WINDOWFOLD: FOLDING SLIDING-WINDOW INFERENCE INTO A SINGLE FORWARD PASS FOR DENSE CLIP

Anonymous authors
Paper under double-blind review

ABSTRACT

Every training-free method that turns CLIP into a dense predictor is evaluated with sliding-window inference: the image is cut into overlapping 224×224 crops at the resolution CLIP was pretrained on, each crop is encoded separately, and the predictions are averaged, so the encoder processes the image area four times over. We show the window is a workaround for a train–test mismatch, and that it is doing two separable things. First, attention trained to spread its mass over 196 keys misallocates it over more: a plain high-resolution forward collapses while windows over the same pixels hold up. Second, a 224 crop silently caps how far a surgery may mix features, a cap worth 8.9 mIoU to a single pass on Cityscapes that no reweighting of attention can express. We replace each with a learned correction fit to the window's own output, with no labels, text or external data: a rank-16 LoRA on the attention projections, and a Gaussian footprint on the last block's mixing whose two to six widths are fit by the same distillation. Across three surgeries and four benchmarks, one forward matches or exceeds sliding-window inference in all 12 settings under the published protocol at 1.5–$2.4\times$ fewer FLOPs. On SC-CLIP, the training-free state of the art, one forward through the full method is above the published numbers on four of five benchmarks and on their average (42.68 against 42.4) with a single configuration. The fitted footprints are themselves a finding: five surgeries trained independently place the attention footprint at 5–7 patches, half a window, fixed in patch units rather than scaling with the image; across four backbones it follows the pretraining crop, not a pixel extent, and on L/14 one forward beats the ensemble on all five benchmarks.

1 INTRODUCTION

CLIP (Radford et al., 2021) was trained to align one global image vector with one sentence, yet its patch tokens carry localized semantics, and a productive line of work extracts that signal without training by editing the last self-attention block: MaskCLIP (Zhou et al., 2022) keeps only the value path, SCLIP (Wang et al., 2024) symmetrizes the attention, ClearCLIP (Lan et al., 2024a) removes the residual and feed-forward paths, and NACLIP (Hajimiri et al., 2025), ProxyCLIP (Lan et al., 2024b) and SC-CLIP (Bai et al., 2025) refine the same block further.

All of them share a detail that none presents as part of the method: none is evaluated on the image. Each uses sliding-window inference: the image is resized to 1.5–$3\times$ the pretraining resolution, 224×224 crops are taken at stride 112, encoded independently, and their logits averaged, so the encoder processes four times the image area, 6 to 55 crops in the settings we report. The window is the resolution CLIP was pretrained at, and it exists because CLIP cannot be handed a bigger patch grid: its position table must be interpolated and its attention was trained to spread a fixed mass over 196 keys. Applied to every baseline alike, it has become invisible.

We ask whether the redundant passes buy information or merely repair a mismatch. The distinction is testable, because the pixels are the same: a single forward at the same resolution sees everything the windows see, through an encoder operating outside its trained regime. If the gap is a mismatch, a small adaptation should close it without labels.

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.

This is what we find, and the finding has two parts because the window does two things. The first is the mismatch just described. A rank-16 LoRA on the attention projections, trained to regress the fused window features with the frozen model as its own teacher on unlabelled images, repairs it; a teacher equally free of the mismatch but blurry, the whole image at the native grid upsampled, recovers almost nothing, and adapting the position grid instead of attention recovers nothing at all. The second thing the window does had not been named: a 224 crop holds only 14×14 tokens, so it silently caps how far a surgery may mix features, and the newest surgeries mix over a full token-by-token affinity matrix. Reimposing the cap by hand is worth 8.9 mIoU to a single pass on Cityscapes while moving the windowed arm by under 0.1, and a LoRA cannot express it. We therefore supply it as a second learned component: a Gaussian footprint on the last block's mixing, one width per mixing site, two to six scalars in all, fit by the same distillation. The two corrections turn out to be close to additive.

Contributions. (i) A diagnosis: the sliding window of training-free dense CLIP is a workaround for a train–test mismatch, and it does two separable things, each isolated by a control: it gives attention the key count it was trained for and it bounds how far features mix (Section 3). (ii) A method, WindowFold, that folds both into one forward with no labels, text or external data: a LoRA and a learned footprint distilled from the window's own output (Section 4). One forward beats the ensemble in all 12 dataset–surgery combinations under the published protocol at 1.5–$2.4\times$ fewer FLOPs, and on SC-CLIP, the training-free state of the art, it is above the published numbers on four of five benchmarks and on their average (42.68 against 42.4) with one configuration. (iii) A finding: five surgeries trained independently put the attention footprint at 5–7 patches, half a window, fixed in patch units, and on SC-CLIP the fit all but switches off one of the surgery's own aggregation steps at high resolution (Section 5.5).

