# DECOUPLING CONTENT AND SPATIAL SUPPORT IN FROZEN VISION LANGUAGE READOUTS

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Dense prediction with a frozen vision–language encoder depends on both the content supplied to its text-aligned head and the spatial support used to aggregate patches. We investigate whether the same representations are suitable for both roles by generalizing the pretrained pooling head into a support-set decoder with independently selectable content and support depths. An exhaustive $13 \times 13$ depth map on SigLIP 2 B/16 reveals distinct preferences: final-layer content and pre-contextual support. Holding content fixed, replacing pre-contextual support with final-layer support reduces VOC20 mIoU by 6.4 points. Statistical controls clarify the distinction: matching content from an earlier layer to the native head input's moments recovers most of its readout loss, while matching final-layer support features to pre-contextual moments substantially narrows the support gap. These interventions distinguish compatibility with the frozen head from the feature geometry used for spatial aggregation. Cross-encoder controls characterize where the shallow-support preference holds. Applying this separation yields SigMAP, a training-free decoder that combines final-layer content with pre-contextual support and improves on identity support across five datasets. Together, these findings identify content compatibility and spatial support quality as distinct criteria for selecting representations under a frozen readout.

## 1 INTRODUCTION

Frozen vision–language encoders provide semantic features, but dense prediction also requires deciding which patches should contribute to each location. The feature representation supplied to a text-aligned readout and the similarities used for grouping patches serve different roles. Encoders with a pooling head make this distinction especially consequential because patch features reach the deployed text space through a learned aggregation interface. SigLIP 2 (Tschannen et al., 2025) offers a useful case study: pooling ground-truth regions through its frozen head classifies 95.88% of VOC20 validation regions correctly (§2.1), motivating a test of how content and aggregation jointly affect dense readout.

Changing the representation used throughout a decoder changes both roles at once. We make their effects separately measurable by generalizing the pooling head from one probe-attention row to an arbitrary row-stochastic support matrix $S$. The resulting support-set decoder exposes content depth $c$ and support depth $s$ as independent axes under the same frozen head. Holding one axis fixed while varying the other distinguishes compatibility with the text interface from the quality of the affinities used for aggregation.

An exhaustive $13 \times 13$ map on SigLIP 2 B/16 reveals that content and support favor different depths under the frozen pooling interface. The best measured configuration combines final-layer content with depth-0 pre-contextual support. Final-layer pairwise similarity provides worse spatial support (Fig. 1); moving content to an earlier layer also reduces readout quality. With the same depth-0 support, reading content from depth 10 instead of the head's native input drops the development-screen score from 87.3 to 22.3 mIoU (§2.3). A representation's usefulness for aggregation therefore does not imply compatibility with the frozen text interface.

Statistical controls clarify this distinction using different reference representations for the two roles. Matching depth-10 content to the native final-layer input's moments recovers 87.1% of the content loss on VOC20 validation. Matching depth-12 support to depth-0 moments recovers 70.4% of the



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

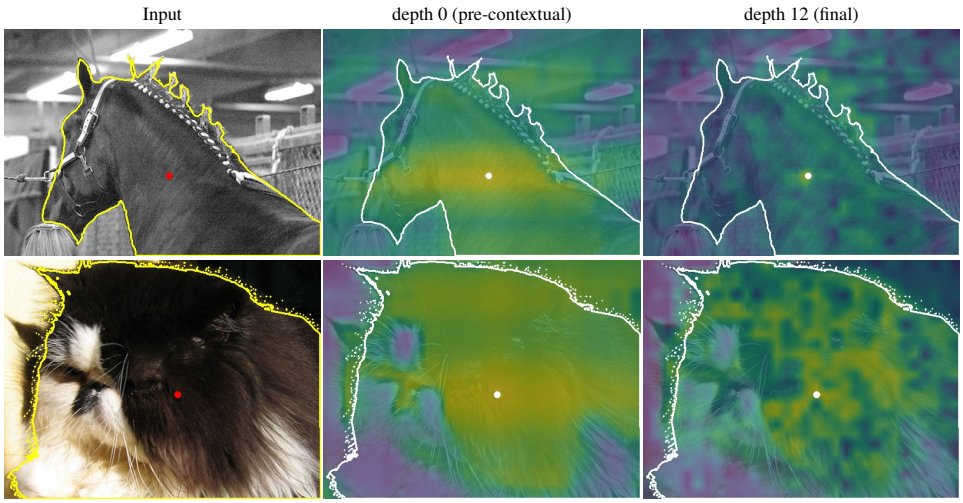


Figure 1: Pre-contextual and final-layer affinities on the same images. In these examples, depth-0 affinity remains object-local, whereas depth-12 affinity spreads beyond object boundaries. The decoder evaluates these representations as spatial support while retaining final-layer content (§2.2).

support gap. These recoveries show that feature statistics matter to both roles, while distinguishing the head's expected content statistics from the geometry used for spatial aggregation. They also explain why poor shallow-layer readout alone does not establish that shallow features lack useful information.

Further controls characterize the phenomenon's scope. Removing the highest-norm tokens from support computation or projecting out a probe-identified global direction does not recover the gap. Across the encoders tested, the support penalty appears on SigLIP 2, SigLIP v1, and CLIP-RN50 and is absent on AIMv2-LiT and CoCa under the same type of support-depth measurement. Section 2.5 relates this descriptive association to the models' alignment routes. Within SigLIP 2, the shallow-to-final support penalty holds across five datasets, although intermediate depth trajectories need not be monotonic.

Based on these results, we retain final-layer content and construct support from the encoder's own pre-contextual patch-embedding output. The resulting SigMAP decoder raises the five-set mean from 35.24 for identity support to 42.82, with the content path fixed. It uses one decoder configuration across datasets and needs neither training nor an external support model. A separate SigLIP v1 evaluation also improves the prior adaptation (Zhou et al., 2026) under its reported protocol.

**Contributions.** (1) We introduce a support-set decoder that makes content depth and support depth independently measurable under a frozen pooling head. (2) Through an exhaustive depth map and statistical controls with different reference representations, we distinguish content compatibility at the head interface from spatial support quality. (3) We derive SigMAP from this separation and demonstrate its benefit over identity support and alternative spatial supports on five datasets, using the encoder's own pre-contextual representation.

