CALIBRATED CONCEPT DENOISING FOR OPEN-VOCABULARY SEMANTIC SEGMENTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Frozen vision-language encoders support open-vocabulary segmentation, but their dense features often mix target evidence with correlated background and competing concepts. This motivates a query-dependent scoring operation: the same channel can be useful for one concept and misleading for another. We propose Calibrated Concept Denoising (CCD), which reweights visual channels for each text query while retaining dense cosine scoring. CCD transfers concept masks from SmartCLIP, refines them with a prior-preserving residual adapter, and estimates masks directly from labeled competitor errors. A labeled set of 1,000 training images per dataset supports adaptation and branch selection; the selected branch is fixed for validation. Under the released ClearCLIP evaluator, CCD improves ViT-B/16 average mIoU by 0.48 points across six benchmarks, including a 2.20-point gain on Cityscapes. Controlled experiments across three feature families and a ViT-L/14 extension characterize where query-conditioned channel reweighting helps.

1 INTRODUCTION

Open-vocabulary semantic segmentation assigns text-defined categories to image regions beyond a fixed training vocabulary. CLIP supports this setting through large-scale image–text alignment Radford et al. (2021). Later methods learn semantic groups from image–text supervision Xu et al. (2022); Cha et al. (2023), expand segmentation supervision Ghiasi et al. (2022), or separate region formation from open-vocabulary recognition Shin et al. (2022); Ding et al. (2022). Transferring image-level CLIP features to dense prediction remains difficult: patch features inherit global context, co-occurring objects, and dataset bias. The resulting score maps often show background activation, blurred boundaries, and confusion between correlated categories.

A second line of work exposes or modifies CLIP features for dense inference. MaskCLIP extracts dense predictions from a frozen encoder Zhou et al. (2022); SCLIP redesigns the final self-attention operation Wang et al. (2024); and CLIP Surgery removes redundant components that interfere with localization Li et al. (2025). ClearCLIP decomposes the final transformer block to suppress class-agnostic dense noise Lan et al. (2024). Related methods use contextual dense prompting Rao et al. (2022), patch aggregation Luo et al. (2023), or one-stage pixel–text alignment Zhou et al. (2023). More recent training-free methods select attention heads Duan et al. (2025) or refine patch affinities using class distributions Kang et al. (2025). These advances motivate a complementary question at the scoring stage: which feature dimensions should contribute to each query? One dimension may support one category while providing false evidence for another.

Learned segmentation modules improve spatial modeling through mask adaptation, side networks, and cost aggregation Liang et al. (2023); Xu et al. (2023); Cho et al. (2024). SmartCLIP instead organizes CLIP dimensions through concept-specific masks Xie et al. (2025). Transferring these masks to patch-level scoring raises a distinct issue: dimensions useful for global recognition may not preserve localization, and masking can damage classes that a strong frozen scorer already separates.

We propose Calibrated Concept Denoising (CCD), a concept-conditioned channel reweighting framework for frozen open-vocabulary segmentation. CCD assigns each text query a non-negative mask and recomputes cosine similarity after masking the visual features. A pretrained SmartCLIP prior provides concept structure; a residual adapter learns task-specific adjustments while preserving

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.

that prior at initialization. A second construction estimates masks directly from channels associated with labeled competitor errors. Both retain the dense scoring interface and require no image decoder.

The choice of mask matters as much as its construction: a transferred prior or adapted mask can improve one dataset and degrade another. We therefore study CCD with a labeled budget of 1,000 target-dataset training images for adaptation and branch selection. The encoders remain frozen, and one branch is selected per dataset for validation. Correction and regression counts expose the local trade-offs of adaptation. Under the released ClearCLIP evaluator, the resulting ViT-B/16 policy improves six-dataset average mIoU by 0.48 points, with its largest gain on Cityscapes. Experiments across feature families examine the broader applicability of the scoring operation.

Contributions. The main contributions of this paper are:

- We formulate query-conditioned channel reweighting for frozen dense vision-language features, complementing modifications to spatial attention.
- We introduce CCD, which transfers and adapts concept masks and constructs masks directly from competitor errors while retaining dense cosine scoring.
- We evaluate dataset-level mask selection with an explicit target-label budget and analyze its correction–regression trade-offs across six benchmarks, three feature families, and two backbone sizes.

2 Related Work

Vision-language models for dense prediction. CLIP learns transferable visual representations from natural language supervision Radford et al. (2021). GroupViT Xu et al. (2022) and TCL Cha et al. (2023) learn grouping or mask-generation mechanisms from image–text pairs. Other approaches retrieve visual evidence, expand image-level supervision, or adapt CLIP to region proposals Shin et al. (2022); Ghiasi et al. (2022); Liang et al. (2023). Side-adapter networks, cost aggregation, and encoder–decoder designs improve spatial modeling Xu et al. (2023); Cho et al. (2024); Xie et al. (2024). CCD learns or transfers query-by-channel weights on frozen features, with predictions produced by dense image–text similarity.

Training-free dense CLIP inference. MaskCLIP extracts dense labels from CLIP Zhou et al. (2022), SCLIP modifies final-layer self-attention Wang et al. (2024), and ClearCLIP decomposes the final transformer block to improve localization Lan et al. (2024). CLIP Surgery removes redundant features Li et al. (2025). DIH-CLIP selects attention heads using their information entropy and incorporates intermediate-layer features Duan et al. (2025). CDAM derives patch affinities from the similarity of predicted class distributions Kang et al. (2025). These methods provide recent context for improving frozen dense representations. CCD intervenes in the channel contributions to each query's score. Its target-label calibration is a distinct source of supervision, stated explicitly in our comparisons.

Concept-conditioned representations. SmartCLIP learns text-conditioned masks through modular image–text alignment Xie et al. (2025). Its concept masks motivate CCD, but global alignment and patch-level discrimination impose different requirements. CCD transfers the pretrained masks to dense visual features, adapts them with pixel labels, and compares them with a direct construction from competitor errors. This separates the source of the concept prior from its utility for dense prediction.

