

Anchor the Head: Alignment Drift in Open-Vocabulary Segmentation

Xuan-Bach Le

Ho Chi Minh City University of Technology (HCMUT), VNU-HCM, Vietnam

lexuanbach@hcmut.edu.vn

Abstract

Open-vocabulary segmenters use text to recognize classes without segmentation training labels. A small head can adapt their predictions to labeled (seen) classes, but it can also suppress held-out (novel) classes even when the backbone remains frozen. We study this tension on region-based and dense vision-language segmenters under matched data and supervision. On ADE20K, an unconstrained head with class-dependent outputs reduces novel-class intersection over union (IoU) from 15.1 to 0.1. Training the head to stay close to frozen class probabilities restores novel-class IoU to 15.3 and raises mean IoU (mIoU) from 14.5 to 15.8. To characterize this change, we measure alignment drift, the mean cosine distance between corrected and frozen probability vectors. Across 38 class-specific settings (32 distinct corrections) in 13 families, drift tracks novel-class retention (Pearson correlation $r=-0.90$). Corrections that apply a shared rule across classes retain novel-class IoU throughout their observed drift range, establishing that drift must be interpreted within the correction family. IoU recovery leaves a coverage gap. Anchored heads predict novel classes over a smaller part of the image, while improved confidence calibration alone can conceal novel-class loss. Dividing all class scores by a shared temperature gives the lowest expected calibration error among the tested in-domain corrections retaining novel-class IoU (27.2 $\rightarrow$ 8.7). Our results support anchoring class-specific heads to frozen predictions and evaluating novel-class accuracy, prediction coverage and confidence calibration together.¹

1. Introduction

Open-vocabulary (OV) semantic segmentation assigns image regions to classes described by text, including classes absent from segmentation training. The Segment Anything Model (SAM) [19] proposes regions without semantic labels. Contrastive Language–Image Pretraining (CLIP) [30] then compares each region with text prototypes, the vector

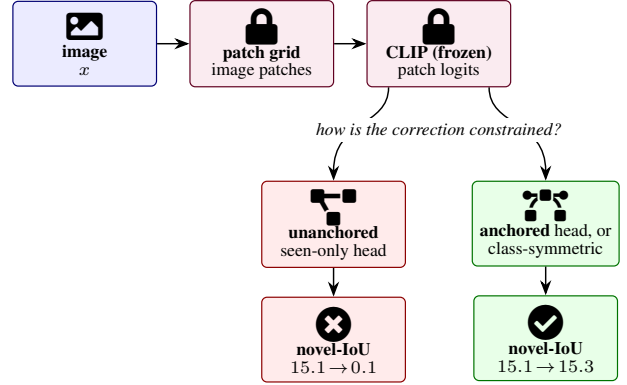


Figure 1. Correction families on frozen dense CLIP. Unanchored heads (red) reduce novel-class IoU, while anchored heads and class-symmetric controls (green) retain it. Values follow Table 5.

representations of class names. Dense-CLIP variants score patches directly. We can adapt these frozen models with a small head trained on labeled *seen* classes. The goal is to retain recognition of held-out *novel* classes.

A head can change which class wins a prediction even when the image representation stays fixed. We ask which constraints preserve novel-class accuracy during adaptation. Within each comparison, we fix the backbone, cached logits (unnormalized class scores), class splits and supervision budget. Our ADE20K splits hold out 30 of 150 classes [48]. We compare class-specific corrections with class-dependent outputs against class-symmetric corrections that apply a shared rule across classes. Anchoring constrains their departure from frozen predictions.

Figure 1 illustrates the resulting distinction. Both branches share frozen patch logits, highlighting how anchoring or class symmetry preserves novel-class predictions after correction. An unconstrained head reduces novel-class IoU from 15.1 to 0.1. The loss persists across our data and objective ablations. We characterize the departure from the frozen predictor through *alignment drift*, measured between corrected and frozen class-probability vectors. Across 38 settings in 13 families, drift correlates with retention at $r=-0.90$, with a family-level correlation of -0.65 (Fig-

¹Code and supplementary material:

<https://github.com/lexuanbach/anchor-the-head>.

ure 4). This association identifies a useful diagnostic within the class-specific family.

Anchoring provides an effective intervention. Distillation, which trains the head to match frozen outputs, restores novel-class IoU to 15.3 at penalty weight $\lambda=30$ and raises mIoU from 14.5 to 15.8. Bounding each score correction by a trust-region radius $\tau=0.5$ also recovers accuracy. Alongside interpolation toward frozen weights [39], these sweeps trace the adaptation–retention trade-off (Figure 3). Class-symmetric controls retain novel-class IoU across their observed drift range, showing why drift alone cannot diagnose loss across correction types. These constraints connect our study to alignment-preserving OV architectures [45, 46].

We contribute a matched study of fitting, class symmetry and anchoring (Table 5), extending it to the released LPOSS label-propagation method [33], a head on its output (Table 6), and cross-dataset transfer. We assess both adaptation gains and retained capabilities. Recovering novel-class IoU can leave a gap in the image area assigned to novel classes, and better confidence calibration can coexist with lost accuracy. We therefore evaluate accuracy, coverage and calibration separately when choosing a correction.

We review related work (Section 2), define the problem (Section 3), and introduce correction families (Section 4). Experiments and deployment guidance follow in Section 5, with limitations in Section 6 and conclusions in Section 7.

2. Related Work

Open-vocabulary segmentation from frozen foundation models. Many open-vocabulary segmenters combine a class-agnostic proposer with a vision-language model. SAM [19, 31] or mask transformers [5] supply regions, and CLIP [30] labels them against text. Training can align pixels and text [13, 23, 43], adapt mask classification [8, 16, 25], or add side networks to frozen CLIP [45]. Other architectures use diffusion features [44] or aggregate cost volumes [6, 41].

Frozen-backbone systems refine the region-to-text interface in several ways. FC-CLIP [46] scores masks with convolutional CLIP, Mask-Adapter [24] aligns proposal masks to CLIP, and LPOSS [33] propagates labels over patch-and-pixel graphs. We use released LPOSS propagation as both a control and a base for our head audit (Table 6). CASS [18], CorrCLIP [47] and PEARL [28] refine the base predictor. ConceptBank [29] recalibrates SAM3 concept prompts using target-domain statistics without learned parameters. Training-free dense methods also modify CLIP attention or features directly [2, 3, 10, 21, 22, 37, 40, 49]. Our dense experiments draw on this family while keeping the base fixed within each comparison.