## 2 Analysis

We use the frozen pooling interface as a measurement instrument to separate two questions: which features can the head read as content, and which features supply useful spatial support? The depth map measures their distinct preferences, and statistical controls examine how each role responds to correction toward its reference representation.

### 2.1 SUPPORT-SET DECODER AS A MEASUREMENT INSTRUMENT

Let a ViT encoder produce patch tokens $X \in \mathbb{R}^{N\times d}$ ($N = 576$ at 384px, patch 16). SigLIP 2 (Tschannen et al., 2025) pools with a MAP head: a learned probe attends over $X$, and the pooled vector is mapped into the text space by the frozen value/output path and MLP. In an oracle diagnostic, pooling ground-truth regions through the frozen head classifies 95.88% of the 2,136 VOC20 validation regions correctly. This region-classification accuracy is not a segmentation score: it shows that useful semantic information is accessible when the aggregation regions are supplied. The support-set decoder below measures how the choice of aggregation affects dense prediction. Throughout, depth 0 denotes the patch-embedding output (linear patch projection plus positional embedding), before any transformer block. Depth $k$ for $1 \le k \le 11$ denotes the output of $k$ transformer blocks. Depth 12 includes all 12 blocks and the encoder's post-layernorm, i.e. the MAP-head input. Both decoder axes use this indexing.

**Generalizing the head.** The probe attention row specifies which patches are summarized together for the image. We replace that row with a row-stochastic matrix $S \in \mathbb{R}^{N\times N}$, where row $i$ specifies the support set for patch $i$. For support features $X^{(s)}$, we cast tokens to FP32, normalize each token to unit L2 norm, and compute cosine affinity $A_{ij} = \langle \hat{x}_i^{(s)}, \hat{x}_j^{(s)} \rangle$. Define the off-diagonal row maximum $m_i = \max_{k\neq i} A_{ik}$ and the binary mask $M_{ij} = \mathbf{1}\{j = i \text{ or } A_{ij} \ge \tau m_i\}$. The support matrix is

$$S_{ij}^{(s)} = \frac{M_{ij}\exp(A_{ij}/T)}{\sum_{k=1}^{N} M_{ik}\exp(A_{ik}/T)}, \tag{1}$$

where $\tau$ is a row-relative threshold coefficient calibrated to the reference mean support degree, and $T > 0$ targets the reference mean row entropy. Self-edges are always retained, so every row has nonempty support. Appendix C.1 specifies calibration and execution. We then pass all rows through the same frozen head:

$$Z = \mathrm{Head}\Big(S^{(s)}\, V(X^{(c)})\Big) \;\longrightarrow\; N \text{ per-patch embeddings} \;\longrightarrow\; N \times C \text{ logits}. \tag{2}$$

Here $X^{(c)}$ are the patch tokens at content depth $c$, $V$ is the head's value projection, and Head comprises the frozen output projection followed by the residual MLP block, $Y + \mathrm{MLP}(\mathrm{LN}(Y))$. Here $Y$ denotes the output-projected aggregate. The learned probe and query/key projections are bypassed; Appendix C.2 gives the complete operation. $S^{(s)}$ is constructed from the pairwise cosine similarities of the patch tokens at support depth $s$ and converted to a row-stochastic matrix by the row-relative mask and temperature calibration of Appendix C.1. Every support source is calibrated toward the same target mean row entropy as the reference final-layer cosine support, using features alone with no label or score in the loop. For support constructed from the host encoder's own representations, no trainable parameters are added and no additional encoder forward is required. The identity $S{=}I$ recovers plain per-patch value readout; other choices of $S$ change only which patches support each decoded location. This turns support into a controlled variable while keeping the pretrained content path and pooling interface fixed.

### 2.2 SPATIAL SUPPORT DEGRADES WITH DEPTH

With support and content decoupled, we evaluate every $(c, s) \in \{0, \ldots, 12\}^2$ pair (Fig. 2); the mIoU obtained with content fixed at the final layer is the *support quality* of depth $s$. The best cell on the development screen is $(c{=}12, s{=}0)$ at 87.33 mIoU. Holding content at the head input, support quality decreases from 87.33 to 80.61 as support moves from depth 0 to depth 12; the same configuration scores 86.85 on VOC20 validation (Table 4).

On validation the same ordering holds: support quality falls from 86.85 to 80.49, a shallow-to-final penalty of 6.4 points. Depth 12 lies within 0.25 of depth 11. The corresponding drops are 3.49, 2.03, 2.26, and 3.32 points on Context59, COCO-Stuff, ADE20K, and Cityscapes (Table 2). Figure 3 shows the changes in predictions when support depth varies with final-layer content fixed. On VOC20, depth-12 support even falls below raw pixel-color support (80.49 vs. 84.84).

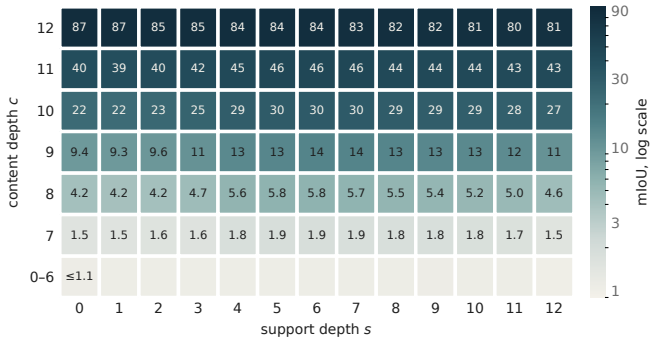


Figure 2: Content–support depth map on the development screen (mIoU, log scale). The best measured configuration combines final-layer content ($c$=12) with pre-contextual support ($s$=0). Moving content to shallow layers does not preserve this advantage, revealing why the two depths must be evaluated independently. Content depths 0–6 are collapsed because all score $\leq 1.1$ mIoU.