3 Method

Calibrated Concept Denoising (CCD) reweights frozen visual channels separately for each text query. Its three mask constructions are a transferred concept prior, a residual adaptation of that prior, and an estimate from labeled competitor errors. A dataset-level calibration step chooses the inference branch.

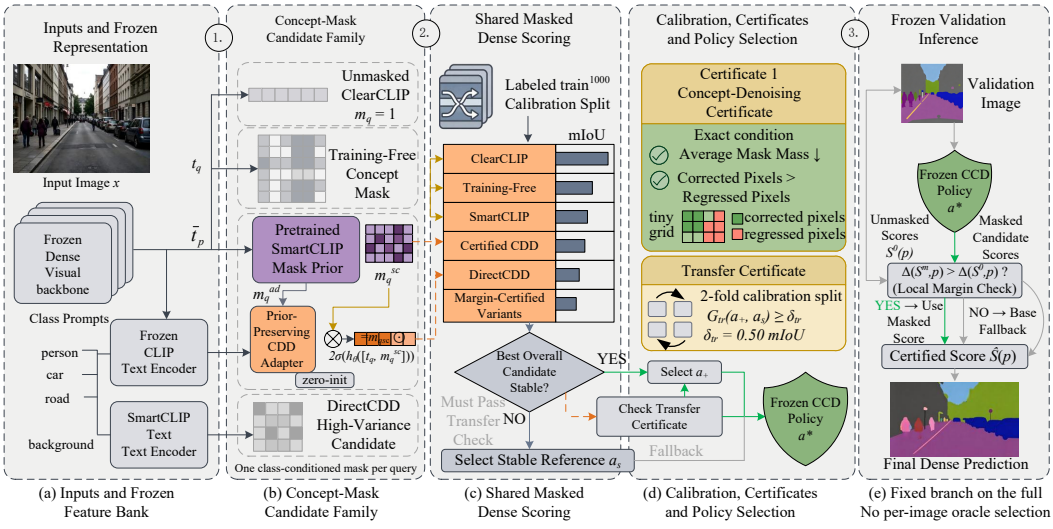


Figure 1: Overview of CCD's candidate scoring and calibration components. Frozen visual features and text embeddings support unmasked, transferred-mask, adapted-mask, and DirectCDD branches. Section 3.4 specifies the dataset-level selection rules used in the reported experiments.

3.1 Concept-Conditioned Dense Scoring

Let $X \in \mathbb{R}^{H\times W\times D}$ be the frozen visual feature map, x_p the feature at location p, and $t_q \in \mathbb{R}^D$ the frozen text embedding of query q. The mapping $\pi(q)$ assigns each of the Q queries to one of C classes. The unmasked scorer is

$$s_q^0(p) = \frac{x_p^\top t_q}{\|x_p\|_2 \|t_q\|_2 \tau}, \qquad S_c^0(p) = \max_{q:\pi(q)=c} s_q^0(p), \qquad \hat{y}^0(p) = \arg\max_c S_c^0(p). \tag{1}$$

For a non-negative mask $m_q \in \mathbb{R}_+^D$, CCD computes

$$s_q^m(p) = \frac{(m_q \odot x_p)^\top t_q^{(a)}}{\|m_q \odot x_p\|_2 \|t_q^{(a)}\|_2 \tau}, \qquad S_c^m(p) = \max_{q:\pi(q)=c} s_q^m(p). \tag{2}$$

Here $\odot$ denotes element-wise multiplication and $\tau > 0$ is a temperature. The branch-specific embedding $t_q^{(a)}$ is either the original frozen CLIP embedding or the frozen SmartCLIP embedding, as specified below. Norms are clamped below by 10^{-6} in the implementation. The main configurations use $\tau = 1$; Appendix A.6 describes relative score scaling in the controlled dense-mask network.

Masking changes the direction of the visual feature used by each query. In particular, $s_q^{\alpha m} = s_q^m$ for $\alpha > 0$ away from the numerical clamp, so selective reweighting depends on relative channel weights. The scorer can be evaluated without materializing a $Q \times H \times W \times D$ tensor: its numerator is $X(m \odot t^{(a)})^\top$, and its squared visual norm is $X^{\odot 2}(m^{\odot 2})^\top$, with X flattened over spatial locations.

3.2 Transferred Masks and Residual Adaptation

SmartCLIP prior. SmartCLIP Xie et al. (2025) learns a text-conditioned mask network through modular image–text alignment and sparsity regularization. Its released model was initialized from CLIP and trained on approximately one million ShareGPT4V image–text pairs. We reuse its pre-trained text encoder and mask network in frozen form. For the prompt "a photo of a {class}," let $\bar{t}_q$ be the full text representation used by its mask network and t_q^{SC} its projected text embedding. The soft mask is

$$m_q^{\mathrm{SC}} = \sigma(g_{\mathrm{SC}}(\bar{t}_q)). \tag{3}$$

The hard-mask variant thresholds each entry at 0.5. Transferred-mask branches pair these masks with either t_q or t_q^{SC} in Eq. (2). All masks must match the channel coordinates and dimension of the feature backbone. Our transfer experiments use the released ViT-B/16 SmartCLIP checkpoint ($D = 512$); the ViT-L/14 extension uses the direct construction below in its own $D = 768$ feature space.