2 Related work

Attention surgery for dense CLIP. MaskCLIP (Zhou et al., 2022) reads the value path of the last block; SCLIP (Wang et al., 2024) uses query–query and key–key attention; ClearCLIP (Lan et al., 2024a) drops the residual and feed-forward network; NACLIP (Hajimiri et al., 2025) biases attention towards a token's neighbourhood; GEM (Bousselham et al., 2024) iterates a self-self attention; ProxyCLIP (Lan et al., 2024b) imports affinities from DINO (Caron et al., 2021); and SC-CLIP (Bai et al., 2025) removes the anomalous tokens that flatten CLIP's attention. All are training-free edits to a frozen encoder, all are evaluated with sliding windows, and our method is orthogonal to the choice among them. NACLIP is the closest relative of our footprint: it also bounds attention with a Gaussian in patch distance, but with a hand-chosen constant width for one site of one surgery, where ours is fit from the window's output, per mixing site, on any surgery, and recovers NACLIP's constant. Trident (Shi et al., 2025) also treats resolution as the problem, splicing sub-image features from CLIP and DINO and aggregating them through SAM; it keeps the sub-image passes and adds two foundation models at inference, where we remove the passes and deploy the original encoder plus a LoRA and six scalars. Methods that train for the task, such as GroupViT (Xu et al., 2022), LSeg (Li et al., 2022), TCL (Cha et al., 2023) and FreeDA (Barsellotti et al., 2024), change what the model knows; we change only how many passes it takes to read it out.

Resolution in ViTs, and self-distillation. FlexiViT (Beyer et al., 2023) and NaViT (Dehghani et al., 2023) train over multiple grids with labels and full fine-tuning; our target comes from the same frozen model under the configuration it was trained for, so no labels are needed. Distilling test-time ensembles into one pass (Hinton et al., 2015) is established; what is specific here is that the expensive procedure repairs a mismatch, which is why a very small adapter suffices and a mismatch-free blurry teacher fails.

3 Diagnosis

3.1 Setup

We evaluate on ADE20K (Zhou et al., 2017), PASCAL-Context-59 (Mottaghi et al., 2014), COCO-Stuff-171 (Caesar et al., 2018), Cityscapes (Cordts et al., 2016) and VOC20 (Everingham et al., 2010), and distil on unlabelled images from COCO (Lin et al., 2014) or from the ADE20K training

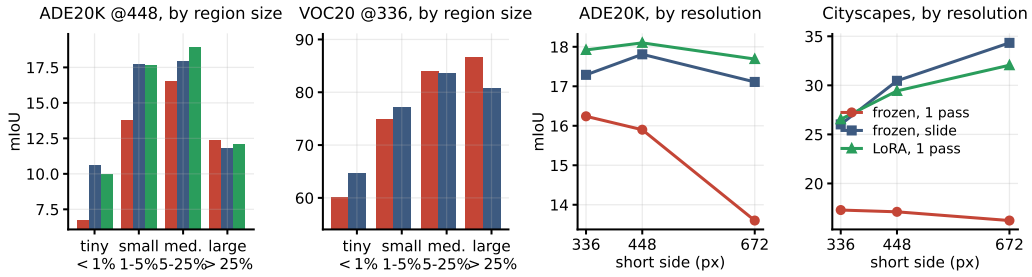


Figure 1: **What the window buys, and what more pixels do.** ClearCLIP, ViT-B/16. *Left two:* pixels split by ground-truth region area; windows win the small bins and lose the large ones, and the adapted single pass keeps both. *Right two:* frozen single-pass accuracy falls as the grid grows while windows over the same pixels hold; the adapted pass tracks or exceeds the ensemble at every resolution.

split (stated per adapter in Section 4.4; no annotation is read), under the standard protocol: short side resized to a fixed length, class embeddings averaged over the 80 OpenAI templates, predictions upsampled to label resolution, mIoU. Sliding windows use 224×224 crops at stride 112 with logit averaging, as in every method's released code. Appendix A checks our reproductions: SCLIP agrees with its published row to within 0.8 mIoU on all five datasets, NACLIP, ProxyCLIP and SC-CLIP match the reproduction in SC-CLIP (Bai et al., 2025) to within 0.6, and MaskCLIP is above the published row on four datasets. All surgeries go through identical code differing only in the last block.

3.2 One pass and many windows are not the same

Table 4 contains the frozen single-pass and frozen sliding-window columns. The gap is large and systematic: on ADE20K at short side 448, ClearCLIP loses 1.9 mIoU when the windows are replaced by one pass, SCLIP 3.9, and on Cityscapes ClearCLIP loses 13.4. Its structure is the first clue. Stratifying pixels by the area of their ground-truth region (Figure 1, left; Appendix D), the entire advantage of windows is on small regions: on ADE20K the tiny bin ($< 1\%$) improves from 6.76 to 10.58 and the small bin from 13.79 to 17.71, while the large bin ($> 25\%$) gets *worse*, 12.37 to 11.77. A 224 crop is a $2\times$ zoom on a small object and a blindfold on a large one; on VOC20, whose objects are large, a frozen single pass already beats the windows for all three surgeries.

3.3 The mismatch grows with the grid

If the single pass were merely a coarser view, its accuracy would be flat in resolution. Instead it collapses while windows over the same pixels hold up (Figure 1, right). From short side 336 to 672, frozen single-pass ClearCLIP falls from 16.24 to 13.60 on ADE20K and from 17.29 to 16.22 on Cityscapes, while windows go from 17.29 to 17.11 and from 26.02 to 34.33. That is the signature of a train–test mismatch that scales with the token count, not of a missing signal.