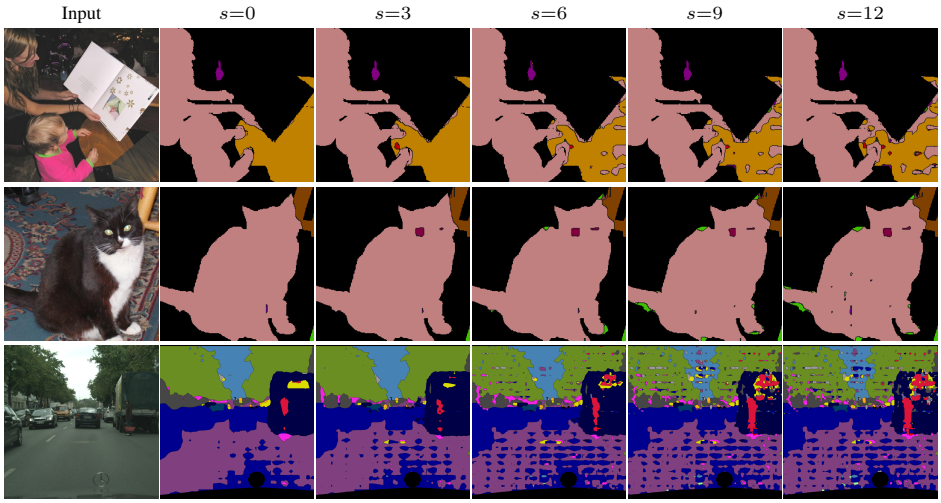


Figure 3: Predictions at different support depths with content fixed at the depth-12 MAP-head input. Each row shows an image followed by predictions using support depths 0, 3, 6, 9, and 12. The examples illustrate changes in spatial coherence; aggregate support quality is reported in Table 2. Support depths 2–7 do not reach the target entropy (Appendix C.1).

### 2.3 Why shallow content fails at the frozen interface

The content axis exposes an interface constraint on the shallow-support solution. Under depth-0 pre-contextual support, reading content from depth 10 instead of the head's native input reduces development-screen performance to 22.3 mIoU, from depth 8 to 4.2, and lower depths approach zero mIoU. This drop is largely a calibration effect at the head interface. On validation, depth-10 content scores 22.9 against 86.85 for the native input; matching its features to the native input recovers 52.3% of the drop by token-norm matching, 75.4% by per-channel affine matching, and 87.1% (to 78.6) by whitening-recoloring to the native input's second moments (Table 1; Appendix C.3). Shallow features are useful as support. Keeping content at the head's native input avoids interface correction; earlier content can recover much of its readout quality after statistical matching.

### 2.4 What explains the support degradation?

We next apply statistical correction to the support axis, using depth 0 as the reference while keeping final-layer content fixed. Matching depth-12 features to depth-0 moments before constructing $S$ recovers 70.4% of the gap on VOC20 validation and 39–75% on the other four sets (Table 7), while

Table 1: Recovery from content- and support-axis interventions on VOC20. Content interventions and support whitening use validation; token removal, global-direction projection, and PCA removal use the development screen. Recovery percentages use the corresponding split's baseline gap.

| Axis | Intervention | Recovered |
|---|---|---|
| Content | Token-norm matching | 52.3% |
| Content | Per-channel affine matching | 75.4% |
| Content | Whitening–recoloring to depth-12 moments | 87.1% |
| Support | Excising the top-8 high-norm tokens | none ($-0.02$ mIoU) |
| Support | Projecting the global direction out | none ($80.61 \rightarrow 80.59$) |
| Support | Removing the top-16 PCA directions | 53.1% |
| Support | Whitening to depth-0 second moments | 70.4% |

Table 2: Support-source comparison on five validation sets (mIoU). All sources use the same frozen-head readout and per-source calibration.

| Support source | VOC20 | C59 | Stuff | ADE | City | Mean |
|---|---|---|---|---|---|---|
| Depth 12 | 80.49 | 33.26 | 22.52 | 21.48 | 38.88 | 39.33 |
| Gaussian prior | **86.99** | 31.34 | 21.47 | 19.42 | 33.60 | 38.56 |
| Mean-RGB | 84.84 | 31.01 | 21.27 | 18.68 | 36.93 | 38.55 |
| Bilateral | 86.54 | 34.91 | 24.03 | 22.13 | 37.71 | 41.06 |
| DINOv2 final | 83.64 | 35.00 | 24.33 | 22.43 | 39.70 | 41.02 |
| DINOv2 patch embed | 84.16 | 32.49 | 22.43 | 21.14 | 39.23 | 39.89 |
| Depth 0 (SigMAP) | 86.85 | **36.75** | **24.55** | **23.74** | **42.20** | **42.82** |

still trailing depth 0. Together with §2.3, this tests two distinct corrections: content toward the native head input, and support toward pre-contextual geometry. Removing the top-16 principal directions also recovers 53% of the development-screen gap. The tested token and direction controls leave the shallow-over-deep ordering unchanged: neither removing the highest-norm tokens from the support matrix nor projecting out a probe-identified global direction recovers the gap. These interventions indicate sensitivity to second-moment structure without identifying a unique mechanism. Matching uses regularized covariance estimates (Appendix C.3), and calibration is re-derived for each transformed support source. Table 1 identifies the evaluation split for each intervention; Appendix A specifies the support controls.

**Is the ordering a preference for smoother supports?** Token uniformity in pure attention networks (Dong et al., 2021) motivates examining support concentration; the theoretical result assumes no skip connections or MLPs. Temperatures are calibrated on development-image features, without labels or segmentation scores, to match the reference mean row entropy (2.56 nats; Appendix C.1). Depths 0 and 12 attain the target on the calibration data, while intermediate depths 2–7 reach the search bound with lower entropies. The shallow-to-final ordering persists under this calibration. Mean row entropy alone is therefore insufficient to characterize support quality: supports with matched mean entropy can still differ in their spatial structure.

**Comparison with alternative spatial supports.** The encoder's intrinsic depth-0 support achieves the highest five-set mean in Table 2: 42.82, compared with 41.06 for bilateral support and 41.02 for DINOv2 final-layer affinity. It has higher reported mIoU than both sources on all five datasets, using the same frozen decoder and per-source calibration. This makes pre-contextual support a practical choice available within the encoder's own forward pass. A position-only Gaussian prior is competitive on VOC20, and raw color is also strong there, but both fall below depth-12 support on each of the other four datasets. Combining position and color with a bilateral kernel improves robustness; the largest margin in favor of intact depth-0 support occurs on Cityscapes. External DINOv2 affinity improves over SigLIP 2 depth-12 support on every dataset and shows a different depth preference: its final-layer affinity is higher than its patch-embedding affinity on four of five datasets. The shallow-support preference is therefore not universal across the tested representations. Table 2 evaluates the utility of intact support sources; it does not isolate the positional and visual components within depth 0 or establish which component dominates its utility.