Seen-class bias in trained open-vocabulary segmenters. SCAN [27] calibrates proposal embeddings with CLIP’s

semantic prior to counter training-class bias. EBSeg [32] adaptively balances image embeddings for training and new classes and aligns their semantic structure with CLIP text embeddings. These architectures establish the value of preserving pretrained semantic information during adaptation.

We extend this architectural motivation into measurable design trade-offs. By fixing the base and supervision, we isolate class specificity, fitting and anchoring across correction families. Our sweeps quantify the anchoring needed for recovery and connect it to drift from frozen predictions. Joint coverage and calibration measurements then expose what IoU recovery leaves unresolved.

Graph smoothing and label propagation. Pairwise smoothing connects fully connected conditional random fields (CRFs) [20], correspondence-based consistency [34], and propagation over feature graphs [33]. We use anchored diffusion and generic graph-Laplacian smoothing as class-symmetric controls. Their comparable retention tests whether the behavior extends across operators.

Our ADE20K [48] and PASCAL VOC [11] evaluations follow the broader dense-prediction literature [4, 7, 26, 42].

Calibration and selective prediction. Temperature scaling divides logits by a positive scalar T . Per-class temperatures T_c and vector scaling introduce class-dependent rescaling, with vector scaling also adding a bias. These are standard post-hoc calibration families [15]. Global temperature also transfers across label sets and distribution shifts in vision-language models [35]. Dense-prediction alternatives include local temperature scaling [9] and selective scaling [36], while open-vocabulary calibration explicitly addresses confidence errors on novel classes [38]. We evaluate calibration error, selective-prediction risk (error among predictions retained by confidence) [1, 12, 17], and novel-class retention separately. Section 5 shows that improved calibration or risk ranking can coexist with a large loss of retention. Weight-space interpolation [39] provides a complementary way to control that trade-off on the correction head (Table S7).

3. Problem Setup

We study corrections to frozen OV segmenters. For an image x , a proposer emits N class-agnostic masks $\{m_i\}_{i=1}^N$. A vision-language model embeds region i as z_i and class c as a text prototype P_c . The vocabulary $\mathcal{C} = \mathcal{C}_s \cup \mathcal{C}_n$ contains seen classes $\mathcal{C}_s$ and held-out novel classes $\mathcal{C}_n$. Dense-CLIP stacks use patches in place of masks. We fit corrections on seen labels and evaluate the full vocabulary, keeping the backbone frozen and withholding novel-class training labels.

We separate segmentation accuracy from confidence quality. Intersection over union (IoU) measures the overlap of predicted and ground-truth regions for a class. We report its mean across classes (mIoU), its mean across novel classes (novel-class IoU), and the harmonic mean of seen and novel means (hIoU). Expected calibration error (ECE) measures the gap between prediction confidence and observed accuracy. Our central question is which constraints retain novel-class accuracy as the correction adapts to seen classes.

4. Correction Families

We separate whether a correction is fitted from how it acts across classes. A *class-symmetric* operator applies a shared transformation to each class channel. A *class-specific* head can alter relative scores through class-dependent parameters or outputs. To expose this distinction, we compare parameter-free diffusion with a trained decoder on a common proposal graph. Section 5 extends the comparison to calibrators and dense heads.

Proposal representation. We use frozen SAM [19] with a base Vision Transformer (ViT-B) to generate masks. Each mask defines a bounding-box crop, with background pixels set to the CLIP mean color. Frozen CLIP [30] with a large ViT and 14-pixel patches (ViT-L/14) encodes each crop as a unit vector $z_i \in \mathbb{R}^D$, where D is the embedding dimension. We form class prototypes by averaging text embeddings over prompt templates and stack them as rows of $P \in \mathbb{R}^{|C| \times D}$. Independent scoring assigns logits $\ell_i^0 = s(z_i P^\top)$ at fixed scale s and selects the highest-scoring class, $\arg \max_c \ell_{ic}^0$. This training-free Proposal+CLIP rule is our baseline.

Proposal graph. Overlapping masks provide complementary evidence about the same image. We connect masks that overlap or share a boundary with weights $w_{ij} = \text{IoU}_{ij} + b_{ij} + [z_i \cdot z_j]_+$. Here IoU_{ij} measures mask overlap, b_{ij} is the boundary-overlap fraction, and $[u]_+ = \max(u, 0)$ retains positive embedding similarity. We normalize each row of the weight matrix to obtain A , which averages neighboring evidence.

Class-symmetric control: anchored graph diffusion. We combine neighboring scores with the original frozen scores at each iteration. Stacking proposal logits into ℓ gives

$$\ell^{(t+1)} = (1 - \alpha) \ell^0 + \alpha A \ell^{(t)}, \quad \ell^{(0)} = \ell^0, \quad (1)$$

where t indexes iterations, $\alpha \in [0, 1]$ weights neighboring evidence, and k is the number of steps. We use $\ell^{(k)}$ for inference. The update learns no parameters, and $\alpha=0$ recovers the baseline. Reinserting frozen scores anchors the

update, while applying the same A to every class makes it class-symmetric. We therefore control both evidence sharing and departure from the frozen predictor.

Class-specific correction: a trained graph decoder.

Our decoder learns to exchange evidence on the same graph. The projection ϕ maps concatenated region and geometric features $[z_i; g_i]$ into node state h_i^0 . Here g_i contains mask area, centroid and bounding-box dimensions. Edge-conditioned messages update the states through gated recurrent units (GRUs). From the final state h_i , a learned map ρ produces $\tilde{z}_i = \text{normalize}(z_i + \epsilon \tanh(\rho(h_i)))$. Normalization restores unit length, and $\epsilon=0.2$ bounds each residual component. We initialize the residual output at zero and regularize it during training. The decoder also learns refined text prototypes $\tilde{P}$, per-region temperatures and scores for ignored regions. Its few million parameters train on majority labels of seen-class masks. This comparison tests whether a bounded learned correction retains the behavior of the frozen representation.

Inference. Softmax normalizes exponentiated logits into class probabilities. We take the maximum probability as confidence. We omit decoder proposals whose ignored-region probability exceeds 0.5. The highest-confidence remaining proposal covering a pixel supplies its label. We compute ECE before painting predictions onto pixels, once per labeled proposal against its majority ground-truth label.