3.4 Training-free repairs are partial

The mismatch has a clear form, so we first try to correct it without training. On a 500-image ADE20K subset at short side 448 (one pass 12.60, windows 14.98), rescaling attention logits so entropy grows like $\log N$ gives 13.15; restoring each head's native entropy per query gives at best 13.32 and hurts when enforced strictly; partitioning the grid into 14×14 blocks with native position tables reproduces independent crops but destroys global context, and averaging that with a full pass reaches 14.65 at twice the cost. Attention scope and position code are co-adapted: changing either alone hurts, and changing both by hand only reproduces the crops, which motivates learning the correction.

	one pass, mixing radius					
	global	14	10	7	4	slide
ADE20K	15.09	16.39	16.82	17.02	**17.03**	18.83
Cityscapes	21.05	26.23	27.36	28.49	**29.90**	41.74
control: the same radius applied inside the windows						
ADE20K	18.83	18.83	18.85	18.80	18.56	—
Cityscapes	41.74	41.74	41.78	41.79	41.67	—

Table 1: **The window also bounds how far features mix.** Frozen SC-CLIP at short side 672, its affinity matrices zeroed outside a Chebyshev radius in patch units. The single pass gains up to $+8.9$ mIoU, monotonically in the bound. The same restriction inside the windows, where a bound of 13 was already in force, changes nothing at radius 14 (no restriction in a 14×14 window), at most 0.05 at 7–10, and costs 0.07–0.27 at 4. No training in any cell.

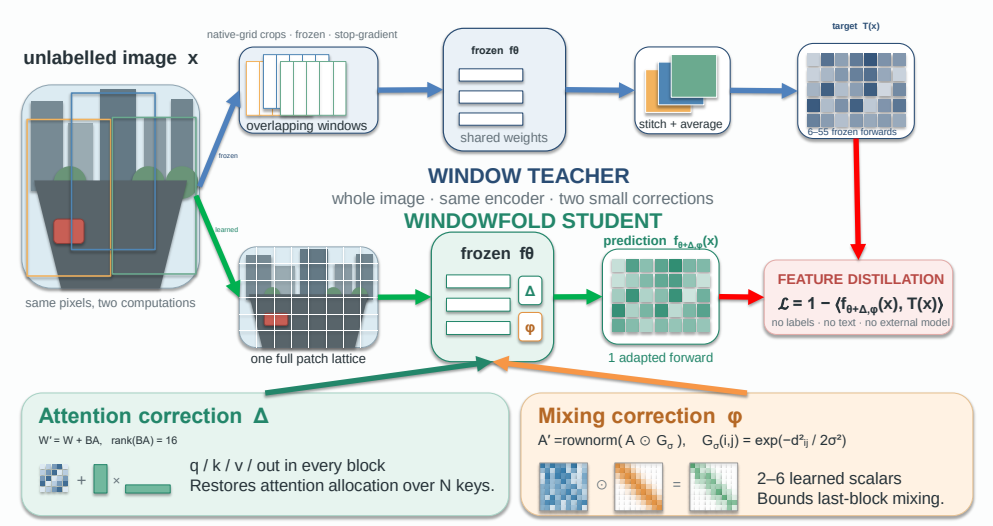


Figure 2: **WindowFold overview.** The frozen teacher encodes native-grid windows of an unlabelled image and stitches them into $T(x)$; the student sees the whole patch lattice once. Feature distillation fits two corrections only: a rank-16 LoRA Δ on every block's attention projections and two to six Gaussian footprint scalars ϕ on the last block's mixing. No labels, text or external models; after training the teacher is discarded.

3.5 The window also bounds how far features mix

Table 1 shows the window does a second, different thing. SC-CLIP mixes features over a full token-by-token affinity matrix, twice. Inside a 224 window that mixing is local by construction: there are only 14×14 tokens. In one pass at short side 672 the same operation pools over 42×56 tokens on ADE20K and 42×84 on Cityscapes. We bounded it by hand, zeroing the affinity outside a fixed Chebyshev radius in patch units, and changed nothing else. The frozen single pass improves monotonically as the bound tightens, by $+1.94$ mIoU on ADE20K and $+8.85$ on Cityscapes; the same bounds inside the windows move them by at most 0.05 down to radius 7, because there the crop already supplies them, and radius 4 costs the windows 0.07–0.27. The window was never only a way to keep the token count at 196; it was also an implicit cap on the spatial extent of aggregation, and for a surgery that aggregates globally that cap is worth more than the token count. The optimum for the single pass, radius 4, is tighter than the window's own 7, while inside the window the same radius is already too tight: the gain is from restoring a missing bound, not from locality as such. A correction to attention cannot express this: a low-rank update reweights keys and has no vocabulary for "not beyond this distance". The method therefore has two parts, one for each thing the window was doing.

Adapted component	ADE20K mIoU
none (frozen, 1 pass)	15.90
learnable position grid only	15.88
LoRA on attention only	**18.23**
both	**18.23**
frozen, slide (4×)	17.81

Table 2: **The mismatch is absorbed by attention.** A learnable position grid trained for 12k steps moves nothing; LoRA alone reproduces the whole effect. ClearCLIP, ADE20K@448.