Table 3: Scope across alignment routes on VOC20 validation. Support gap is shallow-minus-deep support quality under each encoder's frozen readout; positive values indicate worse deep support.

| Encoder | deployed route | support gap |
|---|---|---|
| SigLIP 2 B/16 | bottleneck | $+6.4$ |
| SigLIP v1 B/16 | bottleneck only | $+7.0$ |
| CLIP-RN50 | bottleneck only | $+10.7$ |
| AIMv2-LiT | post hoc, frozen encoder | $-1.0$ |
| CoCa | shared text tower + caption decoder | $-1.9$ |

### 2.5 Scope across alignment routes

We use *bottleneck-aligned* to describe models in which the deployed image–text alignment is trained through the pooling interface into the visual encoder. This operational grouping describes the alignment route; auxiliary pretraining objectives may also affect patch features. We compare it with post-hoc alignment of a frozen encoder and a shared text tower trained through captioning cross-attention.

The support penalty is not a generic property of attention pooling. It is absent in AIMv2-LiT (Fini et al., 2025), whose deployed text alignment is fitted post hoc to a frozen visual space, and in CoCa (Yu et al., 2022), whose contrastive text embedding comes from the same text tower that captions through cross-attention to the patch tokens. It appears on CLIP-RN50 (Radford et al., 2021), a convolutional encoder whose AttentionPool2d is the only route of its contrastive objective, on SigLIP v1 (Zhai et al., 2023), trained with the sigmoid loss alone (VOC20 validation gap $+7.0$, with a positive shallow–deep gap on every set; Appendix B), and on SigLIP 2, whose deployed sigmoid objective (Zhai et al., 2023) likewise reaches the patches only through the MAP head. Its separate LocCa-style captioning decoder is used only during pretraining and is absent from the released model (Wan et al., 2024; Tschannen et al., 2025). The two captioning decoders therefore play different roles: CoCa's shares the text tower that produces its deployed contrastive embedding, whereas SigLIP 2's is a separate head that is not part of the released model. The remaining SigLIP 2 auxiliary objectives, self-distillation and masked prediction on the patch tokens (Tschannen et al., 2025), carry no text alignment. Table 3 summarizes the VOC20 validation shallow-minus-deep support gap under each encoder's own frozen readout (measurements in Appendix B); across the tested encoders, the sign of the gap is associated with whether the deployed alignment reaches the patches only through the pooling bottleneck.

### 2.6 Applying the Separation: SigMAP

We therefore keep semantic content on the final-layer path expected by the frozen head and construct spatial support outside the degraded contextualized stream. Depth 0 is the canonical intrinsic source because it is pre-contextual, requires no learned selector, and is available from the same encoder forward. The support-source controls in Table 2 further show that this choice is more consistent across datasets than hand-crafted position/color priors or external DINOv2 affinity, while requiring neither an external foundation model nor a second encoder forward.

**SigMAP in one frozen forward per crop.** Given the input image, (1) extract pre-contextual patch embeddings $X^{(0)}$ and final-layer features $X^{(F)}$ from the same encoder pass; (2) construct $S^{(0)}$ from pairwise cosine similarities of $X^{(0)}$ using the row-relative mask and calibrated temperature of Eq. 1; and (3) decode $S^{(0)}V(X^{(F)})$ through the unchanged MAP value/output path and MLP. SigMAP changes only the spatial support supplied to the frozen pooling interface: semantic content remains on the final-layer path, while support comes from the encoder's intrinsic pre-contextual representation. SigMAP adds no trainable parameters and requires neither an external model nor an additional encoder forward. Support calibration is derived on development data and shared across all five datasets; prompt templates are selected per dataset on its 40-image training subset (Appendix C.4).

Table 4: Same-backbone comparison on SigLIP 2 B/16 (mIoU). $S{=}I$ exposes the value path per patch; SigMAP replaces identity support with intrinsic depth-0 support while keeping the content path fixed.

| Readout | VOC20 | C59 | Stuff | ADE | City | 5-set |
|---|---|---|---|---|---|---|
| $S{=}I$ (value path per patch) | 75.4 | 29.4 | 20.2 | 18.5 | 32.6 | 35.24 |
| SigMAP | 86.85 | 36.75 | 24.55 | 23.74 | 42.20 | 42.82 |

# 3 Experiments

## 3.1 Experimental Setup

We use the frozen `google/siglip2-base-patch16-384` backbone (Tschannen et al., 2025) with sliding-window inference (crop 384, stride 96, bf16); Cityscapes is evaluated at its native $1024 \times 2048$ resolution. We report VOC20 (Everingham & Winn, 2012), Context59 (Mottaghi et al., 2014), COCO-Stuff (Caesar et al., 2018), ADE20K (Zhou et al., 2017), and Cityscapes (Cordts et al., 2016). Unless otherwise stated, the SigLIP 2 headline results are evaluated on the full validation set with the single configuration of §2.6; control tables explicitly state their alternative support source or backbone. The five-set mean gives equal weight to each dataset's mIoU. One encoder forward is used per crop, not per full image; sliding-window inference requires multiple crop forwards. Training-free denotes the absence of task-specific encoder or head weight updates. Feature-based support calibration and prompt selection are specified in Appendix C.4; diagnostic classifiers do not enter segmentation inference.

All design choices were made on fixed development subsets disjoint from the evaluation data; validation was evaluated once per adopted design choice under pre-specified adoption criteria. Sample identities are pinned by SHA-256, and per-image intersection/union arrays are stored. Implementation and diagnostic details are provided in Appendices C.1 and C.4.

## 3.2 Results

**Main same-backbone results.** Table 4 compares two frozen-head dense readouts on SigLIP 2 B/16. Exposing the pretrained value path per patch with $S{=}I$ reaches 35.24 mean mIoU. With the same content pathway, SigMAP replaces identity support with depth-0 pre-contextual support and reaches 42.82, a $+7.58$-point gain. This comparison measures the benefit of the chosen support within the fixed decoder.