5. Experiments

Setup. We evaluate on ADE20K-150 [48] validation. Our masked-crop pipeline paints and scores predictions at 256×256 , retaining images with at least two SAM proposals. Table 2 and Supp. Tables S2 and S4 use the original cache, whose image count was not recorded (baseline mIoU 8.9). An independent rebuild on different hardware contains 1623 of the 2000 validation images and reaches 10.2 baseline mIoU. SAM mask generation varies across cache builds. Dense CLIP uses a base ViT with 16-pixel patches (ViT-B/16), has no proposer and evaluates all 2000 images.

We hold out 30 ADE classes as novel and retain 120 as seen in each split. Masked-crop splits use seeds 0–4. Dense splits use five sequential permutations from one Random-State stream seeded at 0, and both sets are released. We report mean±standard deviation across splits. Bootstrap confidence intervals (CIs) resample images at seed 0. Calibration and encoder comparisons use 347 images over five splits, with the edge ablation restricted to seed 0. Separate dense-head runs differ by up to 0.1 in repeated rounded metrics (Tables 5–6 and S6–S8).

We report hIoU as $2I_s I_n / (I_s + I_n)$ for seen/novel mIoU I_s, I_n . ECE averages absolute confidence–accuracy gaps

across 15 equal-width bins, weighted by the number of labeled proposals or patches. We report overall, seen and novel ECE separately, with IoU and ECE expressed as percentages. The *collapse index* $C = \text{IoU}_{\text{novel}}^{\text{corrected}} / \text{IoU}_{\text{novel}}^{\text{frozen}}$ measures retention relative to the frozen baseline, with $C=1$ indicating matched IoU. We reserve *novel-class collapse* for reductions to near zero and describe other settings by their IoU or C .

We define *alignment drift* on the common set $\mathcal{V}$ of labeled evaluation patches, excluding ignored locations. Let p_i and p_i^0 be the corrected and frozen probability vectors over the full class vocabulary. Then

$$d = \frac{100}{|\mathcal{V}|} \sum_{i \in \mathcal{V}} \left(1 - \frac{p_i^\top p_i^0}{\|p_i\|_2 \|p_i^0\|_2} \right). \quad (2)$$

Identical distributions give $d=0$. The dot product and Euclidean norms compare distribution direction, with larger d indicating greater change. We use the same valid patch support as the dense metrics, without class identities or novel-class labels.

Label-dependent choices use seen-class data. The 347-image calibration study fits calibrators on seen-class masks from 300 training images. Diffusion uses the fixed defaults $\alpha=0.4$ and $k=2$. We cache proposals from a 16-point SAM grid with embeddings, edges and labels. Masked-crop corrections run on CPU, and the decoder trains for 12 Adam epochs on one GPU with fixed seeds. Table 1 identifies the protocol for each main-paper result table. We compare corrections within each protocol, since changes in cache or scoring resolution also change the baseline.

Primary comparison on ADE20K masked crops. On matched masked-crop scores, graph diffusion changes mIoU by +0.1 relative to independent scoring, within the bootstrap interval (Table 2). Novel-class IoU and hIoU increase on this cache, while the five-split calibration comparison includes zero gain in its interval. ECE falls from 14.6 to 5.9. In contrast, the trained graph decoder reaches 7.1 mIoU and 1.1 novel-class IoU. Our independent cache rebuild confirms this trade-off at 10.2 baseline mIoU, with decoder novel-class IoU falling from 9.5 to 2.5.

Seen/novel trade-off for the trained decoder. We probe this loss by removing decoder components (Supp. Table S4). Three graph variants maintain seen-class IoU between 8.6 and 9.8, against the baseline’s 8.7. Their novel-class IoU falls from 9.4 to at most 1.5, and ECE increases. Bounding and regularizing the residual therefore leaves substantial novel-class loss in this configuration. The sensitivity to altered frozen representations also appears in zero-shot robustness under shift [39] and continual open-vocabulary segmentation [14]. Removing prototype refine-

ment lowers seen-class IoU to 5.5 and novel-class IoU to 0.0.

Calibration baselines and novel-class retention (Supp. B). We next isolate calibration on fixed logits. A global temperature gives the lowest overall ECE on identical frozen logits over five splits, reducing it from 14.6 to 3.5, compared with diffusion’s 5.9. Their novel-class ECE values are statistically indistinguishable (paired t statistic 0.6, four degrees of freedom). Per-class calibrators, however, reduce novel-class IoU and raise novel ECE to 27.9 and 34.0. Class-symmetric diffusion retains novel-class IoU. Table S1 and Figure S1 show why aggregate calibration is insufficient. A lower overall ECE can coexist with larger confidence errors on novel classes.

Diffusion, edge and encoder ablations (Supp. C, E). Our graph and crop ablations isolate the available gains. Moderate diffusion weights ($\alpha=0.4$ – 0.6) retain or improve novel-class IoU. Extreme settings lose at most 0.7. IoU-only edges yield 9.57 against the seed-0 baseline of 9.75, and context-padded crops raise mIoU from 8.9 to 9.2. These ablations delimit the gains available from changing the graph or region crop alone.

Replication on a dense-CLIP base. We remove proposal generation with frozen CLIP ViT-B/16 and SCLIP-style Correlative Self-Attention, which uses patch affinities for dense prediction. Table 3 evaluates five splits at patch resolution, where the baseline reaches 14.5 mIoU. Our dense head is a 170k-parameter multilayer perceptron (MLP) adding a logit residual, distinct from the graph decoder. Per-class temperature reduces novel-class IoU from 15.1 to 0.0, and the MLP reduces it to 0.1. Global temperature gives the lowest calibration error (27.2→8.7), while diffusion attains the highest novel-class IoU (15.7). Diffusion’s masked-crop calibration gain disappears on this base, separating its retention behavior from its calibration effect.

Replication on PASCAL VOC2012. We test the same distinction on VOC2012 [11] at higher accuracy, with 6 of 20 object classes held out. The dense-CLIP base reaches 81.8 mIoU at full resolution and 75.3 at the patch resolution used in Table 4. Per-class temperature lowers novel-class IoU from 75.2 to 0.0, and the decoder reaches 16.1. Global temperature reduces ECE from 43.4 to 4.3. Diffusion retains the novel classes, reaching 77.1 novel-class IoU. Figure 2 illustrates the difference on a held-out dining table, which diffusion preserves and per-class temperature repaints as seen. In the final row, both retain the seen dog, while diffusion also retains the novel sofa. These examples expose selective loss of novel objects within an otherwise recognizable scene.