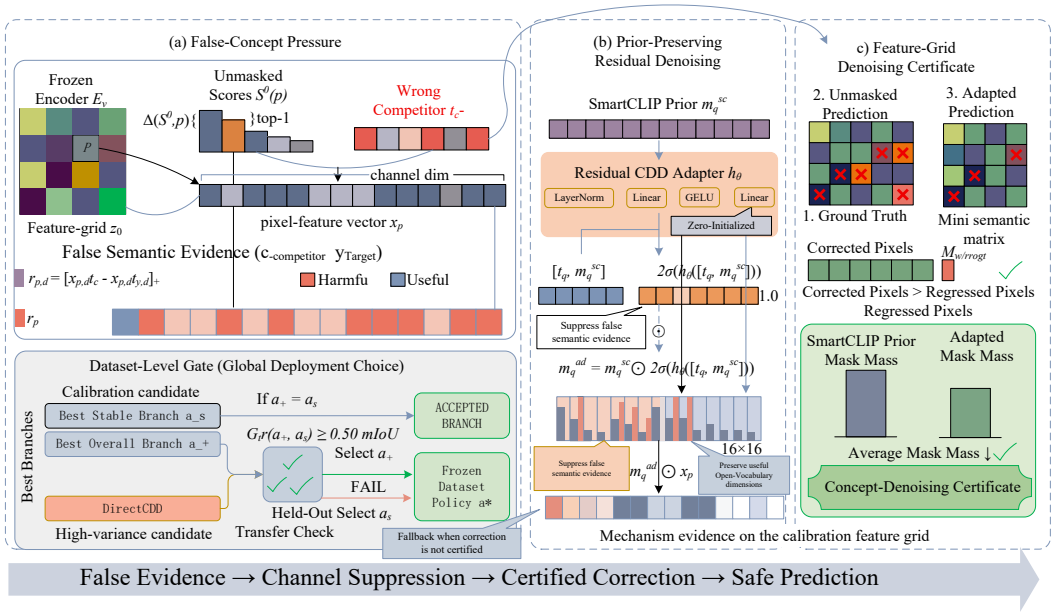


Figure 2: Concept-mask adaptation and diagnostics. A SmartCLIP prior initializes the residual adapter. False-concept pressure describes channel contributions to competitor errors; mask changes, corrections, and regressions characterize the adapted predictions. Sections 3.3–3.5 define these quantities and their roles in the pipeline.

Prior-preserving adapter. A small network h_θ adapts the prior:

$$m_q^{\mathrm{ad}} = m_q^{\mathrm{SC}} \odot 2\sigma\big(h_\theta([t_q, m_q^{\mathrm{SC}}])\big)\,. \tag{4}$$

The adapter consists of LayerNorm, a $2D$-to-D linear layer, GELU, and a D-to-D output projection. Only the output projection's weights and bias are initialized to zero, giving $m_q^{\mathrm{ad}} = m_q^{\mathrm{SC}}$ initially. The multiplicative factor lies in $(0, 2)$ and can suppress or amplify channels relative to the prior. The encoders and g_{SC} remain frozen.

For labeled feature-grid locations Ω_{fit}, we train the adapter with class-weighted cross-entropy:

$$\mathcal{L}_{\mathrm{CDD}} = -\frac{\sum_{p\in\Omega_{\mathrm{fit}}} w_{y_p} \log \mathrm{softmax}(S^m(p))_{y_p}}{\sum_{p\in\Omega_{\mathrm{fit}}} w_{y_p}}. \tag{5}$$

Here y_p is the target class and w_c is inversely proportional to its feature-grid frequency, with class counts clamped to at least one and weights normalized to mean one. We use AdamW for one epoch, learning rate 10^{-3}, batch size two, and seed zero. The Cityscapes adapter uses the original CLIP text embeddings for both its input and dense scoring.

3.3 Direct Masks from Competitor Errors

DirectCDD constructs masks from labeled feature-grid errors without a SmartCLIP prior or a learned mask network. This estimator uses one frozen text embedding t_c per class. For a misclassified location, let $c_p^- = \widehat{y}^0(p)$. We define the positive channel contribution to the raw competitor–target contrast as

$$r_{p,d} = \Big[x_{p,d}(t_{c_p^-,d} - t_{y_p,d})\Big]_+, \qquad [z]_+ = \max(0, z). \tag{6}$$

This *false-concept pressure* is a channel statistic for constructing the direct mask and interpreting adapter changes; it is not an extra term in the adapter loss. It measures the unnormalized contrast, whereas the final decision uses the normalized scores in Eq. (2).

Let $\mathcal{E}_c$ collect labeled locations incorrectly predicted as class c. The direct mask averages pressure within each competitor class and rescales it:

$$R_{c,d} = \frac{\sum_{p\in\mathcal{E}_c} r_{p,d}}{\max(1, |\mathcal{E}_c|)}, \qquad m_{c,d}^{\mathrm{dir}} = 1 - \frac{R_{c,d}}{\max(\epsilon, \max_j R_{c,j})}, \qquad \epsilon = 10^{-6}. \tag{7}$$

Classes with no observed errors retain the all-ones mask. DirectCDD evaluates two interventions: replacing the score vector when its top-two margin increases, or replacing only the score of the original competitor. A two-fold procedure retains competitor classes whose intervention improves mIoU in both directions and selects the intervention with the larger mean fold gain. The mask is then estimated on all training images. Appendix A.2 specifies the fold construction and score updates.

3.4 Calibration and Local Gating

Dataset-level selection. Let $\mathcal{A}$ contain the unmasked scorer, transferred-mask policies, the prior adapter, DirectCDD, and two text-derived mask references. The primary ViT-B/16 configuration selects

$$a^\star = \arg\max_{a \in \mathcal{A}} \mathrm{mIoU}_{\mathrm{train1000}}(a). \tag{8}$$

The unmasked branch provides the fallback when no candidate improves its score. Appendix A.1 defines the candidate configurations; Table 2 reports the selected branch on each dataset. The labeled training set supports both fitting and selection, so its mIoU is a selection statistic. Full-validation images are evaluated with the resulting fixed policy.

The ViT-L/14 extension compares DirectCDD with the unmasked scorer. It selects DirectCDD when the mean two-fold operator gain on train1000 reaches 0.50 mIoU points and otherwise retains the unmasked branch. The masks for each fold are estimated on the other fold; competitor and operator selection use those same fold results. This is an empirical selection rule.

Local margin gate. Transferred-mask and adapter branches can use a fixed local gate. For a score vector S, define $\Delta(S, p)$ as the difference between its largest and second-largest entries. The gate returns

$$\widehat{S}(p) = \begin{cases} S^m(p), & \Delta(S^m, p) > \Delta(S^0, p), \\ S^0(p), & \text{otherwise.} \end{cases} \tag{9}$$

This selects between masked and unmasked scores using their relative margins; its empirical effect is assessed through corrections and regressions.