**Off-image assignment analysis.** We define the off-image assignment rate as the fraction of evaluated non-void pixels whose predicted class is absent from that image's ground-truth class set. If deep support aggregates patches across unrelated regions, their mixed content may favor classes absent from the image. We therefore test whether deep support increases off-image assignments relative to depth 0. This holds on all five validation sets (Fig. 4), where the off-image assignment rate falls by 19–47% of its depth-12 value after support relocation. On every set, depth-12 support whitened to depth-0 second moments lies between the two endpoints, closing 47–79% of the off-image assignment gap while recovering 39–75% of the mIoU gap (Table 7, Appendix D); the direction of each whitened control was specified before evaluation. Whitening shifts both measures toward the depth-0 endpoint on every dataset.

**Protocol-matched comparison (SigLIP v1).** For a separate within-family replication, we evaluate SigMAP on the SigLIP v1 B/16 backbone used by the prior SigLIP adaptation (Zhou et al., 2026), with the same resolution and prompt templates (Table 5); these values are not directly comparable to the SigLIP 2 results in Table 4. This prior-work-matched protocol is distinct from the cross-encoder scope protocol used in Table 3 and Appendix B. On the five-set mean over VOC20, Context59, COCO-Stuff, ADE20K, and Cityscapes, the $S{=}I$ readout reaches 30.28 against 26.9 for the prior adaptation. SigMAP reaches 31.93, a further $+1.65$ over the same frozen value-path readout.

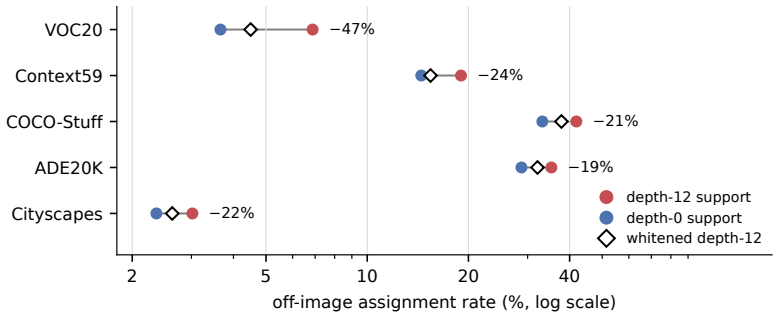


Figure 4: Off-image assignment under depth-12, whitened depth-12, and depth-0 support across five validation sets (log scale; lower is better).

Table 5: Protocol-matched comparison on the separate SigLIP v1 B/16 backbone (five-set mean mIoU). These values are not directly comparable to the SigLIP 2 results in Table 4; prior rows are from Zhou et al. (2026).

| Method | 5-set mean mIoU |
|---|---|
| Vanilla readout | 22.1 |
| ClearCLIP transplant | 2.3 |
| Prior SigLIP adaptation | 26.9 |
| $S{=}I$ control | 30.28 |
| **SigMAP** | **31.93** |

## 4 Related Work

**Modifying CLIP's last layers.** MaskCLIP (Zhou et al., 2022), SCLIP (Wang et al., 2024), GEM (Bousselham et al., 2024), NACLIP (Hajimiri et al., 2025), ClearCLIP (Lan et al., 2024a), ProxyCLIP (Lan et al., 2024b), and ResCLIP (Yang et al., 2025) adapt CLIP's attention or patch-correlation structure for dense inference. These methods modify dense readout from CLIP; several alter attention-based aggregation within the contextualized stream, while ProxyCLIP imports affinity from an external encoder, corresponding to an external $S$. We generalize a frozen pooling head to expose content and support as independent experimental variables. This provides a common readout for comparing native and external support sources and for measuring depth effects with the other role fixed. Table 5 includes the ClearCLIP transplant under the protocol-matched SigLIP-v1 comparison.

**Earlier features for dense prediction.** Text-based decomposition attributes CLIP's representation to layers, attention heads, and image patches (Gandelsman et al., 2024); LHT-CLIP analyzes visual discriminability and semantic alignment across layers, heads, and tokens (Zhou et al., 2026). Several lines of work use information beyond CLIP's final representation. CLIPer (Sun et al., 2025) explicitly combines features from multiple depths, while Perception Encoder (Bolya et al., 2025) shows that strong spatial representations can reside in intermediate layers. CAT-Seg (Cho et al., 2024) aggregates image–text costs, and CorrCLIP (Zhang et al., 2025) reconstructs patch correlations using off-the-shelf foundation models. Our depth map separates the role in which a representation is useful: pre-contextual features supply strong support on SigLIP 2, while content must be evaluated for compatibility with the frozen head. Statistical controls further show recovery using different reference representations for the two roles. Related support sources include the bilateral position$\times$color affinity of dense CRFs (Krähenbühl & Koltun, 2011), evaluated through the same decoder in Table 2, and CLIP-DINOiser (Wysoczańska et al., 2024), which learns DINO-like affinity from CLIP features under DINO guidance. SigMAP uses the encoder's own pre-contextual patch-embedding output without training or a guiding model.

**Registers and anomalous tokens.** High-norm artifact tokens have been linked to global information storage and dense-prediction artifacts (Darcet et al., 2024). Related CLIP adaptations address

global or anomalous patch behavior, including CLIPtrase (Shao et al., 2024) and SC-CLIP (Bai et al., 2025). We test the effect of excluding the highest-norm tokens from the support matrix and of removing a probe-identified global direction. Neither tested intervention recovers the support gap on SigLIP 2 (Appendix A). These results concern the specified interventions, rather than ruling out every role of global information.

**Distillation on SigLIP 2.** SPAR (Kombol et al., 2026) distills a single-pass student from a densely strided sliding-window teacher on SigLIP 2, using the pooling head's value path. SigMAP uses this frozen value path with spatial support selected from the content–support analysis.

