	8.44	8.53	−0.15	0.00	N/A	N/A	N/A
OVS	CAT-Seg	1.41	14.79	−14.85	11.21	20.93	3.92	1.97
OVS	H-CLIP	33.04	30.35	+3.04	24.85	92.05	45.37	18.13
OVS	SAN	17.03	26.89	−10.96	11.66	53.70	21.97	9.49
OVS	SED	34.31	35.09	−0.80	19.62	79.39	41.47	17.59
COVSS	**COVSS baseline**	**41.40**	**35.89**	**+6.17**†	**26.85**	**93.79**	**51.76**	**22.27**
Offline	Joint training	43.30	N/A	N/A	N/A	94.55	54.97	23.90
		ADE150						
Family	Method	Final All	Diag	BWT	Pre	VOC20	PC59	PC459
CSS	MiB	8.99	18.26	−10.30	N/A	N/A	N/A	N/A
CSS	PLOP	4.95	8.07	−3.47	N/A	N/A	N/A	N/A
CSS	SDR	10.42	15.35	−5.78	N/A	N/A	N/A	N/A
CSS	SimCIS	0.54	2.36	−2.11	N/A	N/A	N/A	N/A
CSS	DKD	26.24	28.68	−2.71	0.00	N/A	N/A	N/A
CSS	ALIFE	31.94	**36.25**	−4.79	0.00	N/A	N/A	N/A
Cross-over	LBD	4.89	7.29	−2.67	0.00	N/A	N/A	N/A
OVS	CAT-Seg	0.85	10.74	−15.57	11.36	18.90	2.84	1.48
OVS	H-CLIP	25.00	22.41	+2.87	17.37	87.77	35.44	13.46
OVS	SAN	7.71	19.22	−12.78	6.43	64.47	18.25	6.33
OVS	SED	29.82	30.41	−0.83	18.56	78.84	39.80	16.95
COVSS	**COVSS baseline**	**36.72**	31.79	**+5.48**†	**21.32**	**89.85**	**45.34**	**20.75**
Offline	Joint training	41.56	N/A	N/A	N/A	92.42	51.14	23.89

Table 3 shows final gains over first supervision for eight COCO groups and nine ADE groups out of ten, alongside local declines and recoveries. All six external trajectories end above their first checkpoint (Fig. 3a,b), although their peaks differ. Appendix B provides additional matrices and curves.

5.3 Sources of Subsequent Gains

Later knowledge can correct an old category's excessive predictions rather than simply strengthen its response. TP, FP, and FN denote true-positive, false-positive, and false-negative pixel counts. With fixed ground truth, let $G_c = TP_c + FN_c$, $D_c = G_c + FP_c$, $a_c = TP'_c - TP_c$, and $r_c = FP_c - FP'_c$ for old category c; primes denote the later checkpoint. Direct algebra gives

$$\Delta \,\mathrm{IoU}_c = \frac{a_c D_c + TP_c r_c}{D_c(D_c - r_c)}. \tag{8}$$

For positive denominators, IoU can increase even with $a_c \leq 0$ when false-positive reduction compensates for lost correct coverage. Among 834 IoU-increasing old-class transitions on COCO and ADE, 404 (48.44%) have no recall increase, which requires an absolute FP reduction under the fixed-pixel evaluation conditions (Fig. 3c).

Fixed-image tracking further resolves the changes. At COCO $T_6 \rightarrow T_7$, a frisbee case loses 55 true-positive pixels but removes 5,316 false positives; 2,795 are correctly reassigned to current-class

Table 3: Reference category-group trajectories of the COVSS baseline (80k per stage, order seed 0). Bold: first supervision; arrows: change from the preceding checkpoint. These are single runs, not element-wise averages of the ten-run main table.