3.5 Adapter Diagnostics

On a labeled diagnostic grid Ω_{diag}, let N_{cor} count locations changed from an incorrect unmasked prediction to a correct adapted prediction, and N_{reg} count the reverse. We report

$$M(m) = \frac{1}{QD} \sum_{q,d} m_{q,d}, \qquad \Delta_M = D\big[M(m^{\mathrm{SC}}) - M(m^{\mathrm{ad}})\big], \qquad N_{\mathrm{net}} = N_{\mathrm{cor}} - N_{\mathrm{reg}}. \tag{10}$$

Thus Δ_M is the mean reduction in summed channel weights per query, not a count of removed channels. The conjunction $\Delta_M > 0$ and $N_{\mathrm{net}} > 0$ summarizes contraction and net corrections in the reported parameterization. Mask mass is scale-dependent, so selective channel effects are examined through the mask-change plots and prediction changes. These diagnostics characterize the fitted masks; they do not enter the cross-entropy objective.

4 Experiments

4.1 Experimental Setup

We evaluate on the validation splits of PASCAL VOC20 (Everingham et al., 2010), PASCAL Context59 (Mottaghi et al., 2014), COCO-Object (Lin et al., 2014), COCO-Stuff164k (Caesar et al., 2018), ADE20K (Zhou et al., 2017), and Cityscapes (Cordts et al., 2016). Our setting uses frozen encoders with labeled target-dataset calibration. For each dataset, train1000 contains 1,000 images from its training split with dense segmentation labels. These labels support adapter fitting, direct-mask estimation, and branch selection; the 200-image adapter diagnostics are drawn from this same set. The adapter is fitted and scored on train1000, while DirectCDD uses two folds within it for competitor and operator selection. The reported setting does not evaluate transfer to classes withheld from this target supervision.

Method	Encoder	VOC20	Context 59	COCO-Object*	COCO-Stuff	ADE20K	City-scapes	Avg5	Avg6*
GroupViT	ViT-S/16	79.7	23.4	–	15.3	9.2	11.1	27.7	–
CoCu	ViT-S/16	–	–	–	13.6	11.1	15.0	–	–
TCL	ViT-B/16	77.5	30.3	–	19.6	14.9	23.1	33.1	–
CLIP	ViT-B/16	41.8	9.2	–	4.4	2.1	5.5	12.6	–
MaskCLIP	ViT-B/16	74.9	26.4	–	16.4	9.8	12.6	28.0	–
ReCo	ViT-B/16	57.7	22.3	–	14.8	11.2	21.1	25.4	–
CLIPSurgery	ViT-B/16	–	–	–	21.9	–	31.4	–	–
SCLIP	ViT-B/16	80.4	34.2	–	22.4	16.1	32.2	37.1	–
ClearCLIP	ViT-B/16	80.9	35.9	–	23.9	16.7	30.0	37.5	–
DIH-CLIP†	ViT-B/16	84.9	39.7	–	26.7	19.6	40.2	42.22	–
ClearCLIP*	ViT-B/16	80.92	35.83	33.04	23.89	16.65	30.08	37.47	36.74
CCD* (1,000 labels)	ViT-B/16	80.92	35.93	33.45	24.27	16.42	32.28	37.96	37.21

Table 1: Open-vocabulary segmentation (mIoU, %). Published references follow Lan et al. (2024); † denotes DIH-CLIP's own published results Duan et al. (2025). Asterisks mark our runs using the released ClearCLIP evaluator. CCD uses 1,000 densely labeled target-training images per dataset; ClearCLIP* uses none. Avg5 excludes COCO-Object; Avg6 includes it.

Dataset	U	P_A	P_B	P_{ad}	DirectCDD	Selected
VOC20	**81.58**	79.06	81.56	81.56	80.99	U
Context59	32.60	32.70	32.61	32.61	**32.84**	DirectCDD
COCO-Object	32.28	32.91	**32.91**	32.91	32.05	P_B
COCO-Stuff	23.21	23.66	23.66	23.66	**23.88**	DirectCDD
ADE20K	11.78	11.51	11.78	11.78	**11.89**	DirectCDD
Cityscapes	28.68	27.69	29.08	**31.05**	29.59	P_{ad}

Table 2: ViT-B/16 candidate selection on train1000 (mIoU, %). Bold marks the selected candidate. U is unmasked; P_A, P_B, and P_{ad} are fixed candidate policies defined in Appendix A.1.

We use two evaluation protocols. The primary comparison runs ClearCLIP and CCD through the released ClearCLIP evaluator, with its class-name files, resizing, sliding-window inference, and label mappings. The controlled protocol evaluates exported frozen feature bundles with a shared evaluator to examine individual mask configurations. These protocols have different absolute scores and are reported separately. The metric is full-validation mIoU in percentage points. Published zero-shot methods are included as performance references; the matched ClearCLIP run establishes the change obtained by the full calibrated CCD pipeline.

4.2 Benchmark Comparison and Selected Policies

The main comparison follows ClearCLIP's background-free five-dataset protocol, with COCO-Object added for the complete reproduced runs. The primary CCD row uses the recorded train1000-only branch selector. Appendix A.3 separately documents the earlier selection variant that also consulted validation feature-bundle scores.

With ViT-B/16 and the same released evaluator, CCD improves Avg5 by 0.49 points and Avg6 by 0.48 points (computed before rounding). Dataset changes are $+0.00$, $+0.10$, $+0.41$, $+0.38$, -0.23, and $+2.20$ in table order. Cityscapes accounts for most of the gain; ADE20K shows a small regression. DIH-CLIP reports higher scores on the five overlapping background-free datasets, providing current performance context for this ClearCLIP-based study.

Table 2 reports the training scores used for branch selection. COCO-Object selects a soft transferred mask with frozen SmartCLIP text and unconditional replacement; the tied policies deploy the same branch. Cityscapes selects the prior adapter with CLIP text and a margin gate. Context59, COCO-Stuff, and ADE20K select DirectCDD with CLIP text and the competitor-only update, while VOC20 retains the unmasked scorer. The two text-only candidates do not exceed the listed winner on any dataset. Thus the aggregate result reflects the calibrated candidate family.