Table 1. Evaluation protocols for main-paper result tables. Methods within each comparison share the evaluation cache and scoring resolution.

Result table	Dataset and stack	Evaluation images	Scoring resolution	Splits
2	ADE20K masked crops	SAM-cacheable subset of 2000	256×256	seed 0
3, 5, 6	ADE20K dense CLIP	all 2000 validation images	CLIP patch grid	5
4	VOC2012 dense CLIP	full validation set	CLIP patch grid	5

Table 2. Frozen SAM+CLIP corrections on ADE20K masked crops, using the original cache and seed-0 split.

Method	mIoU	hIoU	novel-IoU	ECE↓
Proposal+CLIP (independent)	8.9	9.1	9.4	14.6
Graph diffusion ($\alpha=0.4, k=2$)	9.0	9.4	10.0	5.9
Trained graph decoder	7.1	2.0	1.1	23.8

Table 3. Dense-CLIP corrections on ADE20K at patch resolution over five seen/novel splits.

Calibrator / head	mIoU	hIoU	ECE↓	novel-ECE↓	novel-IoU
Uncalibrated	14.5	14.7	27.2	29.7	15.1
Global temperature	14.5	14.7	8.7	10.0	15.1
Per-class temperature	4.1	0.0	15.8	43.8	0.0
Graph diffusion	15.0	15.2	27.8	30.6	15.7
Trained decoder	5.6	0.3	2.6	31.6	0.1

Table 4. Dense-CLIP corrections on VOC2012 at patch resolution, with 6 of 20 classes held out per split.

Calibrator / head	mIoU	ECE↓	novel-ECE↓	novel-IoU
Uncalibrated	75.3	43.4	43.6	75.2
Global temperature	75.3	4.3	4.2	75.2
Per-class temperature	43.8	15.4	58.8	0.0
Graph diffusion	77.2	44.6	44.9	77.1
Trained decoder	51.7	12.7	16.7	16.1

Generic smoothing and trust-region constraints. We remove the frozen anchor from graph smoothing to obtain a generic Laplacian control on dense ADE20K (Supp. Table S6). It reaches 16.3 novel-class IoU and 15.5 mIoU, compared with 15.7 and 15.0 for anchored diffusion. Retention is thus shared by two smoothing operators with different anchoring choices.

For the trained head, recovery depends on the residual bound. A logit-space trust region of ± 2 yields 5.8 novel-class IoU against the 15.1 baseline. Tightening the bound τ produces monotone recovery (Supp. Table S8). At $\tau=0.5$, the head reaches 14.6 novel-class IoU and 15.2 mIoU.

Text-conditioned calibration and risk-coverage. We fit a text-conditioned calibrator on seen classes, predicting $T_c = \text{softplus}(\text{MLP}(P_c))$ from class text prototype P_c . The

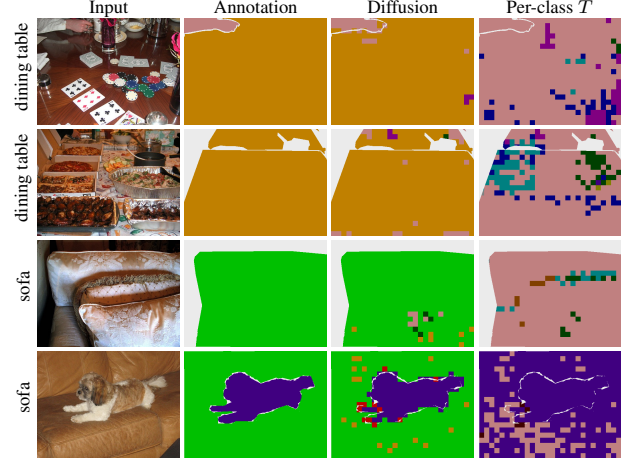


Figure 2. VOC2012 predictions for held-out classes, identified by row labels.

softplus function $\log(1 + e^u)$ converts an MLP output u into a positive temperature. Novel classes receive temperatures through their text, yet novel-class IoU falls to 1.9. Text conditioning therefore extends the observed failure beyond independently fitted class parameters.

We rank predictions by confidence and measure error as the retained fraction increases. The area under this risk-coverage curve (AURC, lower is better) ranks global temperature first among the calibration and diffusion controls at 48.1. Diffusion (49.2) and the uncalibrated base (49.9) differ by 0.7. The two trained corrections reach lower AURC (34.2 and 38.2) while remaining confident on dominant seen classes. Neither ECE nor AURC captures their novel-class loss.

Alignment-preserving learned corrections. We test anchoring under matched dense-CLIP splits and supervision (Table 5, Supp. Table S7). B1 minimizes $\mathcal{L} = \text{CE}_{\text{seen}} + \lambda \text{KL}$, combining seen-label cross-entropy with Kullback–Leibler divergence from frozen to corrected probabilities over all classes and sampled patches. B2 penalizes cosine distance between corrected and frozen embeddings. B3 mixes output probabilities, and B4 interpolates toward the uncorrected head using WiSE-FT [39]. B5 temporarily withholds seen classes as pseudo-novel

Table 5. Alignment-preserving corrections on ADE20K dense CLIP under matched supervision. Results are mean $\pm$ std over five splits.