## 5 Conclusion

Generalizing a frozen pooling head into a support-set decoder makes content compatibility and spatial support quality independently measurable. On SigLIP 2 B/16, the exhaustive depth map favors final-layer content with pre-contextual support. Statistical controls connect these preferences to different reference representations: content is corrected toward the native head input, while support is corrected toward depth-0 moments. These interventions recover 87.1% of the depth-10 content loss and 70.4% of the VOC20 support gap (39–75% on the other sets), demonstrating that feature statistics affect both roles. Cross-encoder controls characterize the scope of the support-depth effect.

SigMAP implements this separation directly: it keeps final-layer content unchanged and supplies support from the encoder's own pre-contextual representation. On SigLIP 2 B/16, it reaches 42.82 mean mIoU across five datasets, improving the same-backbone $S{=}I$ value-path readout by 7.58 points without training, an external support model, or an additional encoder forward per crop. Compatibility with the frozen content path and the quality of spatial support should be measured separately.

**Limitations.** The alignment-route comparison spans five encoders with different architectures, data, and training objectives. Its descriptive association does not isolate the route as a cause, and attention pooling alone is not a sufficient predictor. Statistical matching recovers only part of the support gap, and the negative token/direction controls exclude only the tested interventions. Depth 0 combines positional and visual information; the support-source comparison does not establish how much each contributes. The content sweep concerns the frozen ViT interface; CLIP-RN50 contributes only a support-axis comparison because its stages change channel width. Finally, one forward per crop still requires multiple forwards for a full image under sliding-window evaluation.

### Ethics statement

This work analyzes and repurposes publicly released pretrained encoders on standard public benchmarks; no new data involving human subjects was collected.

### Reproducibility statement

All design decisions use development data disjoint from reported evaluations. Validation is evaluated once per retained module under pre-specified adoption criteria, with sample hashes and per-image intersection/union arrays stored alongside the experiment artifacts. Code, prompt templates, sample identifiers, calibration derivations, and source records for reported numbers will be released; the five-set headline evaluation re-derives bit-identically from the current evaluation code.

### The use of AI tools

AI assistants supported code drafting, literature-assisted discussion, methodological feedback, interpretation of experimental results, manuscript organization, and language editing. The authors are responsible for the experimental decisions, validation of reported results, source verification, and all final claims.

## REFERENCES

Sule Bai, Yong Liu, Yifei Han, Haoji Zhang, Yansong Tang, Jie Zhou, and Jiwen Lu. Self-calibrated CLIP for training-free open-vocabulary segmentation. *IEEE Transactions on Image Processing*, 34:8271–8284, 2025.

Daniel Bolya, Po-Yao Huang, Peize Sun, Jang Hyun Cho, Andrea Madotto, Chen Wei, Tengyu Ma, Jiale Zhi, Jathushan Rajasegaran, Hanoona Bangalath, Junke Wang, Marco Monteiro, Hu Xu, Shiyu Dong, Nikhila Ravi, Shang-Wen Li, Piotr Dollár, and Christoph Feichtenhofer. Perception encoder: The best visual embeddings are not at the output of the network. In *Advances in Neural Information Processing Systems*, volume 38, pp. 60884–60937, 2025.

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*, pp. 3828–3837, 2024.

Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. COCO-Stuff: Thing and stuff classes in context. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1209–1218, 2018.

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

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.

Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. In *International Conference on Learning Representations*, 2024.

Yihe Dong, Jean-Baptiste Cordonnier, and Andreas Loukas. Attention is not all you need: Pure attention loses rank doubly exponentially with depth. In *International Conference on Machine Learning*, pp. 2793–2803, 2021.

Mark Everingham and John Winn. The PASCAL visual object classes challenge 2012 (VOC2012) development kit. Technical report, Pattern Analysis, Statistical Modelling and Computational Learning, 2012. URL https://www.robots.ox.ac.uk/~vgg/projects/pascal/VOC/voc2012/htmldoc/index.html.

Enrico Fini, Mustafa Shukor, Xiujun Li, Philipp Dufter, Michal Klein, David Haldimann, Sai Aitharaju, Victor G. Turrisi da Costa, Louis Béthune, Zhe Gan, Alexander Toshev, Marcin Eichner, Moin Nabi, Yinfei Yang, Joshua Susskind, and Alaaeldin El-Nouby. Multimodal autoregressive pre-training of large vision encoders. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9641–9654, 2025.

Yossi Gandelsman, Alexei A. Efros, and Jacob Steinhardt. Interpreting CLIP's image representation via text-based decomposition. In *International Conference on Learning Representations*, 2024.

Sina Hajimiri, Ismail Ben Ayed, and Jose Dolz. Pay attention to your neighbours: Training-free open-vocabulary semantic segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 5061–5071, 2025.

Naomi Kombol, Ivan Martinović, Siniša Šegvić, and Giorgos Tolias. SPAR: Single-pass any-resolution ViT for open-vocabulary segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27579–27589, 2026.

Philipp Krähenbühl and Vladlen Koltun. Efficient inference in fully connected CRFs with Gaussian edge potentials. In *Advances in Neural Information Processing Systems*, volume 24, pp. 109–117, 2011.

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, 2024a.

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*, pp. 70–88, 2024b.

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.

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

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

Tong Shao, Zhuotao Tian, Hang Zhao, and Jingyong Su. Explore the potential of CLIP for training-free open vocabulary semantic segmentation. In *European Conference on Computer Vision*, pp. 139–156, 2024.

Lin Sun, Jiale Cao, Jin Xie, Xiaoheng Jiang, and Yanwei Pang. CLIPer: Hierarchically improving spatial representation of CLIP for open-vocabulary semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 23199–23209, 2025.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, et al. SigLIP 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025.

Bo Wan, Michael Tschannen, Yongqin Xian, Filip Pavetic, Ibrahim Alabdulmohsin, Xiao Wang, André Susano Pinto, Andreas Steiner, Lucas Beyer, and Xiaohua Zhai. LocCa: Visual pretraining with location-aware captioners. In *Advances in Neural Information Processing Systems*, volume 37, pp. 116355–116387, 2024.

Feng Wang, Jieru Mei, and Alan Yuille. SCLIP: Rethinking self-attention for dense vision-language inference. In *European Conference on Computer Vision*, pp. 315–332, 2024.