Family	ΔCtx.	ΔObj.	ΔCity	ΔAvg.
MaskCLIP-style	+0.47	+6.17	+0.59	+2.41
SCLIP-style	+0.00	+5.58	+0.88	+2.16
ClearCLIP-style	+0.18	+6.36	+6.01	+4.18

Table 3: CCD masks applied to controlled feature families. Entries are absolute mIoU gains (%); Ctx., Obj., and City denote Context59, COCO-Object, and Cityscapes. Each change is measured within its feature-bundle protocol. The ClearCLIP-style Cityscapes entry uses the prior adapter (15.20 to 21.21), not the separate dense-mask network in Table 4.

Split	MaskCLIP	SCLIP	ClearCLIP	CCD
VOC20	53.58	64.70	75.10	**75.98**
Context59	23.02	26.63	31.48	**31.67**
COCO-Object	15.10	19.29	22.46	**28.82**
COCO-Stuff	14.73	17.12	20.74	**21.01**
ADE20K	10.85	11.03	14.27	**14.43**
Cityscapes	22.99	12.99	15.20	**24.53**
Avg.	23.38	25.29	29.87	**32.74**

Table 4: Controlled feature-bundle results under the shared project evaluator (mIoU, %). The controlled policy uses a separate dense-mask network on Cityscapes; these scores provide the reference for the controlled ablations.

Split	ClearCLIP	CCD	Δ
VOC20	79.99	79.99	+0.00
Context59	29.39	30.19	+0.80
COCO-Object	30.12	31.16	+1.04
COCO-Stuff	19.89	20.28	+0.39
ADE20K	15.02	15.25	+0.23
Cityscapes	27.96	27.96	+0.00
Avg5	34.45	34.73	+0.28
Avg6	33.73	34.14	+0.41

Table 5: ViT-L/14 transfer on full validation splits with the official runner (mIoU, %). The train1000 policy accepts DirectCDD when its mean two-fold operator gain reaches 0.50 points; otherwise CCD retains ClearCLIP.

Dataset	ΔFold	ΔVal.	Selected	ΔPolicy
VOC20	+0.18	−0.11	–	+0.00
Context59	+0.81	+0.80	✓	+0.80
COCO-Object	+0.98	+1.04	✓	+1.04
COCO-Stuff	+0.58	+0.39	✓	+0.39
ADE20K	+0.57	+0.23	✓	+0.23
Cityscapes	+0.28	−0.43	–	+0.00

Table 6: ViT-L/14 selection analysis (mIoU points). ΔFold is the DirectCDD candidate's mean training-fold gain; ΔVal. is its full-validation gain. Selection uses ΔFold ≥ 0.50. ΔPolicy is the resulting validation gain after selection.

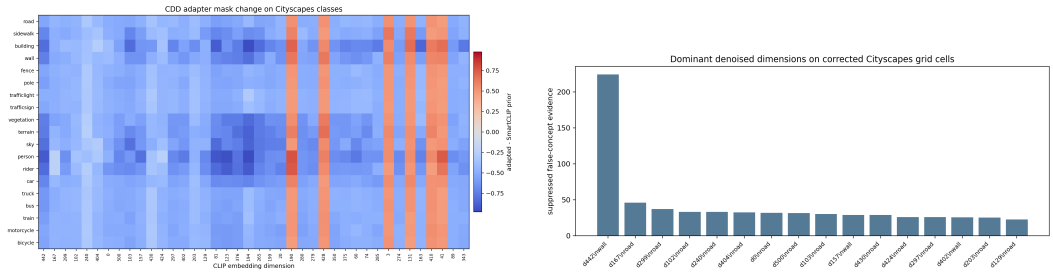
4.3 Feature Families and Backbone Extension

Table 3 tests channel masking across three feature families on Context59, COCO-Object, and Cityscapes. Full-validation changes are non-negative in all nine tested pairs, with the largest gains on COCO-Object and Cityscapes. Each change is measured against the corresponding unmasked features under the same controlled evaluator.

Table 4 reports a separate six-dataset controlled policy. Its Cityscapes branch is a directly trained dense-mask network with relative score temperature two, reaching 24.53 mIoU; the prior adapter used in the plug-in table reaches 21.21. The controlled policy improves average mIoU by 2.86 points using unrounded scores. Appendix A.6 defines the dense-mask network and its temperature selection.

Table 5 evaluates the direct mask construction with ViT-L/14. Context59, COCO-Object, COCO-Stuff, and ADE20K select DirectCDD; VOC20 and Cityscapes retain ClearCLIP. Estimating masks directly in the new backbone's feature space improves Avg6 by 0.41 points without transferring ViT-B/16 masks.

Selection and validation gains. Table 6 compares ViT-L/14 candidates before and after selection. All four accepted candidates improve validation mIoU. VOC20 and Cityscapes have positive fold



(a) Mask change on selected CLIP dimension IDs.

(b) Corrected-error dimensions.

Figure 3: Post-hoc channel diagnostics for the Cityscapes prior adapter. The left panel shows class-conditioned mask changes. The right panel ranks channels by $\sum_{p\in\mathcal{C}} r_{p,d}[m^{\mathrm{SC}}_{c_p^-,d} - m^{\mathrm{ad}}_{c_p^-,d}]_+$, where $\mathcal{C}$ contains corrected locations. The ranking highlights associations between suppression and corrected errors.

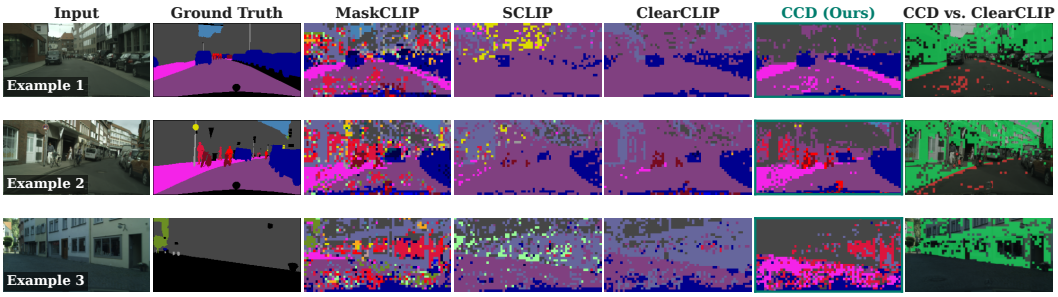


Figure 4: Cityscapes predictions under the matched ViT-B/16 feature-bundle protocol. Columns compare MaskCLIP, SCLIP, ClearCLIP, and the CCD prior adapter on the same three validation images. The final column shows CCD's changes to ClearCLIP: green marks corrections and red marks regressions. These examples illustrate high net-correction cases and are ordered by net corrections.