Correction	mIoU	seen-IoU	novel-IoU	hIoU	ECE $\downarrow$	novel-ECE $\downarrow$	drift	novel%	C
<i>Class-symmetric (no head)</i>									
Uncalibrated (frozen baseline)	14.5 $\pm$ 0.0	14.3 $\pm$ 0.3	15.1 $\pm$ 1.0	14.7 $\pm$ 0.4	27.2 $\pm$ 0.0	29.7 $\pm$ 9.2	0.0	15.6	1.00
Global temperature	14.5 $\pm$ 0.0	14.3 $\pm$ 0.3	15.1 $\pm$ 1.0	14.7 $\pm$ 0.4	8.7 $\pm$ 0.7	10.0 $\pm$ 3.2	13.4	15.6	1.00
Anchored graph diffusion	15.0 $\pm$ 0.0	14.8 $\pm$ 0.3	15.7 $\pm$ 1.0	15.2 $\pm$ 0.4	27.8 $\pm$ 0.0	30.6 $\pm$ 9.4	0.2	15.6	1.04
Generic graph Laplacian	15.5 $\pm$ 0.0	15.3 $\pm$ 0.3	16.3 $\pm$ 1.1	15.8 $\pm$ 0.4	28.5 $\pm$ 0.0	31.5 $\pm$ 9.7	1.3	15.6	1.08
<i>Learned, seen-only, class-specific</i>									
Seen-only residual head	6.5 $\pm$ 0.3	8.1 $\pm$ 0.3	0.1 $\pm$ 0.1	0.1 $\pm$ 0.3	3.9 $\pm$ 2.5	34.2 $\pm$ 5.2	56.9	0.0	0.01
<i>Learned with explicit alignment preservation</i>									
B1 frozen-logit distillation ($\lambda=0.1$)	8.7 $\pm$ 0.4	10.1 $\pm$ 0.3	3.3 $\pm$ 1.8	4.7 $\pm$ 2.1	2.8 $\pm$ 1.2	29.2 $\pm$ 5.6	52.6	0.2	0.22
B1 frozen-logit distillation ($\lambda=0.3$)	10.2 $\pm$ 0.5	11.4 $\pm$ 0.5	5.4 $\pm$ 1.9	7.1 $\pm$ 1.8	4.9 $\pm$ 2.0	22.6 $\pm$ 6.7	48.4	0.6	0.35
B1 frozen-logit distillation ($\lambda=1.0$)	12.3 $\pm$ 0.6	13.2 $\pm$ 0.4	8.6 $\pm$ 2.1	10.3 $\pm$ 1.6	13.6 $\pm$ 2.1	12.9 $\pm$ 5.9	39.0	1.6	0.57
B1 frozen-logit distillation ($\lambda=3.0$)	14.6 $\pm$ 0.5	15.2 $\pm$ 0.5	12.5 $\pm$ 2.1	13.6 $\pm$ 1.3	24.2 $\pm$ 1.8	11.1 $\pm$ 4.3	25.6	3.6	0.83
B1 frozen-logit distillation ($\lambda=10.0$)	15.8 $\pm$ 0.3	16.1 $\pm$ 0.4	14.8 $\pm$ 1.1	15.4 $\pm$ 0.6	33.6 $\pm$ 1.2	18.8 $\pm$ 10.2	10.0	7.4	0.98
B1 frozen-logit distillation ($\lambda=30.0$)	15.8 $\pm$ 0.2	15.9 $\pm$ 0.4	15.3 $\pm$ 0.8	15.6 $\pm$ 0.3	35.1 $\pm$ 1.0	25.6 $\pm$ 9.7	2.5	11.0	1.01
B2 embedding distillation ($\lambda=0.0$)	11.1 $\pm$ 0.9	13.1 $\pm$ 1.1	2.7 $\pm$ 0.7	4.5 $\pm$ 1.0	11.2 $\pm$ 4.0	43.6 $\pm$ 8.1	57.8	0.5	0.18
B2 embedding distillation ($\lambda=1.0$)	12.1 $\pm$ 0.8	13.9 $\pm$ 0.9	4.5 $\pm$ 1.2	6.7 $\pm$ 1.4	6.0 $\pm$ 1.8	32.7 $\pm$ 7.2	53.5	1.0	0.29
B2 embedding distillation ($\lambda=10.0$)	14.4 $\pm$ 0.6	15.3 $\pm$ 0.5	10.6 $\pm$ 2.5	12.4 $\pm$ 1.8	13.9 $\pm$ 2.3	14.4 $\pm$ 3.7	38.4	3.4	0.70
<i>Episodic head / learned affinity</i>									
B5 pseudo-novel episodic training	6.8 $\pm$ 0.6	8.4 $\pm$ 0.7	0.6 $\pm$ 0.6	1.1 $\pm$ 1.0	6.7 $\pm$ 3.9	28.9 $\pm$ 7.7	52.9	0.0	0.04
B6 learned class-agnostic affinity	15.0 $\pm$ 0.0	14.8 $\pm$ 0.3	15.7 $\pm$ 1.0	15.2 $\pm$ 0.4	27.8 $\pm$ 0.0	30.5 $\pm$ 9.4	0.2	15.6	1.04

Drift: probability-space distance in Eq. (2). *novel%*: labeled patches predicted as a novel class (15.6 for the frozen baseline). C : collapse index, novel-IoU relative to that baseline.

and preserves their relative probabilities. B6 learns class-agnostic graph affinities, the weights that connect patches.

At $\lambda=30$, distillation retains 15.3 novel-class IoU ($C=1.01$), raises mIoU from 14.5 to 15.8, and raises seen-class IoU from 14.3 to 15.9. B6 retains 15.7 novel-class IoU by learning affinities within anchored diffusion. We therefore recover retention while learning either constrained scores or shared affinities. B5 reaches 0.6 novel-class IoU. Its penalty preserves relative probabilities within pseudo-novel seen classes, leaving competition with the actual novel classes unconstrained.

Novel-class coverage under anchoring. We then ask how much of the scene receives novel labels. We measure novel-class coverage as the fraction of labeled evaluation patches assigned to any novel class, reported as novel% in Table 5. This prediction rate measures the spatial extent of novel assignments. Recall instead conditions on ground-truth novel locations, while risk-coverage analysis measures the fraction of predictions retained after confidence filtering.

Stronger distillation increases the novel prediction rate from 0.2% at $\lambda=0.1$ to 3.6% at $\lambda=3$, 7.4% at $\lambda=10$ and 11.0% at $\lambda=30$. Even at $C=1.01$, the head assigns fewer locations to novel classes than the frozen baseline (15.6%). Matching aggregate novel-class IoU therefore coexists with a smaller predicted novel region. Class-symmetric controls, including B6, retain the baseline prediction rate. Our results therefore distinguish recovering overlap accuracy from restoring the spatial extent of novel predictions.

The anchoring frontier. In Figure 3, we plot seen-class IoU horizontally and novel-class IoU vertically to expose the anchoring trade-off. The output gate g mixes probabilities, while the residual scale η scales the learned logit correction (full grid in Supp. Table S7). These two paths move monotonically between $C=1.0$ at the frozen endpoint and $C=0.005$ at the head endpoint. Distillation strength λ and trust-region size τ produce corresponding monotone changes in novel-class IoU (Table S8). The upper-right anchored points show that recovery can accompany gains on seen classes, beyond simply returning to the frozen predictor.