COCO

	T_0	T_1	T_2	T_3	T_4	T_5	T_6	T_7	T_8	T_9
C_0	**23.3**	26.3↑	29.1↑	28.7↓	32.3↑	33.3↑	31.7↓	34.1↑	34.9↑	35.1↑
C_1	16.8	**22.8**↑	23.6↑	23.0↓	25.7↑	27.0↑	25.5↓	28.6↑	31.1↑	30.6↓
C_2	26.7	30.5↑	**37.1**↑	38.0↑	39.4↑	39.8↑	41.4↑	45.2↑	49.2↑	49.4↑
C_3	27.9	30.6↑	32.1↑	**38.4**↑	40.2↑	41.4↑	40.7↓	42.1↑	44.7↑	44.5↓
C_4	21.8	26.6↑	28.1↑	28.3↑	**36.4**↑	37.5↑	36.5↓	39.8↑	42.0↑	41.6↓
C_5	17.8	22.4↑	22.5↑	25.7↑	28.5↑	**34.6**↑	35.3↑	37.2↑	37.6↑	38.6↑
C_6	19.1	22.0↑	23.5↑	23.9↑	26.6↑	26.9↑	**32.9**↑	37.7↑	39.5↑	39.6↑
C_7	26.0	29.2↑	32.2↑	34.3↑	34.9↑	36.4↑	36.3↓	**46.0**↑	48.0↑	48.3↑
C_8	20.4	22.6↑	23.5↑	25.0↑	27.0↑	30.9↑	27.3↓	30.4↑	**48.7**↑	47.8↓
C_9	19.3	21.8↑	25.0↑	25.7↑	27.6↑	28.0↑	28.1↑	32.8↑	35.1↑	**38.8**↑

ADE

	T_0	T_1	T_2	T_3	T_4	T_5	T_6	T_7	T_8	T_9
C_0	**23.3**	22.7↓	23.6↑	27.9↑	27.3↓	27.1↓	28.1↑	29.6↑	31.1↑	32.6↑
C_1	15.6	**25.6**↑	24.7↓	28.2↑	28.0↓	26.5↓	26.9↑	29.9↑	30.4↑	31.7↑
C_2	9.2	13.0↑	**21.6**↑	25.0↑	26.6↑	26.2↓	26.8↑	26.0↓	25.5↓	26.5↑
C_3	14.5	19.4↑	19.7↑	**36.1**↑	35.8↓	34.2↓	36.0↑	37.6↑	36.5↓	40.4↑
C_4	12.5	16.2↑	19.0↑	22.0↑	**28.5**↑	30.5↑	31.2↑	34.5↑	31.8↓	35.2↑
C_5	16.3	17.8↑	19.2↑	19.3↑	22.4↑	**29.4**↑	30.3↑	32.3↑	33.0↑	37.0↑
C_6	21.5	26.1↑	29.0↑	32.3↑	30.7↓	31.4↑	**45.3**↑	50.2↑	49.3↓	48.7↓
C_7	20.0	20.1↑	22.0↑	25.6↑	25.2↓	22.9↓	23.3↑	**35.1**↑	35.6↑	38.2↑
C_8	18.8	23.0↑	24.4↑	28.4↑	28.0↓	26.7↓	27.6↑	27.2↓	**40.3**↑	44.1↑
C_9	12.8	12.6↓	16.1↑	18.8↑	19.0↑	19.2↑	21.7↑	25.4↑	23.4↓	**32.8**↑

person and 2,020 to future-class tree. Holding the vocabulary fixed, replacing only the old-class logit worsens IoU, whereas replacing competing logits produces the gain. These cases identify a mechanism of response correction, not a guarantee that every update or class benefits. Appendix D includes counterexamples, output interventions, and exact evaluation scopes. Figure 3d summarizes both selected cases.

This behavior is consistent with an OVS model reusing representations across category queries. Complement supervision retains negative-region constraints, while spatial/query aggregation and an adaptive decoder can reorganize competing predictions. Component pairing shows that the earliest decoder with final DAFA retains 90.6%/91.6% of final internal performance on COCO/ADE, while later decoder adaptation still helps. These findings support compatibility across stages without attributing the entire gain to one module.

5.4 Robustness to Task Organization

At 10k iterations per stage, all three class orders yield positive BWT: 4.21–7.32 on COCO and 2.74–6.87 on ADE. Initial acquisition and final performance remain order dependent. In five-, ten-, and twenty-task COCO streams, final performance exceeds the first checkpoint in every case, while Final All is 32.75, 33.65, and 30.13. Finer tasks do not become easier merely because the total training budget increases. Appendix E reports full trajectories, making clear the different total budgets and the coexistence of larger subsequent gains with external-vocabulary declines.