Dataset	Val.	Corrected	Regressed	Net	Ratio	Δ_M
VOC20	71.18	1047	2799	-1752	0.37	-113.72
Context59	30.61	8247	9803	-1556	0.84	-75.61
COCO-Object	21.27	12517	49361	-36844	0.25	-75.85
COCO-Stuff	19.25	6069	6893	-824	0.88	-118.91
ADE20K	13.81	4800	5593	-793	0.86	-30.79
Cityscapes	21.21	50519	24989	25530	2.02	102.56

Table 7: Prior-adapter validation mIoU (%) under the controlled evaluator and diagnostics on 200 images from each train1000 set. $\Delta_M = D[M(m^{\mathrm{SC}}) - M(m^{\mathrm{ad}})]$ is a summed-weight reduction per query; Ratio is corrected divided by regressed locations.

gains of 0.18 and 0.28 but validation changes of -0.11 and -0.43; the threshold retains their unmasked predictions. Gain magnitudes also vary: ADE20K's 0.57-point fold gain becomes 0.23 on validation, while COCO-Object's 0.98-point gain becomes 1.04.

4.4 Mask Changes and Prediction Changes

On 200 Cityscapes training images, the prior adapter corrects 50,519 feature-grid locations and regresses 24,989, for 25,530 net corrections and a ratio of 2.02. The mean summed mask weight decreases from 267.53 to 164.97 per query, a drop of 102.56 (normalized mean drop 0.2003 for $D = 512$). The channel ranking in Figure 3 associates several suppressed dimensions with corrected false *wall* and *road* predictions. The selected validation examples in Figure 4 illustrate both corrections and regressions. These observations describe the learned mask's behavior; the adapter itself is optimized by cross-entropy.

Configuration	VOC	Ctx.	Obj.	Stuff	ADE	City	Avg.	ΔAvg.
Full policy	75.98	31.67	28.82	21.01	14.43	24.53	**32.74**	–
$\tau_m = 1$	75.98	31.67	28.82	21.01	14.43	22.66	32.43	−0.31
w/o dense	75.98	31.67	28.82	21.01	14.43	16.23	31.35	−1.38
Text-only	75.10	31.48	27.84	20.74	14.27	15.20	30.77	−1.97
Unmasked	75.10	31.48	22.46	20.74	14.27	15.20	29.87	−2.86

Table 8: Controlled configuration comparisons (mIoU, %). Dataset order follows Table 1. The full policy uses $\tau_m = 2$; the $\tau_m = 1$ row uses the same checkpoint. Changes are computed before rounding.

Step	Selective	Gate	Mask	COCO-Object	Avg.
0	✓	✓	✓	27.84	**30.77**
1	–	✓	✓	27.84	29.74
2	–	–	✓	27.61	29.30
3	–	–	–	22.46	29.87

Table 9: Cumulative text-only mask ablations (mIoU, %). Each step removes one component: vocabulary-level selective activation, local gating, then masking. Avg. covers six datasets under the controlled evaluator.

Table 7 shows that this pattern is specific to the Cityscapes adapter: the other five adapters increase mask mass and introduce more regressions than corrections on the diagnostic images. This motivates evaluating multiple mask constructions rather than applying residual adaptation unconditionally.

4.5 Configuration Comparisons

Mask construction and relative temperature. Table 8 reports absolute scores under the controlled evaluator. Replacing the Cityscapes dense-mask network with the top-200 text mask (w/o dense) reduces Cityscapes mIoU from 24.53 to 16.23 and the six-dataset average by 1.38 points. The text-only selective policy reaches 30.77 on average, 1.97 points below the full policy; its largest gaps are on Cityscapes and COCO-Object. The unmasked row removes both masking and SmartCLIP text substitution and scores 29.87.

Temperature acts on the dense branch before the margin gate: masked scores are divided by τ_m, while unmasked scores retain unit temperature. On the same seed-zero checkpoint, changing τ_m from one to two raises Cityscapes mIoU from 22.66 to 24.53. This scaling preserves the masked branch's argmax but changes which locations pass the relative-margin comparison. The resulting 1.87-point gain isolates the effect of score scaling on this checkpoint. Appendix A.6 gives the training setup and temperature search.

Selective activation and local gating. Table 9 examines the text-only reference, whose mask construction requires no target labels. Disabling vocabulary-level selective activation lowers the average from 30.77 to 29.74 while leaving COCO-Object unchanged. Removing local gating further lowers it to 29.30. Finally, removing the mask recovers 29.87 on average but reduces COCO-Object from 27.61 to 22.46. Thus this mask helps COCO-Object even without gating, whereas selective activation and gating improve its aggregate behavior across datasets.

5 Conclusion

CCD reweights query-specific channels in frozen dense vision-language features using transferred masks, residual adaptation, and competitor-error estimates. With 1,000 labeled training images per dataset, calibrated branch selection improves average mIoU across six benchmarks, with its largest gain on Cityscapes and a regression on ADE20K. Feature-family experiments and a ViT-L/14 extension characterize its applicability, while correction–regression diagnostics expose the trade-offs of each mask.

AI USE STATEMENT

Generative AI tools were used solely to assist with and polish the writing of this manuscript. The authors reviewed and revised all AI-assisted content and take full responsibility for the final manuscript.

REPRODUCIBILITY STATEMENT

The paper specifies the frozen dense CLIP scoring path, SmartCLIP mask prior, prior-preserving residual adapter, direct mask estimator, calibration protocol, controlled feature bundles, six segmentation benchmarks, and evaluation metrics. The appendix records the controlled protocol, policy audit, feature families, and ablations. This repository contains the manuscript source, figures, tables, and references; it does not include the complete training and feature-export code, checkpoints, or all pretrained model assets. Reproducing the reported numbers requires those artifacts, the public datasets, and the model versions and evaluator settings specified in the paper.