Dataset	frozen 1 pass	resize teacher	window teacher	frozen slide
ADE20K	15.90	16.15	**18.15**	17.81
Context-59	34.85	35.02	**37.68**	35.91
COCO-Stuff	23.23	24.46	**25.93**	24.67

Table 3: **The gain comes from the windows, not from fine-tuning.** The resize teacher is the whole image at the native grid, upsampled: same images, loss and steps, no mismatch, no crop detail. It recovers almost none of the gap. ClearCLIP@448.

4 METHOD

4.1 OBJECTIVE

Figure 2 gives the complete training and deployment path. Let f_θ be the frozen encoder with a surgery applied to its last block and x an image with patch lattice $g_h \times g_w$. The teacher is the sliding-window computation on that lattice, $T(x) = \mathrm{norm}\big(\sum_w P_w^\top f_\theta(x_w) \,/\, \sum_w P_w^\top \mathbf{1}\big)$, over 224 crops x_w at stride 112, with P_w placing each 14×14 output onto the lattice; crop and stride are multiples of the patch size, so nothing is resampled. The student is the same encoder and surgery run once on the whole image with two learned corrections, a low-rank update Δ on the attention projections and a footprint ϕ on the last block's mixing, trained to minimise $\mathcal{L} = \mathbb{E}_x\big[1 - \langle f_{\theta+\Delta,\phi}(x), T(x)\rangle\big]$ per lattice cell after ℓ_2 normalisation. No labels, text or external data are involved.

4.2 ATTENTION: A LOW-RANK CORRECTION

We place a LoRA of rank 16 on the query, key, value and output projections of every block: 48 sites, 1.78M trainable parameters against 86M frozen. The obvious alternative, a learnable position grid initialised from the interpolated table, is inert (Table 2: 15.88 against a frozen 15.90 after 12k steps), while LoRA alone gives 18.23 and adding the grid changes nothing. The interpolated position code is a serviceable description of where each patch is; what fails when the grid grows is how attention allocates probability across a larger key set, and a low-rank correction to the projections restores it.

4.3 MIXING EXTENT: A LEARNED FOOTPRINT

Section 3.5 shows the second thing the window does is worth up to 8.9 mIoU and cannot be expressed by a LoRA, so we supply it directly and let the teacher set it. Every surgery's last block mixes patch features through one or more affinity matrices $A \in \mathbb{R}^{N\times N}$: the self-self attention of ClearCLIP, SCLIP and NACLIP, the proxy affinities of ProxyCLIP, and for SC-CLIP the attention plus two similarity-map aggregations. We multiply each by a Gaussian in patch distance,

$$G_\sigma(i,j) = \exp\left(-\frac{d_{ij}^2}{2\sigma^2}\right), \qquad \log\sigma = a + \tfrac{b}{2}\log\tfrac{N}{N_{\text{native}}}, \tag{1}$$

applied as a logit bias $\log G_\sigma$ where the affinity passes through a softmax and as a multiplicative weight where it is row-normalised. Each mixing site has its own (a, b): one pair for most surgeries, three for SC-CLIP, so the footprint is two to six scalars. The parameterisation leaves both hypotheses the diagnosis raises representable and assumes neither: $b = 0$ is a footprint fixed in patch units, which is what the window imposes; $b = 1$ is one fixed as a fraction of the image side. We initialise at $\sigma_0 = 10$ patches and $b = 0$, wide enough to be close to no bound on the grids we train on, and let the loss decide. The footprint is trained jointly with the LoRA from the same teacher with no extra supervision.

4.4 TRAINING OVER SEVERAL TOKEN COUNTS, AND IMPLEMENTATION

A student trained at one crop shape is sharp at one token count: an adapter trained on 448^2 crops agrees with its teacher at cosine 0.9768 on a 28×28 lattice and 0.9466 on the 28×56 lattice

		short side 336 (published protocol)				short side 448			
Dataset	Surgery	1 pass	slide	WindowFold	Δ	1 pass	slide	WindowFold	Δ
ADE20K	MaskCLIP	13.27	13.65	**14.45**	+0.80	12.86	13.35	**14.54**	+1.20
	SCLIP	14.09	16.48	**17.31**	+0.83	13.03	16.93	**17.35**	+0.42
	ClearCLIP	16.24	17.29	**17.92**	+0.63	15.90	17.81	**18.10**	+0.28
Context-59	MaskCLIP	27.82	28.48	**30.40**	+1.92	26.85	27.57	**30.35**	+2.78
	SCLIP	31.77	34.02	**35.40**	+1.38	31.13	34.17	**35.68**	+1.51
	ClearCLIP	35.11	35.77	**37.42**	+1.65	34.85	35.91	**37.89**	+1.98
COCO-Stuff	MaskCLIP	18.67	19.30	**20.58**	+1.28	17.72	18.62	**20.54**	+1.92
	SCLIP	20.84	23.15	**24.08**	+0.93	19.91	23.30	**24.46**	+1.16
	ClearCLIP	23.64	24.59	**25.78**	+1.19	23.23	24.67	**26.09**	+1.41
Cityscapes	MaskCLIP	23.75	26.17	**27.89**	+1.72	24.73	27.52	**30.35**	+2.83
	SCLIP	15.97	26.50	**26.77**	+0.27	15.59	**30.72**	30.17	−0.55
	ClearCLIP	17.29	26.02	**26.54**	+0.52	17.11	**30.46**	29.43	−1.03