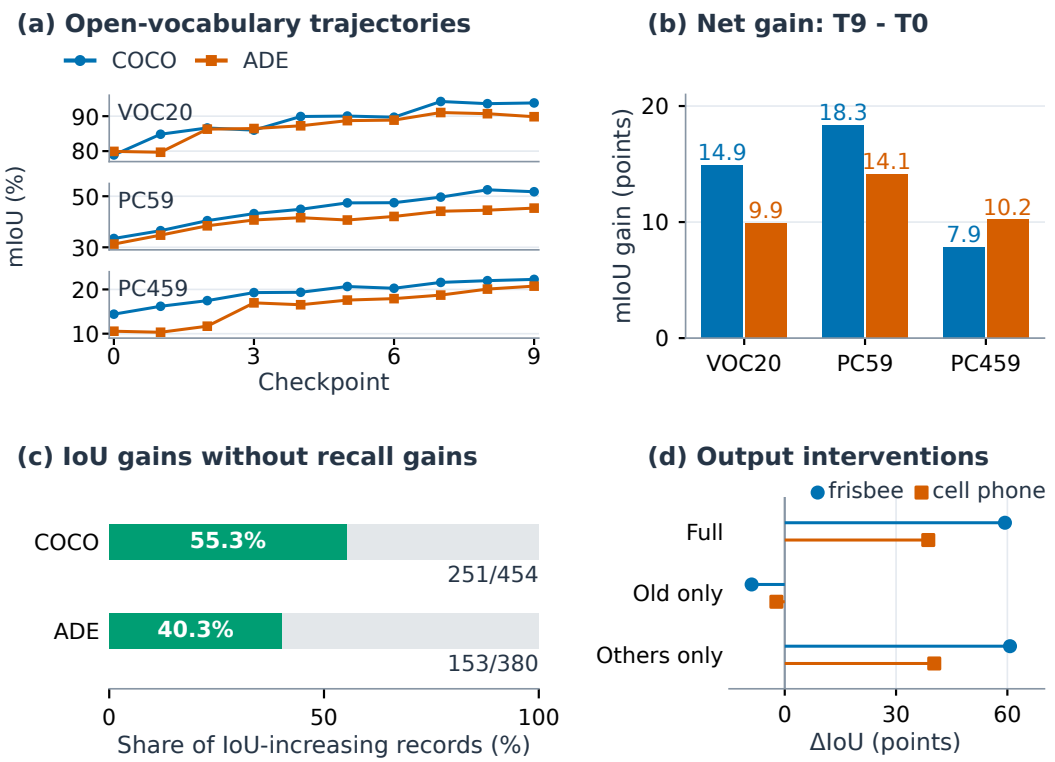


Figure 3: Performance development and diagnostics of the COVSS baseline (80k per task; single reference runs). (a) External mIoU on separate vertical scales. (b) Final-minus-initial gains. (c) IoU-increasing old-class records without recall gains; counts need not represent distinct classes. (d) Two fixed COCO cases at $T_6 \to T_7$, relative to T_6 under the same 171-class vocabulary. Full replaces all logits; Old only replaces the target-class logit; Others only replaces the remaining logits. Full diagnostics are in Appendix D.

6 Discussion and Conclusion

COVSS contributes a benchmark for studying scene understanding that develops under incremental, partial identity supervision. Its construction reuses established annotations to make the learning sequence, information boundary, and internal/external evaluation comparable across methods. The reference implementation demonstrates one effective approach; the contrasting CSS and OVS trajectories establish a broader research question than improving a single architecture.

The results suggest that continual learning should preserve correct discrimination while allowing later knowledge to revise earlier errors. Shared open-vocabulary prediction provides one path for such revision, but neither an open interface nor a language prior ensures beneficial updates. The algebraic analysis identifies how IoU gains can occur; it does not prove which parameter updates will produce them. Stronger current-stage acquisition without losing subsequent gains is an immediate direction for the baseline.

The benchmark uses controlled class increments and existing finite datasets. Unequal or variable-length increments, conventional CSS task organizations, changing environments, and evolving annotation priorities are important extensions. External vocabulary tests also do not exhaust open-world behavior or guarantee safety in embodied applications. Studying these richer conditions can determine when the observed compatibility between learning, retention, and open prediction extends to longer-running visual systems.

AI Use Statement

AI tools assisted with structural and language editing, LaTeX conversion, and scripts for result organization and layout checks. Authors remain responsible for experiments and interpretation.

Reproducibility Statement

The separate supplementary material details class partitions, stage-selection and identity rules, configurations, metrics, run-level results, and trajectories, distinguishing training randomness, category order, and single-run diagnostics.