REFERENCES

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.

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*, pp. 11165–11174, 2023.

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.

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

Jian Ding, Nan Xue, Gui-Song Xia, and Dengxin Dai. Decoupling zero-shot semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11583–11592, 2022.

Songsong Duan, Xi Yang, and Nannan Wang. DIH-CLIP: Unleashing the diversity of multi-head self-attention for training-free open-vocabulary semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22794–22803, 2025.

Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2): 303–338, 2010.

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.

Dong Un Kang, Hayeon Kim, and Se Young Chun. Class distribution-induced attention map for open-vocabulary semantic segmentations. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=CMqOfvD3tO.

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*, pp. 143–160. Springer, 2024.

Yi Li, Hualiang Wang, Yiqun Duan, Jiheng Zhang, and Xiaomeng Li. A closer look at the explainability of contrastive language-image pre-training. *Pattern Recognition*, 162:111409, 2025.

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.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pp. 740–755. Springer, 2014.

Huaishao Luo, Junwei Bao, Youzheng Wu, Xiaodong He, and Tianrui Li. Segclip: Patch aggregation with learnable centers for open-vocabulary semantic segmentation. In *International Conference on Machine Learning*, pp. 23033–23044. PMLR, 2023.

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*, pp. 891–898, 2014.

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.

Yongming Rao, Wenliang Zhao, Guangyi Chen, Yansong Tang, Zheng Zhu, Guan Huang, Jie Zhou, and Jiwen Lu. Denseclip: Language-guided dense prediction with context-aware prompting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 18082–18091, 2022.

Gyungin Shin, Weidi Xie, and Samuel Albanie. Reco: Retrieve and co-segment for zero-shot transfer. *Advances in Neural Information Processing Systems*, 35:33754–33767, 2022.

Feng Wang, Jieru Mei, and Alan L. Yuille. SCLIP: rethinking self-attention for dense vision-language inference. In *ECCV (21)*, volume 15079 of *Lecture Notes in Computer Science*, pp. 315–332. Springer, 2024.

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*, pp. 3426–3436, 2024.

Shaoan Xie, Lingjing Lingjing, Yujia Zheng, Yu Yao, Zeyu Tang, Eric P Xing, Guangyi Chen, and Kun Zhang. Smartclip: Modular vision-language alignment with identification guarantees. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 29780–29790, 2025.

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.

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*, pp. 633–641, 2017.

Chong Zhou, Chen Change Loy, and Bo Dai. Extract free dense labels from clip. In *European conference on computer vision*, pp. 696–712. Springer, 2022.

Ziqin Zhou, Yinjie Lei, Bowen Zhang, Lingqiao Liu, and Yifan Liu. Zegclip: Towards adapting clip for zero-shot semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11175–11185, 2023.

A Implementation and Additional Results

A.1 Candidate Definitions

The primary ViT-B/16 selector compares seven recorded candidates: the unmasked scorer U, two fixed feature-calibrated policies P_A and P_B, a policy P_{ad} that replaces P_B's Cityscapes branch with the prior adapter, DirectCDD, a text-only selective mask, and a top-200 text mask. Table 10 expands the two policy names into their actual mask, text, and gating choices. P_A and P_B correspond to the archived v3 and v4 configurations; P_{ad} corresponds to v5. Each is fixed before comparison in the released evaluator on train1000.

Dataset	P_A	P_B
VOC20	Hard prior / CLIP text / margin	Unmasked
Context59	Soft prior / SmartCLIP text / margin	Unmasked
COCO-Object	Soft prior / SmartCLIP text / replace	Same as P_A
COCO-Stuff	Soft prior / SmartCLIP text / replace	Same as P_A
ADE20K	Soft prior / SmartCLIP text / margin	Unmasked
Cityscapes	Dense-mask network / CLIP text / margin	Hard prior / CLIP text / margin

Table 10: Expanded candidate policies. "Prior" denotes the frozen SmartCLIP mask, "margin" uses Eq. (9), and "replace" always uses the masked score. The dense-mask network uses $\tau_m = 2$.

The transferred-mask menu varies soft versus thresholded masks, original CLIP versus SmartCLIP text embeddings, and unconditional replacement versus the margin gate. The feature policy P_A also includes the text-only references and the Cityscapes dense-mask network. The restricted policy P_B compares the unmasked scorer with hard-prior/original-text/margin, soft-prior/SmartCLIP-text/replace, and soft-prior/SmartCLIP-text/margin, requiring at least 0.50 mIoU points of improvement on training features. No SmartCLIP weights are updated in these transferred-mask candidates.

Text-only references. For the top-200 reference, $m_{q,d}$ is one on the 200 largest entries of $|t_q|$ and zero elsewhere. The selective reference defines $v_{q,d} = |t_{q,d} - \frac{1}{Q-1}\sum_{j\neq q} t_{j,d}|$ and rescales each row of v to $[0,1]$, giving $m_q = 1 + 0.8\,\mathrm{minmax}(v_q)$. Both references mask the visual and text vectors in the cosine score. The selective reference interpolates unmasked and masked class scores with $g_p = (1 - \mathrm{minmax}_p\,\Delta(S^0, p))^{1/2}$, where min–max normalization is within each image. It activates this interpolation only when the mean nearest-neighbor cosine similarity among query embeddings is at most 0.88; otherwise it returns the unmasked scores. Mask construction uses text alone, although selecting either reference in the primary policy uses labels.

A.2 DirectCDD Fold Procedure

DirectCDD divides the ordered 1,000-image training bundle into two contiguous 500-image folds. It builds the mask in Eq. (7) on one fold and evaluates interventions on the other, then swaps the folds. For each intervention, a competitor class is retained only if its mask has at least one suppressed channel and applying the intervention to that class alone increases mIoU in both directions. Let $\mathcal{K}$ be the resulting class set.