Table 4: **One adapted forward against four frozen ones.** Full validation sets; windows are 224 crops at stride 112 with logit averaging; short side 336 is the protocol SCLIP and ClearCLIP publish under. WindowFold here is the rank-16 LoRA distilled on 5000 unlabelled COCO images, one adapter per surgery reused across all cells. It wins all 12 cells at 336 and 10 of 12 at 448; the exceptions are Cityscapes, where distilling on unlabelled Cityscapes images instead brings ClearCLIP to 30.25 against 30.46.

of a 2:1 street scene. We therefore cycle over crop shapes 336^2, 448^2, 448×896 and 672^2 with per-shape batch sizes 8, 6, 2, 2, which lifts agreement on the wide lattice to 0.9637. Training uses unlabelled pixels only, with no annotation file opened. The LoRA-only adapters of MaskCLIP, SCLIP, ClearCLIP, NACLIP and ProxyCLIP are distilled on the 5000 images of the COCO 2017 validation split; every SC-CLIP adapter and every adapter that includes a footprint is distilled on the 20,210 images of the ADE20K training split, disjoint from the ADE20K validation images we evaluate on; for those rows the ADE20K column is therefore in-domain unlabelled distillation. All runs take 12k steps with AdamW, peak learning rate 10^{-4} (10^{-3} for the footprint's scalars), one-cycle schedule and gradient clipping at 1.0; 40–60 minutes per surgery on one RTX 4090, teacher recomputed online.

5 Experiments

5.1 Main result

Table 4 is the central table: four datasets, three surgeries, two resolutions. Under the published protocol (short side 336, where SCLIP and ClearCLIP report), one adapted forward beats four frozen ones in all twelve combinations, by +0.27 to +1.92 mIoU; at short side 448 it wins ten of twelve, by up to +2.83. One adapter per surgery is trained once on unlabelled COCO and reused across all cells with nothing tuned per dataset. The gain is largest where the frozen window is weakest (MaskCLIP, +1.20 to +2.83), consistent with a mismatch whose size does not depend on the surgery. The two negative cells are Cityscapes at 448, where the ensemble gains 4.4 mIoU from the extra resolution and the COCO-trained adapter 2.9; distilling on unlabelled Cityscapes images instead brings ClearCLIP@448 to 30.25 against a windowed 30.46. Appendix H shows the three arms on images.

5.2 Where the teacher's advantage comes from

Perhaps any self-distillation of a frozen CLIP would improve dense features, making the window teacher incidental. Table 3 tests this with a teacher that shares everything with the window teacher except the detail: the whole image at the native 224 grid, encoded once and bilinearly upsampled, same images, loss, optimiser and steps. It yields +0.25, +0.17 and +1.23 mIoU on ADE20K, Context-59 and COCO-Stuff, against +2.25, +2.83 and +2.70 for the window teacher.

	square only	mixed aspect	mixed asp.+res	frozen slide
ADE20K @448	**18.23**	18.15	18.10	17.81
Context-59 @448	37.53	37.68	**37.89**	35.91
COCO-Stuff @448	25.75	25.93	**26.09**	24.67
Cityscapes @448	27.20	28.80	29.43	**30.46**
ADE20K @672	**18.42**	17.81	17.69	17.11
Context-59 @672	**37.23**	36.34	36.22	33.39
COCO-Stuff @672	**25.18**	24.75	24.51	22.99
Cityscapes @672	28.08	30.59	32.05	**34.33**

Table 5: **Shape diversity is a trade, not an enabler.** ClearCLIP, COCO-trained. Every recipe is stable at 672 where the frozen pass collapses; mixed shapes buy 2:1 data and cost 0.6–1.0 on square-ish data at 672, where the square recipe extrapolates best. The ensemble degrades from 448 to 672 on the three square-ish datasets.

	ADE20K		Cityscapes	
	@448	@672	@448	@672
NACLIP				
frozen, 1 pass	16.55	14.27	25.19	23.24
frozen, slide	17.52	16.59	33.64	36.51
LoRA, square	**18.22**	**18.58**	31.95	33.69
LoRA, mixed	18.13	17.65	**34.14**	**37.04**
ProxyCLIP				
frozen, 1 pass	18.35	16.27	24.65	25.80
frozen, slide	18.68	17.55	**34.49**	**37.13**
LoRA, square	**19.41**	**19.75**	28.46	30.56
LoRA, mixed	18.89	18.52	29.27	31.87
SC-CLIP				
frozen, 1 pass	18.34	15.09	23.29	21.05
frozen, slide	20.07	18.83	**38.95**	**41.74**
LoRA, square	20.30	**20.14**	32.02	31.49
LoRA, mixed	20.11	18.79	33.70	33.82
+ footprint	**20.33**	19.36	36.46	39.14

Table 6: **The three strongest surgeries.** All carry the mismatch (frozen single pass loses 2.3, 2.1, 3.3 mIoU on ADE20K from 448 to 672) and all are repaired on ADE20K with a margin that widens with resolution. Last row: the full method, which closes most of SC-CLIP's Cityscapes deficit. SC-CLIP adapters are distilled on unlabelled ADE20K training images (Section 4.4).