References

Donghyeon Baek, Youngmin Oh, Sanghoon Lee, Junghyup Lee, and Bumsub Ham. Decomposed knowledge distillation for class-incremental semantic segmentation. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://arxiv.org/abs/2210.05941.

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*, 2018.

Fabio Cermelli, Massimiliano Mancini, Samuel Rota Bulò, Elisa Ricci, and Barbara Caputo. Modeling the background for incremental learning in semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

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*, 2024.

Anurag Das, Yongqin Xian, Yang He, Zeynep Akata, and Bernt Schiele. Urban scene semantic segmentation with low-cost coarse annotation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023.

Matthias De Lange, Rahaf Aljundi, Marc Masana, Sarah Parisot, Xu Jia, Aleš Leonardis, Gregory Slabaugh, and Tinne Tuytelaars. A continual learning survey: Defying forgetting in classification tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(7):3366–3385, 2022.

Zheng Ding, Jieke Wang, and Zhuowen Tu. Open-vocabulary universal image segmentation with MaskCLIP. *arXiv preprint arXiv:2208.08984*, 2022.

Arthur Douillard, Yifu Chen, Arnaud Dapogny, and Matthieu Cord. PLOP: Learning without forgetting for continual semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

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.

Zhongrui Gui, Shuyang Sun, Runjia Li, Jianhao Yuan, Zhaochong An, Karsten Roth, Ameya Prabhu, and Philip H. S. Torr. kNN-CLIP: Retrieval enables training-free segmentation on continually expanding large vocabularies. *Transactions on Machine Learning Research*, 2024.

Dongjun Hwang, Yejin Kim, Minyoung Lee, Seong Joon Oh, and Junsuk Choe. OVS meets continual learning: Towards sustainable open-vocabulary segmentation. In *Advances in Neural Information Processing Systems*, 2025.

Boyi Li, Kilian Q. Weinberger, Serge Belongie, Vladlen Koltun, and René Ranftl. Language-driven semantic segmentation. In *International Conference on Learning Representations*, 2022.

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*, 2023.

Zihan Lin, Zilei Wang, and Xu Wang. Towards continual universal segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 29417–29427, 2025.

David Lopez-Paz and Marc'Aurelio Ranzato. Gradient episodic memory for continual learning. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Umberto Michieli and Pietro Zanuttigh. Continual semantic segmentation via repulsion-attraction of sparse and disentangled latent representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

Roozbeh Mottaghi, Xianjie Chen, Xiaobai Liu, Nam-Gyu Cho, Seong-Whan Lee, Sanja Fidler, Raquel Urtasun, and Alan Yuille. The role of context for object detection and semantic segmentation in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014.

Youngmin Oh, Donghyeon Baek, and Bumsub Ham. ALIFE: Adaptive logit regularizer and feature replay for incremental semantic segmentation. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://arxiv.org/abs/2210.06816.

Zelin Peng, Zhengqin Xu, Zhilin Zeng, Yu Huang, Yaoming Wang, and Wei Shen. Parameter-efficient fine-tuning in hyperspherical space for open-vocabulary semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning*, 2021.

Gyungin Shin, Weidi Xie, and Samuel Albanie. All you need are a few pixels: Semantic segmentation with PixelPick. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 2021.

Gido M. van de Ven, Tinne Tuytelaars, and Andreas S. Tolias. Three types of incremental learning. *Nature Machine Intelligence*, 4:1185–1197, 2022.

Liyuan Wang, Xingxing Zhang, Hang Su, and Jun Zhu. A comprehensive survey of continual learning: Theory, method and application. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(8):5362–5383, 2024.

Ruitao Wu, Yifan Zhao, and Jia Li. Learning yourself: Class-incremental semantic segmentation with language-inspired bootstrapped disentanglement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025.

Bin Xie, Jiale Cao, Jin Xie, Fahad Shahbaz Khan, and Yanwei Pang. SED: A simple encoder-decoder for open-vocabulary semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

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*, 2023.

Hongmei Yin, Tingliang Feng, Fan Lyu, Fanhua Shang, Hongying Liu, Wei Feng, and Liang Wan. Beyond background shift: Rethinking instance replay in continual semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9839–9848, 2025.

Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ADE20K dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

Yuchen Zhu, Cheng Shi, Dingyou Wang, Jiajin Tang, Zhengxuan Wei, Yu Wu, Guanbin Li, and Sibei Yang. Rethinking query-based transformer for continual image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.