Monika Wysoczańska, Oriane Siméoni, Michaël Ramamonjisoa, Andrei Bursuc, Tomasz Trzciński, and Patrick Pérez. CLIP-DINOiser: Teaching CLIP a few DINO tricks for open-vocabulary semantic segmentation. In *European Conference on Computer Vision*, pp. 320–337, 2024.

Yuhang Yang, Jinhong Deng, Wen Li, and Lixin Duan. ResCLIP: Residual attention for training-free dense vision-language inference. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 29968–29978, 2025.

Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. CoCa: Contrastive captioners are image-text foundation models. *Transactions on Machine Learning Research*, 2022.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11975–11986, 2023.

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*, pp. 24677–24687, 2025.

Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ADE20K dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 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, 2022.

Jinxin Zhou, Jiachen Jiang, and Zhihui Zhu. Improving visual discriminability of CLIP for training-free open-vocabulary semantic segmentation. *Transactions on Machine Learning Research*, 2026.

## A Additional Support Interventions

All controls in this appendix use the development screen defined in Appendix C.4. Final-layer content and the frozen readout remain fixed while support features or support mixing are changed.

**High-norm-token exclusion.** For each crop, we exclude the eight largest-norm support tokens from both rows and columns of the affinity matrix. Thresholding and row softmax are recomputed on the retained $568 \times 568$ matrix, which is embedded back into the $576 \times 576$ support. Each excluded position receives an identity row, retaining its own value-path prediction. Degree and entropy calibration are re-derived for the retained tokens. The encoder features are unchanged; the intervention isolates these tokens from support mixing. Depth-12 performance changes from 80.61 to 80.59 mIoU ($-0.02$ in Table 1).

**Global-direction projection.** Projecting a probe-identified global direction out of depth-12 support gives 80.59 mIoU when the direction is probed at depth 12 and 80.63 with the block-11 direction, compared with 80.61 before projection. Both remain close to the original endpoint and below depth-0 support at 87.33.

**Mean-direction and principal-component removal.** Per-image mean-token direction removal decreases performance at every tested depth. Removing top-$k$ PCA components also decreases performance at shallow and middle depths. At depth 12, removing the top 16 components instead raises mIoU from 80.61 to 84.18, recovering 53.1% of the gap to depth 0. Together with the validation whitening results in Appendix D, this depth-dependent response connects support quality to multivariate feature statistics.

## B Cross-Encoder Controls

We compare shallow and final support with each encoder's final content and frozen readout held fixed. Table 6 gives the VOC20 validation endpoints underlying the corresponding rows of Table 3. SigLIP 2 uses depth 0 and post-LN depth 12 support; CLIP-RN50 uses the stem and stage 4. For AIMv2-LiT and CoCa, the pairs are $(F, 0)$ and $(F, F)$, where $F$ denotes the encoder's final content depth.

Table 6: VOC20 validation support endpoints. Each row holds the encoder and final content fixed; the gap is shallow minus final mIoU.

| Encoder | Shallow support | Final support | Gap |
|---|---|---|---|
| SigLIP 2 B/16 | 86.85 | 80.49 | $+6.36$ |
| CLIP-RN50 | 56.72 | 46.03 | $+10.69$ |
| AIMv2-LiT | 76.03 | 77.07 | $-1.04$ |
| CoCa | 70.99 | 72.89 | $-1.90$ |

**SigLIP v1 replication.** SigLIP v1 B/16-384 (Zhai et al., 2023) uses calibration re-derived for its own features. The shallow-versus-final support comparison was specified before measurement. The VOC20 validation gap is $+7.02$ mIoU. Across the five validation sets, depth-0 support reaches

37.44 mean mIoU against 34.30 for depth-12 support, with $S{=}I$ at 32.68. The per-dataset shallow-minus-final gaps are 7.02, 2.68, 1.87, 1.79, and 2.35 points for VOC20, Context59, COCO-Stuff, ADE20K, and Cityscapes, respectively. This scope evaluation uses the release protocol; the separate comparison with prior SigLIP adaptations follows their resolution and prompts (Table 5).

**Scope of the comparison.** AIMv2-LiT aligns a frozen visual space post hoc, while CoCa trains its shared text tower through both contrastive alignment and captioning cross-attention. Their endpoint results delimit the shallow-support preference observed in SigLIP 2, SigLIP v1, and CLIP-RN50. For CLIP-RN50, all four adjacent support-depth drops on the development screen are significant under paired bootstrap with Holm correction. Its changing channel widths restrict the common-readout comparison to the support axis. These experiments characterize an alignment-route association across the tested encoders, whose architectures, data, and training objectives also vary.

# C Implementation and Diagnostic Protocol

## C.1 Support construction and calibration

Equation 1 uses a row-relative threshold. We compute cosine affinities in FP32, exclude the diagonal when finding $m_i$, retain edges satisfying $A_{ij} \geq \tau m_i$, and set $M_{ii} = 1$. The original cosine logits are divided by $T$, excluded edges are masked to $-\infty$, and row-wise softmax produces $S$.

Calibration uses the development features specified in Appendix C.4, with 576 patch rows per image. The final-layer reference $(\tau, T) = (0.85, 0.15)$ defines the target mean degree and row entropy (approximately 22.40 and 2.56 nats, respectively). A 20,000-bin reverse-cumulative histogram of off-diagonal ratios $A_{ij}/m_i$ determines $\tau$ by matching the reference mean degree. Ratios are clipped to $[0, 1]$ only for histogram construction; inference uses the original cosine affinities and row-relative thresholds.

With $\tau$ fixed, 22 geometric-bisection steps over $T \in [10^{-4}, 1]$ target the reference mean entropy, using neither class labels nor segmentation scores. The depth-0 execution settings are $\tau = 0.8384$ and $T = 0.088345$. Depths 0 and 12 attain the target on the calibration data. At depths 2–7, $T$ reaches the upper bound before attaining the target, with achieved entropies of 2.24–2.55 nats. Calibration is repeated for alternative sources and depths, and each resulting configuration is shared across all five benchmarks.