Alignment-drift diagnostics and confound ablation. Using the drift defined in Eq. (2), we examine retention alongside the novel prediction rate. We also record the change in the gap between the largest seen and novel probabilities, which measures movement toward seen-class predictions. Across 38 measured class-specific settings (32 distinct corrections, 13 families), drift correlates with C at $r=-0.90$. Restricting the analysis to distinct corrections gives $r \approx -0.88$. To account for correlated sweeps, we bootstrap the 13 families, obtaining a 95% CI of $[-0.96, -0.74]$. Correlation between family means is -0.65 (CI $[-0.95, -0.23]$).

Figure 4 separates class-specific settings from class-symmetric controls. Its descending fit describes the former, while the controls establish its scope. Global temperature and LPOSS retain novel-class IoU at drift values of 13.4 and 50.6, respectively. The remaining class-symmetric corrections also retain it over their observed drift range. Within

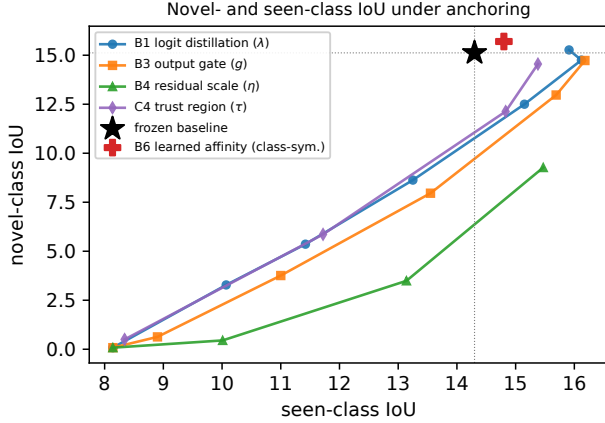


Figure 3. Seen–novel IoU trade-offs under distillation (λ), output gating (g), residual scaling (η) and trust-region (τ) sweeps. The star marks the frozen baseline.

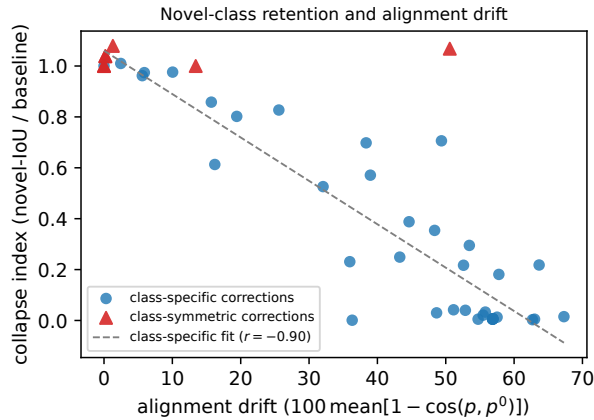


Figure 4. Novel-class retention versus alignment drift across 38 class-specific settings in 13 families ($r=-0.90$). Class-symmetric controls retain novel-class IoU across the observed drift range.

the class-specific family, the reduction accompanies a shift toward seen classes. Drift must therefore be interpreted together with the structure of the correction.

We test alternative explanations by changing sampling, loss, logit scale, supervision breadth (30 to 120 seen classes), and patch-label purity one factor at a time (Supp. Table S8). Novel-class IoU remains at or below 0.3 in these variations except the narrowest supervision setting. With 30 seen classes, it reaches 3.3 with a standard deviation of 4.3, against the 15.1 baseline ($C=0.22$). Increasing supervision to 120 classes lowers C to 0.01. Broader seen supervision thus leaves the loss unresolved. In this ablation, tightening the trust region supplies the monotone recovery.

LPOSS under the matched protocol. We evaluate the released LPOSS propagation [33] on the same cached logits, images and five splits as the controls (Table 6). LPOSS uses affinities from the self-supervised DINO ViT-B/16 encoder. Its graph uses $k_{nn}=400$ nearest neighbors, similarity exponent $\gamma=3$ and propagation weight $\alpha=0.95$, solved by conjugate gradients. Diffusion and the generic Laplacian use CLIP-feature graphs with $k_{nn}=8$. We preserve the four upstream propagation functions and replace the GPU neighbor-search kernel with exact nearest-neighbor selection.

LPOSS reaches 15.1 mIoU and 16.1 novel-class IoU, comparable to the generic Laplacian’s 15.5 and 16.3. Despite the largest class-symmetric drift (50.6), LPOSS retains novel-class IoU ($C=1.07$). Adding the same seen-only head raises seen-class IoU to 17.0 but lowers novel-class IoU from 16.1 to 10.7. A stronger propagation base therefore remains sensitive to a class-specific correction fitted on seen labels.

DINO-proxy patch affinities. We further strengthen the base by replacing CLIP affinities with DINO ViT-B/16 affinities, reaching 15.5 mIoU and 16.3 novel-class IoU. This is our implementation of ProxyCLIP-style attention [22], selected on training data. Per-class temperature reduces novel-class IoU to 0.0, and the seen-only head reduces it to 0.1. Global temperature lowers ECE from 30.7 to 9.7, while class-symmetric corrections retain novel-class IoU. Distillation with $\lambda \geq 10$ reaches 16.4 to 16.6 mIoU against the 15.5 baseline, retaining 94 to 98% of baseline novel-class IoU. Drift tracks retention within this run ($r=-0.93$ across 22 settings). The qualitative findings thus extend across the three dense ADE20K bases at 14.5, 15.1 and 15.5 mIoU.

Cross-vocabulary transfer (COCO-Stuff→ADE20K). In Table S9, we fit corrections on COCO-Stuff-171 patches from 600 COCO val2017 images and apply them to all 2000 ADE20K validation images. The backbone and 600-image fitting budget match the in-domain study. Fitting uses no ADE labels, although the datasets share some class names. Class-indexed temperatures and logit heads require a vocabulary mapping or parameters for ADE-only classes before transfer.

Class-symmetric and text-conditioned corrections transfer directly. We observe no severe novel-class loss comparable to the in-domain head. Global temperature lowers ADE ECE from 27.2 to 8.4 with unchanged mIoU. The text-conditioned calibrator reaches 6.5 ECE and 14.3 mIoU. A class-agnostic embedding head improves mIoU from 14.5 to 14.9 without anchoring and to 15.6 with an embedding cosine-distance penalty. These results establish a cross-dataset use case for corrections whose parameters can be

Table 6. LPOSS and correction controls on identical ADE20K dense-CLIP logits at patch resolution over five splits. C denotes retention relative to the frozen baseline.