5.3 Recipes, and resolution as a knob

Table 5 compares training recipes. At 448 shape diversity helps where it covers the aspect ratio (27.20 → 29.43 on Cityscapes) and is neutral elsewhere; at 672 the square recipe is best on the three square-ish datasets by 0.7–1.0 despite never seeing a 672 crop. We adopt the mixed recipe as the one never far from best. More important, the window ensemble itself degrades from 448 to 672 on all three square-ish datasets (17.81 → 17.11, 35.91 → 33.39, 24.67 → 22.99), because a finer lattice leaves less context per pinned crop. The adapted pass has no such constraint: its margin widens with resolution (+1.62 → +3.84 on Context-59, square recipe), and one adapted forward at 672 beats the ensemble at 448 even on Cityscapes (32.05 against 30.46). Resolution is a knob a windowed pipeline cannot offer, because for it resolution only adds windows.

5.4 The strongest surgeries, including the state of the art

Table 6 adds NACLIP, the strongest purely internal surgery; ProxyCLIP, whose last-block attention is read off a frozen DINOv2, so that the LoRA reaches only the blocks below it; and SC-CLIP, the training-free state of the art among methods that add no second foundation model at inference (Trident, which adds DINO and SAM, is higher; Section 5.6), whose port reproduces the published row to within 0.2 mIoU on four of five benchmarks (Table 9). All three carry the mismatch and SC-CLIP carries the most: the frozen single pass falls by 2.3, 2.1 and 3.3 mIoU on ADE20K from 448 to 672, so the surgery that is best under the ensemble is worst without it; the more machinery a surgery builds on the frozen encoder, the more of its accuracy is contingent on the trained token count. All three are repaired on ADE20K with a margin that widens with resolution; the best single-pass number anywhere is SC-CLIP on square crops at 672, 20.14 against 18.83 for its own ensemble. That ProxyCLIP is helped at all says the adapter fixes something underneath the external attention; a ClearCLIP-trained adapter moved onto other surgeries recovers 48–61% of a matched adapter's gain (Appendix G). Cityscapes remains the exception: the LoRA-only pass loses to the ensemble by 6 mIoU on ProxyCLIP and 8 on SC-CLIP, the deficit Section 3.5 diagnosed.

Surgery	site	learned σ_{28}	learned b	σ at grid 21×28	σ at grid 42×56
ClearCLIP	attn	7.0	−0.11	7.1	6.6
SCLIP	attn	6.7	−0.22	6.9	5.9
NACLIP	attn	5.7	−0.06	5.8	5.5
ProxyCLIP	attn	5.2	−0.51	5.6	3.9
SC-CLIP	attn	7.4	−0.21	7.6	6.6
	pre	1.4	−1.50	1.7	0.6
	post	4.1	−0.45	4.4	3.2

Table 7: **What the footprint learns.** Init $\sigma = 10$, $b = 0$; fit jointly with the LoRA. σ_{28}: width on the 28×28 training grid; b: exponent on token count. Five surgeries put the attention footprint at 5–7 patches with every b negative; SC-CLIP's first aggregation is driven to a 3×3 neighbourhood, its second to the radius of Table 1.

		ADE20K @448	ADE20K @672	Cityscapes @448	Cityscapes @560	Cityscapes @672
SC-CLIP	frozen, 1 pass	18.34	15.09	23.29	22.53	21.05
	frozen, slide	20.07	18.83	**38.95**	**40.96**	**41.74**
	LoRA	20.11	18.79	33.70	34.23	33.82
	footprint	19.39	16.94	29.45	29.94	28.91
	LoRA + foot.	**20.33**	**19.36**	36.46	38.55	39.14
ClearCLIP	frozen, slide	17.81	17.11	**30.46**	32.68	**34.33**
	LoRA	18.10	17.69	29.43	30.95	32.05
	footprint	16.83	—	21.12	21.77	—
	LoRA + foot.	**18.27**	**18.08**	29.23	**32.14**	33.82

Table 8: **The two corrections are separable and nearly additive.** Mixed recipe; training corpus per adapter as in Section 4.4. On SC-CLIP at 672 the footprint alone is worth $+7.9$ mIoU on Cityscapes, the LoRA alone $+12.8$, both $+18.1$. On ClearCLIP, whose one site is left at $\sigma \approx 7$, the footprint is close to inert.

5.5 What the footprint learns, and what it is worth

Table 7 reports what the footprint settles on. Three things stand out. The first is agreement: five surgeries, trained independently with the width free to go anywhere from one patch to the whole grid, put the attention footprint between 5 and 7 patches. Seven is half a window, and the four plain attentions land within 1.5 of it. NACLIP's authors chose $\sigma = 5$ by hand for a 14×14 grid; the fit, told nothing about that, picks 5.7. The width is a property of the encoder, of how far a CLIP token's useful context extends in patches, rather than of any one surgery. The second is the sign of b: every exponent is negative, and the four plain attentions have $|b| \leq 0.22$, so the footprint is fixed in patch units and does not grow with the image. This is the window's hypothesis, not the alternative, which we had also tested directly by scaling NACLIP's prior with the image and found to cost accuracy (Appendix F).