## C.2 Frozen head operation

For content tokens $X$, the frozen value projection is split into attention heads:

$$V_h = XW_{V,h}^\top + b_{V,h}, \qquad \widetilde{V}_h(i) = \sum_j S_{ij} V_h(j). \tag{3}$$

The same spatial support $S$ is used for every head. Concatenation and the output projection are followed by the residual MLP block:

$$Y = \mathrm{Concat}_h(\widetilde{V}_h) W_O^\top + b_O, \qquad Z = Y + \mathrm{MLP}(\mathrm{LN}(Y)). \tag{4}$$

The value/output weights and biases, LayerNorm, and MLP are all frozen. The learned probe and query/key projections are bypassed, and no additional post-head LayerNorm is applied. Identity support gives the $S = I$ baseline; SigMAP supplies the support of Eq. 1.

## C.3 Statistical matching

**Fitting the moments.** Statistics are fitted without labels on the development images specified in Appendix C.4, using all 576 patch tokens per image ($n = 576{,}000$). One set of moments is pooled across the development sample and reused on validation; validation images do not enter this fit. Encoder features are extracted in bf16, while statistics and transforms are computed in FP32. For row-vector features collected as $X \in \mathbb{R}^{n \times d}$, the mean and centered population covariance are

$$\mu = \frac{1}{n} \sum_{a=1}^{n} x_a, \qquad C = \frac{1}{n} X^\top X - \mu^\top \mu. \tag{5}$$

The covariance is regularized with trace-scaled shrinkage:

$$\widetilde{C} = 0.999C + 0.001\,\frac{\mathrm{tr}(C)}{d}I. \tag{6}$$

For the eigendecomposition $\widetilde{C} = Q\,\mathrm{diag}(\lambda)Q^\top$, define

$$R_C(p) = Q\,\mathrm{diag}\big(\max(\lambda, 10^{-6})^p\big)\,Q^\top. \tag{7}$$

The eigenvalue floor is applied elementwise; no rank truncation is used.

**Whitening–recoloring.** With source and target moments indexed by $s$ and $t$, respectively, the transform is

$$x' = (x - \mu_s)R_{C_s}(-\tfrac{1}{2})R_{C_t}(\tfrac{1}{2}) + \mu_t. \tag{8}$$

Under the row-vector convention, source whitening precedes target recoloring, and the target mean is restored. The transform uses the regularized covariances defined above, so matching is approximate for the original empirical covariances. No additional LayerNorm is applied after the transform.

**Norm and channel controls.** Token-norm matching rescales each source token to the target's mean token norm:

$$x' = x\,\frac{\mathbb{E}_t\|x_t\|_2}{\max(\|x\|_2, 10^{-12})}. \tag{9}$$

Per-channel affine matching uses development means and standard deviations:

$$x'_j = (x_j - \mu_{s,j})\frac{\sigma_{t,j}}{\sigma_{s,j}} + \mu_{t,j}. \tag{10}$$

**Application to content and support.** For the depth-10 content control, the source is the output after ten transformer blocks and the target is the post-LN depth-12 MAP-head input. Depth-0 support is unchanged, so its original threshold and temperature are retained. The transformed content is passed through the frozen value path without token-wise L2 normalization. For the support control, post-LN depth-12 features are transformed to depth-0 moments while the original final-layer content remains fixed. The transformed support tokens are L2 normalized to construct cosine affinities, and threshold/temperature calibration is repeated without labels. The resulting support-control configuration uses $\tau = 0.8586$ and $T = 0.098269$. Both axes use the recovery definitions in Appendix D.

### C.4 Data splits and evaluation protocol

The development screen contains the first 1,000 images of the VOC2012 segmentation training split and is disjoint from the segmentation validation sets. Diagnostic runs use seed 0 and bf16 encoder computation, with FP32 affinity construction and statistical transforms. Sample identities, including the 40-image training subsets for prompt selection, are pinned by SHA-256. Table 1 identifies the evaluation split for each intervention. Per-image intersection/union arrays, the measurement ledger, and written predictions are retained with the experiment artifacts.

**Prompt selection.** For each dataset, prompt templates are selected by an entropy criterion on 40 training images. Ground-truth masks are used to exclude void pixels; semantic class labels are not used in the selection criterion. The selected template lists accompany the code. Encoder and head weights remain frozen throughout calibration, prompt selection, and evaluation.

**External support and evaluation records.** External support uses `facebook/dinov2-base` (Oquab et al., 2024) without registers. Final support uses its last hidden state; patch-embedding support uses its embedding output interpolated to the $24 \times 24$ grid. Both are evaluated through the frozen decoder of §2.1 with source-specific calibration from Appendix C.1. The development map, support-source controls, and headline evaluation use the protocol in §3.1; their records include sample identifiers and per-image intersection/union arrays.

Table 7: Whitening depth-12 support to depth-0 second moments across five validation sets. The first three numeric columns are off-image assignment rates (%); assignment and score recovery are the fractions of the corresponding depth-12 to depth-0 gaps removed.

| Dataset | depth-12 | whitened | depth-0 | Assignment recovery | Whitened mIoU | Score recovery |
|---|---|---|---|---|---|---|
| VOC20 | 6.88 | 4.50 | 3.66 | 74% | 84.97 | 70% |
| Context59 | 19.01 | 15.43 | 14.48 | 79% | 34.61 | 39% |
| COCO-Stuff | 41.89 | 37.82 | 33.18 | 47% | 23.39 | 43% |
| ADE20K | 35.30 | 32.10 | 28.76 | 49% | 22.68 | 53% |
| Cityscapes | 3.02 | 2.63 | 2.36 | 59% | 41.37 | 75% |

## D Statistical Recovery Across Datasets

**Recovery definitions.** Let $q_b$, $q_t$, and $q_r$ denote the baseline, transformed, and reference mIoU on the same evaluation split. Score recovery is

$$R_{\text{score}} = 100\,\frac{q_t - q_b}{q_r - q_b}. \tag{11}$$

For support interventions, the baseline and reference use depth-12 and depth-0 support with content fixed. For content interventions, they use depth-10 and native depth-12 content with depth-0 support fixed. If $a_b$, $a_t$, and $a_r$ are the corresponding off-image assignment rates, assignment recovery is $100(a_b - a_t)/(a_b - a_r)$. Recovery measures the fraction of the baseline-to-reference metric gap removed by an intervention. For example, VOC20 support whitening gives $100(84.97 - 80.49)/(86.85 - 80.49) = 70.4\%$. Table 7 rounds recovery to whole percentages.