Method	mIoU	seen-IoU	novel-IoU	hIoU	ECE↓	drift	C
Frozen baseline	14.5±0.0	14.3±0.3	15.1±1.0	14.7±0.4	27.2±0.0	0.0	1.00
Anchored diffusion	15.0±0.0	14.8±0.3	15.7±1.0	15.2±0.4	27.8±0.0	0.2	1.04
Generic graph Laplacian	15.5±0.0	15.3±0.3	16.3±1.1	15.8±0.4	28.5±0.0	1.3	1.08
LPOSS (authors’ code)	15.1±0.0	14.9±0.3	16.1±1.1	15.4±0.4	53.2±0.0	50.6	1.07
baseline + seen-only head	6.5±0.3	8.1±0.4	0.1±0.2	0.2±0.3	4.0±2.5	56.9	0.01
LPOSS + seen-only head	15.8±0.6	17.0±0.6	10.7±2.2	13.0±1.6	44.6±2.9	49.4	0.71

applied across vocabularies.

Deployment guidance. Our results suggest choosing the correction according to the deployment objective. When confidence calibration is the priority, fit a global temperature on seen-class data. It gives the lowest in-domain ECE among corrections retaining novel-class IoU and transfers across vocabularies. For a learned class-specific head, we recommend constraining adaptation with frozen-logit distillation or a tight trust region. The tested settings $\lambda \geq 10$ and $\tau = 0.5$ provide starting points for validation on the target stack. Assess novel prediction rate alongside IoU when spatial coverage matters, since an anchored head can recover IoU while assigning fewer locations to novel classes.

Monitor drift against the frozen predictor on a fixed evaluation support. This requires no novel-class labels and complements ECE and AURC, which miss the observed retention loss. Interpret drift within the correction family. Its association with retention is strong for class-specific heads, whereas class-symmetric corrections retain novel-class IoU throughout their tested drift range. We therefore use drift as a family-specific diagnostic, with retention checked separately. Where labeled evaluation data are available, report retention, coverage and calibration together to make the deployment trade-off explicit.

6. Limitations

Scale and scope. We isolate post-hoc corrections over frozen stacks and evaluate five bases spanning ADE20K masked crops, dense CLIP, LPOSS and DINO-proxy propagation, and VOC. ADE20K bases span about 9–15.5 mIoU and VOC 75–82. Our frontier and diagnostics use five dense-regime splits. Trained systems such as FC-CLIP, CAT-Seg and Mask-Adapter operate at higher accuracy and adapt components held fixed here. CorrCLIP and CASS provide additional training-free bases for extending the audit.

Coverage and drift scope. Anchored heads recover IoU more fully than coverage, assigning 7 to 11% of evaluation patches to novel classes against the baseline’s 15.6%.

For applications prioritizing novel-class coverage, we recommend evaluating this gap when choosing a correction. Our controlled head variations support the drift association, which remains correlational. Global temperature demonstrates that substantial drift can coexist with unchanged novel-class IoU, and averaging over patches can conceal losses on boundaries or small objects.

Open questions. We set anchor strength explicitly and obtain similar results for $\lambda \geq 10$ on the two bases where it is swept. Selecting λ from a drift budget measured on seen data offers a label-free direction for further study. Low-rank adaptation (LoRA), learned prompt embeddings and full-resolution anchored heads remain to be evaluated. Combining global temperature with an anchored head would connect the calibration and retention improvements observed separately here.

7. Conclusion

Across five bases, we find that fitting class-specific corrections on seen labels reduces novel-class IoU. Our drift analysis tracks this loss within the class-specific family ($r = -0.90$ pointwise, -0.65 across families). We recover retention with distillation and tight residual constraints while improving accuracy. The coverage gap shows why IoU alone cannot characterize recovery. Class-symmetric controls retain novel-class IoU throughout the observed drift range, and global temperature supplies effective calibration. For these frozen stacks, we recommend a concrete design choice: constrain class-specific heads toward the frozen predictions and evaluate novel-class retention alongside calibration and coverage. Drift provides an additional diagnostic when novel labels are unavailable.

Acknowledgments

We acknowledge Ho Chi Minh City University of Technology (HCMUT), VNU-HCM for supporting this study. We thank VNG Corporation for sponsoring the GPU resources used in this work.

Societal Impact

The audit identifies a failure mode that aggregate calibration metrics can conceal. Reporting novel-class retention alongside calibration makes this failure visible during model assessment. Operational use of calibrated abstention requires domain-specific validation, particularly where missed novel objects carry a high cost.

References

- [1] Anastasios N. Angelopoulos and Stephen Bates. Conformal prediction: A gentle introduction. *Foundations and Trends in Machine Learning*, 16(4):494–591, 2023. 2
- [2] Walid Bousselham, Felix Petersen, Vittorio Ferrari, and Hilde Kuehne. Grounding everything: Emerging localization properties in vision-language transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2
- [3] Junbum Cha, Jonghwan Mun, and Byungseok Roh. Learning to generate text-grounded mask for open-world semantic segmentation from only image-text pairs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2
- [4] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, et al. DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018. 2
- [5] Bowen Cheng, Ishan Misra, Alexander G. Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2022. 2
- [6] 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 (CVPR)*, 2024. 2
- [7] Marius Cordts, Mohamed Omran, Sebastian Ramos, et al. The Cityscapes dataset for semantic urban scene understanding. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016. 2
- [8] Jian Ding, Nan Xue, Gui-Song Xia, and Dengxin Dai. Decoupling zero-shot semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 11583–11592, 2022. 2
- [9] Zhipeng Ding, Xu Han, Peirong Liu, and Marc Niethammer. Local temperature scaling for probability calibration. In *IEEE/CVF International Conference on Computer Vision*, 2021. 2
- [10] Zheng Ding, Jieke Wang, and Zhuowen Tu. Open-vocabulary universal image segmentation with MaskCLIP. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, 2023. 2
- [11] 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. 2, 4
- [12] Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems*, 2017. 2
- [13] 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*, 2022. 2
- [14] Zhongrui Gui, Shuyang Sun, Runjia Li, Jianhao Yuan, Zhaochong An, Karsten Roth, Ameya Prabhu, and Philip Torr. kNN-CLIP: Retrieval enables training-free segmentation on continually expanding large vocabularies. *Transactions on Machine Learning Research*, 2024. 4
- [15] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning*, 2017. 2
- [16] Cong Han, Yujie Zhong, Dengjie Li, Kai Han, and Lin Ma. Open-vocabulary semantic segmentation with decoupled one-pass network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 2
- [17] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *International Conference on Learning Representations*, 2017. 2
- [18] Chanyoung Kim, Dayun Ju, Woojung Han, Ming-Hsuan Yang, and Seong Jae Hwang. Distilling spectral graph for object-context aware open-vocabulary semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15033–15042, 2025. 2
- [19] Alexander Kirillov, Eric Mintun, Nikhila Ravi, et al. Segment anything. In *International Conference on Computer Vision*, 2023. 1, 2, 3
- [20] Philipp Krähenbühl and Vladlen Koltun. Efficient inference in fully connected CRFs with gaussian edge potentials. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2011. 2
- [21] Mengcheng Lan, Chaofeng Chen, Yiping Ke, Xinjiang Wang, Litong Feng, and Wayne Zhang. ClearCLIP: Decomposing CLIP representations for dense vision-language inference. In *European Conference on Computer Vision*, 2024. 2
- [22] Mengcheng Lan, Chaofeng Chen, Yiping Ke, Xinjiang Wang, Litong Feng, and Wayne Zhang. ProxyCLIP: Proxy attention improves CLIP for open-vocabulary segmentation. In *European Conference on Computer Vision*, 2024. 2, 7
- [23] Boyi Li, Kilian Q. Weinberger, Serge Belongie, Vladlen Koltun, and René Ranftl. Language-driven semantic segmentation. In *International Conference on Learning Representations (ICLR)*, 2022. 2
- [24] Yongkang Li, Tianheng Cheng, Bin Feng, Wenyu Liu, and Xinggang Wang. Mask-Adapter: The devil is in the masks for open-vocabulary segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. 2
- [25] Feng Liang, Bichen Wu, Xiaoliang Dai, et al. Open-vocabulary semantic segmentation with mask-adapted CLIP.