The third is SC-CLIP, whose three sites the fit treats very differently. Its attention goes to 7.4, with the others. Its second aggregation goes to 4.1, the radius the hand search of Table 1 found best. Its first aggregation, which mixes the outlier-repaired features through a mid-layer similarity map, is driven to $\sigma = 1.4$ on the training grid and 0.6 at 672: a 3×3 neighbourhood, then nearly the identity. This operation, introduced to smooth features across similar tokens, is helpful inside a window and harmful across a street scene, and the fit's remedy is to all but remove it at high resolution, a decision no hand search over three coupled radii would have reached.

Table 8 says what it is worth, and that the two corrections repair two different things. On SC-CLIP at 672 the footprint alone, six scalars and no LoRA, lifts Cityscapes by $+7.9$ mIoU, the same as the hand radius did; the LoRA alone by $+12.8$; the two together by $+18.1$, within 2.6 of their sum, which two routes to one repair could not produce. The deficit the rest of the paper could not close falls from 7.9 to 2.6 (39.14 against the ensemble's 41.74), and on ADE20K the pair is best at both resolutions. On ClearCLIP, whose one site is left at $\sigma \approx 7$, the footprint moves things by ±0.2 at 448 and $+0.4$ and $+1.8$ at 672: where mixing was already local there was less to bound. Trained without the LoRA the attention site drifts wide ($\sigma = 16$ on ClearCLIP, 14 on SC-CLIP): the 5–7 patch figure is the footprint of a repaired attention, not of a broken one. The fit is robust to initialisation ($\sigma_0 = 4, 10, 20$), seed and training corpus: every width reproduces within half a patch and the five-benchmark average moves by at most 0.06 (Appendix C).

5.6 Where this sits in the literature

Table 9 places the method next to the field under the published protocol. The LoRA alone beats the window ensemble on 26 of 30 cells, every exception being Cityscapes; the footprint raises every per-surgery average, by $+4.3$ and $+4.4$ on ProxyCLIP and SC-CLIP, which mix globally. On SC-CLIP (Bai et al., 2025) one forward through the full method is above the published state of the art on VOC20 (85.77 against 84.3), Context-59 (41.09 against 40.1) and COCO-Stuff (27.86 against

		V20	PC59	Stuff	ADE	City†	avg
MaskCLIP	slide	64.11	28.48	19.30	13.65	28.41	30.79
	LoRA	**66.42**	**30.40**	**20.58**	**14.45**	**31.76**	**32.72**
SCLIP	slide	81.52	34.02	23.15	16.48	**32.92**	37.62
	LoRA	82.80	35.40	24.08	17.31	31.53	38.22
	WindowFold	**83.82**	**35.49**	**24.14**	**17.37**	32.65	**38.69**
ClearCLIP	slide	82.44	35.77	24.59	17.29	**32.68**	38.55
	LoRA	83.58	**37.42**	**25.78**	17.92	30.95	39.13
	WindowFold	**84.31**	37.28	25.59	**18.01**	32.14	**39.47**
NACLIP	slide	77.21	35.05	23.87	17.38	35.57	37.82
	LoRA	79.33	36.72	**24.96**	17.91	35.94	38.97
	WindowFold	**79.96**	**36.77**	24.83	**18.02**	**36.04**	**39.12**
ProxyCLIP	slide	83.05	37.13	25.71	18.68	**36.27**	40.17
	LoRA	84.62	**38.13**	**26.64**	**18.90**	30.86	39.83
	WindowFold	**84.85**	38.12	26.41	18.80	35.29	**40.69**
SC-CLIP (Bai et al., 2025), as reported		84.3	40.1	26.6	20.1	41.0	42.4
SC-CLIP	slide	84.51	40.19	27.40	20.05	**40.96**	42.62
	LoRA	**86.39**	**41.21**	**27.91**	20.00	34.23	41.95
	WindowFold	85.77	41.09	27.86	**20.14**	38.55	**42.68**

Table 9: **Where this sits.** Published protocol, ViT-B/16; †Cityscapes at short side 560. "LoRA" is the attention correction alone; WindowFold adds the footprint (MaskCLIP has no mixing site). SC-CLIP's first row is our port of the published row above it. Training corpus per row as in Section 4.4: COCO for the non-SC-CLIP LoRA rows, unlabelled ADE20K training images otherwise.

26.6), level on ADE20K, and 2.4 short on Cityscapes; its average is 42.68 against 42.4 published. Trident (Shi et al., 2025), which adds DINO and SAM at inference, is higher still.

Other backbones. On ViT-L/14 the mismatch is larger and so is the repair (Appendix E): the full method beats the ensemble on all five benchmarks for all three surgeries we ran, by +1.46 on ClearCLIP, +3.10 on NACLIP and +3.67 on SCLIP. On ViT-B/32 the footprint closes the gap the LoRA alone left (+0.71 on average). The fitted footprints are 6.5–8.4 patches on L/14's 16-patch grid, 5–7 on B/16's 14 and 3.6 on B/32's 7: half the pretraining grid on each, which for these 224-crop models is also a fixed 112 pixels. ViT-L/14@336 (24-patch grid) separates the two: ClearCLIP's footprint moves from 8.4 to 15.1 patches and NACLIP's from 6.5 to 11.1, so the width follows the pretraining crop.