The margin intervention uses the masked score vector only where $\widehat{y}^0(p) \in \mathcal{K}$ and $\Delta(S^m, p) > \Delta(S^0, p)$. The competitor-only intervention keeps the unmasked vector except for its original top class:

$$\widetilde{S}_c(p) = \begin{cases} S^m_c(p), & c = \widehat{y}^0(p) \in \mathcal{K}, \\ S^0_c(p), & \text{otherwise.} \end{cases} \tag{11}$$

This operation replaces that score; cosine normalization does not require the replacement to be smaller. We compute the mean mIoU gain over the two folds using the retained classes, choose the better intervention, and rebuild the mask on all 1,000 images. An empty retained set or non-positive best mean gain returns the unmasked scorer. The primary ViT-B/16 DirectCDD checkpoints for Context59, COCO-Stuff, and ADE20K all select the competitor-only intervention. The ViT-L/14 policy additionally requires a 0.50-point mean fold gain. The folds support both class/operator selection and its reported gain, rather than a second, independent estimate of the entire selection procedure.

A.3 Training-Set Policy Audit

The primary train1000 candidate scores and selected branches are reported in Table 2. Duplicate scores correspond to identical deployed branches or differences below reporting precision.

Earlier cross-evaluator audit. An earlier recorded variant (v11) screened DirectCDD with feature-bundle gains on the validation split. In that selector, the feature-bundle gain had to match or exceed the official train1000 gain over its reference, after rounding to two decimal places. The feature-bundle gains were -0.05, $+0.02$, and -0.41 points for Context59, COCO-Stuff, and ADE20K; the corresponding train1000 gains were $+0.14$, $+0.22$, and $+0.11$. This substituted P_A, P_B, and P_B, respectively, yielding the diagnostic row below. Because the audit uses validation labels, it is reported separately from the primary train1000-only selection result (v10).

Selection	VOC20	Context59	COCO-Object	COCO-Stuff	ADE20K	Cityscapes	Avg5	Avg6
Training selection	80.92	35.93	33.45	24.27	16.42	32.28	37.96	37.21
Validation-feature audit	80.92	36.05	33.45	24.35	16.66	32.28	38.05	37.29

Table 11: Recorded official-evaluator results for the two selection procedures. The second row is a diagnostic with validation-informed branch selection.

The ViT-L/14 fold gains, validation gains, and branch decisions are reported in Table 6.

A.4 Controlled Feature-Bundle Protocol

The controlled evaluator reuses exported dense features and applies the same label mappings and score-map evaluation to every configuration within a comparison. Its feature extraction and resizing differ from the released ClearCLIP evaluator. The six-dataset policy below uses transferred masks on five datasets and a separately trained dense-mask network on Cityscapes.

Method	VOC20	Context59	COCO-Object	COCO-Stuff	ADE20K	Cityscapes	Avg.
MaskCLIP	53.58	23.02	15.10	14.73	10.85	22.99	23.38
SCLIP	64.70	26.63	19.29	17.12	11.03	12.99	25.29
ClearCLIP	75.10	31.48	22.46	20.74	14.27	15.20	29.87
Training-free gated mask	72.70	29.70	27.84	19.60	13.40	15.21	29.74
Training-free selective mask	75.10	31.48	27.84	20.74	14.27	15.20	30.77
Controlled mask policy	**75.98**	**31.67**	**28.82**	**21.01**	**14.43**	**24.53**	**32.74**

Table 12: Controlled feature-bundle results under the shared project evaluator (mIoU, %). These internal reproductions support method ablations. Avg. averages the six datasets.

The mean improvement over the controlled ClearCLIP baseline is 2.86 points, computed from unrounded scores. This policy and the prior-adapter plug-in study use different Cityscapes checkpoints: 24.53 and 21.21 mIoU, respectively.

A.5 Prior-Adapter Diagnostics

The six-dataset adapter validation scores and training-image diagnostics are reported in Table 7.

The Cityscapes adapter's official-evaluator mIoU is 31.05 on train1000 and 32.28 on validation; the unmasked reference scores 28.68 and 30.08. Its controlled feature-bundle score is 21.21. These are evaluations of the same adapter with different feature and inference pipelines. The diagnostic count uses the first 200 training images. An additional adapter trained with an explicit pressure penalty produces 28,005 corrections, 40,148 regressions, and 16.80 controlled mIoU despite a larger reduction in mask mass. This comparison illustrates that mask contraction alone does not identify an effective mask.

A.6 Dense-Mask Network and Additional Comparisons

The controlled Cityscapes dense-mask network maps the original text embedding through LayerNorm, a D-to-D linear layer, GELU, and a zero-initialized D-to-D output layer, followed by a

sigmoid. It is trained directly with class-weighted pixel cross-entropy and has no SmartCLIP prior. Its selected checkpoint uses one epoch, learning rate 10^{-3}, and seed zero.

Relative score temperature. For this branch, the masked scores are divided by τ_m while the unmasked scores retain unit temperature, before the margin gate is applied. The training set chooses τ_m from $\{0.75, 1, 1.25, 1.5, 2, 3\}$, selecting two. A common positive scaling would leave argmax predictions and the relative margin gate unchanged; this branch-specific scaling changes which score vector is selected. The same seed-zero checkpoint obtains 22.66 mIoU at $\tau_m = 1$ and 24.53 at $\tau_m = 2$.

The per-dataset configuration comparison is reported in Table 8. Its temperature row holds the checkpoint fixed; the unmasked row removes both masking and SmartCLIP text substitution.

The cumulative text-only ablations appear in Table 9; the mask and gate definitions are given in Appendix A.1.

Method or branch	Cityscapes mIoU
ClearCLIP	15.20
Training-free top-k	16.23
Dense mask network, seed 0, $\tau_m = 1$	22.66
MaskCLIP	22.99
Dense mask network, seed 0, $\tau_m = 2$	**24.53**

Table 13: Cityscapes feature-bundle branches (mIoU, %). The two dense-mask rows use the same checkpoint and differ only in relative score temperature.

The dense-mask network and the SmartCLIP prior adapter differ in parameterization and initialization; their absolute scores characterize candidate behavior rather than isolating the causal contribution of the pretrained prior.