- In *IEEE Conference on Computer Vision and Pattern Recognition*, 2023. 2
- [26] Tsung-Yi Lin, Michael Maire, Serge Belongie, et al. Microsoft COCO: Common objects in context. In *European Conference on Computer Vision*, 2014. 2
- [27] 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 (CVPR)*, pages 3491–3500, 2024. 2
- [28] Gensheng Pei, Xiruo Jiang, Xinhao Cai, Tao Chen, Yazhou Yao, and Byeungwoo Jeon. PEARL: Geometry aligns semantics for training-free open-vocabulary semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17927–17937, 2026. 2
- [29] Gensheng Pei, Xiruo Jiang, Yazhou Yao, Xiangbo Shu, Fumin Shen, and Byeungwoo Jeon. Taming SAM3 in the wild: A concept bank for open-vocabulary segmentation. *arXiv preprint arXiv:2602.06333*, 2026. 2
- [30] Alec Radford, Jong Wook Kim, Chris Hallacy, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 2021. 1, 2, 3
- [31] Nikhila Ravi et al. SAM 2: Segment anything in images and videos. In *International Conference on Learning Representations*, 2025. 2
- [32] Xiangheng Shan, Dongyue Wu, Guilin Zhu, Yuanjie Shao, Nong Sang, and Changxin Gao. Open-vocabulary semantic segmentation with image embedding balancing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 28412–28421, 2024. 2
- [33] Vladan Stojnić, Yannis Kalantidis, Jiří Matas, and Giorgos Tolias. LPOSS: Label propagation over patches and pixels for open-vocabulary semantic segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. 2, 7
- [34] Boyuan Sun, Yuqi Yang, Le Zhang, Ming-Ming Cheng, and Qibin Hou. CorrMatch: Label propagation via correlation matching for semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2
- [35] Weijie Tu, Weijian Deng, Dylan Campbell, Stephen Gould, and Tom Gedeon. An empirical study into what matters for calibrating vision-language models. In *International Conference on Machine Learning (ICML)*, 2024. 2
- [36] Dongdong Wang, Boqing Gong, and Liqiang Wang. On calibrating semantic segmentation models: Analyses and an algorithm. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23652–23662, 2023. 2
- [37] Feng Wang, Jieru Mei, and Alan Yuille. SCLIP: Rethinking self-attention for dense vision–language inference. In *European Conference on Computer Vision*, 2024. 2
- [38] Shuoyuan Wang, Jindong Wang, Guoqing Wang, Bob Zhang, Kaiyang Zhou, and Hongxin Wei. Open-vocabulary calibration for fine-tuned CLIP. In *International Conference on Machine Learning (ICML)*, 2024. 2
- [39] Mitchell Wortsman, Gabriel Ilharco, Jong Wook Kim, Mike Li, Simon Kornblith, Rebecca Roelofs, Raphael Gontijo-Lopes, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, and Ludwig Schmidt. Robust fine-tuning of zero-shot models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2, 4, 5
- [40] Monika Wysockańska, Oriane Siméoni, Michaël Ramamonjisoa, Andrei Bursuc, Tomasz Trzcinski, and Patrick Pérez. CLIP-DINOiser: Teaching CLIP a few DINO tricks for open-vocabulary semantic segmentation. In *European Conference on Computer Vision (ECCV)*, 2024. 2
- [41] 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 (CVPR)*, 2024. 2
- [42] Enze Xie, Wenhai Wang, Zhiding Yu, et al. SegFormer: Simple and efficient design for semantic segmentation with transformers. In *Advances in Neural Information Processing Systems*, 2021. 2
- [43] Jiarui Xu, Shalini De Mello, Sifei Liu, et al. GroupViT: Semantic segmentation emerges from text supervision. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2022. 2
- [44] Jiarui Xu, Sifei Liu, Arash Vahdat, Wonmin Byeon, Xiaolong Wang, and Shalini De Mello. Open-vocabulary panoptic segmentation with text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2
- [45] 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 (CVPR)*, 2023. 2
- [46] Qihang Yu, Ju He, Xueqing Deng, Xiaohui Shen, and Liang-Chieh Chen. Convolutions die hard: Open-vocabulary segmentation with single frozen convolutional CLIP. In *Advances in Neural Information Processing Systems*, 2023. 2
- [47] Dengke Zhang, Fagui Liu, and Quan Tang. CorrCLIP: Reconstructing patch correlations in CLIP for open-vocabulary semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 24677–24687, 2025. 2
- [48] Bolei Zhou, Hang Zhao, Xavier Puig, et al. Scene parsing through ADE20K dataset. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 1, 2, 3
- [49] Chong Zhou, Chen Change Loy, and Bo Dai. Extract free dense labels from CLIP. In *European Conference on Computer Vision (ECCV)*, 2022. 2