Cost. In FLOPs the single pass is cheaper at every resolution in the paper, by 1.5–2.4×; in wall clock it is 1.2–1.8× faster on B/16 and 1.2–2.1× on L/14 (Appendix B).

6 Limitations

Adapters leave a residual deficit on Cityscapes, whose scale distribution is unusual: SC-CLIP trails its own 55-window ensemble by 2.6 mIoU at 672, and closing this to 0.9 needs unlabelled in-domain images (Appendix C). The footprint-bearing adapters were distilled on unlabelled ADE20K training images, so their ADE20K numbers are in-domain in that sense. The footprint is one fixed shape and needs a mixing site to bound, which MaskCLIP lacks; SC-CLIP is run only at ViT-B/16; and quadratic attention overtakes the ensemble's redundancy near short side 1344.

7 Conclusion

The sliding window of training-free dense CLIP repairs a train–test mismatch rather than gathering information, and it does two separable things: restoring attention's trained key count and bounding how far a surgery may mix features. A rank-16 LoRA and two to six footprint scalars, distilled from the window's own output without labels, fold both into one forward that beats the ensemble in all 12 dataset–surgery combinations at 1.5–2.4× fewer FLOPs; the fitted footprint, half the pretraining grid whatever the image size, says what was inside the window.

AI USE STATEMENT

Generative AI tools (large language model assistants operated through a coding environment) were used throughout this work: for research ideation and execution, including the design of the radius control and the learned footprint, the implementation of the training and evaluation code and the port of SC-CLIP; for drafting and polishing the text of every section; and for retrieval of published numbers and related work. All AI-generated code was run end to end by the authors and its outputs checked against published reproductions where they exist (Appendix A); every number in the paper is the output of that code on the stated data, and no result was produced or altered by an AI tool. AI-drafted text was reviewed, edited and verified against the experimental logs by the authors, and related work surfaced by AI tools was read in the original before being cited. We take responsibility for the final content of this work, including text, claims and artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

All numbers in the main text and appendix are produced by one evaluation script with the same text prompts (the 80 OpenAI templates), the same $224/112$ sliding window ($336/168$ for ViT-L/14@336) and the same logit averaging for every arm; the protocol is stated in Section 3.1 and reproductions against published rows are in Appendix A. Adapters are trained once per surgery on unlabelled images only (COCO 2017 validation split or ADE20K training split, as stated per adapter in Section 4.4) with the hyper-parameters of that section and reused unchanged across datasets and resolutions; the footprint is trained jointly with the LoRA from the same teacher with the initialisation stated in Section 4.3. The supplementary material contains the training and evaluation code, the ViT-B/16 and ViT-B/32 adapter and footprint checkpoints behind the main tables, the per-cell evaluation logs and the scripts that produced them; the remaining checkpoints will be released with the code.

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.

Luca Barsellotti, Roberto Amoroso, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. Training-free open-vocabulary segmentation with offline diffusion-augmented prototype generation. In *CVPR*, 2024.

Lucas Beyer, Pavel Izmailov, Alexander Kolesnikov, Mathilde Caron, Simon Kornblith, Xiaohua Zhai, Matthias Minderer, Michael Tschannen, Ibrahim Alabdulmohsin, and Filip Pavetić. FlexiViT: One model for all patch sizes. In *CVPR*, 2023.

Walid Bousselham, Felix Petersen, Vittorio Ferrari, and Hilde Kuehne. Grounding everything: Emerging localization properties in vision-language transformers. In *CVPR*, 2024.

Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. COCO-Stuff: Thing and stuff classes in context. In *CVPR*, 2018.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, 2021.

Junbum Cha, Jonghwan Mun, and Byungseok Roh. Learning to generate text-grounded mask for open-world semantic segmentation from only image-text pairs. In *CVPR*, 2023.

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 *CVPR*, 2016.

Mostafa Dehghani, Basil Mustafa, Josip Djolonga, Jonathan Heek, Matthias Minderer, Mathilde Caron, Andreas Steiner, Joan Puigcerver, Robert Geirhos, Ibrahim Alabdulmohsin, Avital Oliver, Piotr Padlewski, Alexey Gritsenko, Mario Lučić, and Neil Houlsby. Patch n' pack: NaViT, a vision transformer for any aspect ratio and resolution. In *NeurIPS*, 2023.

Mark Everingham, Luc Van Gool, Christopher K. I. Williams, John Winn, and Andrew Zisserman. The PASCAL visual object classes (VOC) challenge. *IJCV*, 88(2):303–338, 2010.

Sina Hajimiri, Ismail Ben Ayed, and José Dolz. Pay attention to your neighbours: Training-free open-vocabulary semantic segmentation. In *WACV*, 2025.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv:1503.02531*, 2015.

Mengcheng Lan, Chaofeng Chen, Yiping Ke, Xinjiang Wang, Litong Feng, and Wayne Zhang. ClearCLIP: Decomposing CLIP representations for dense vision-language inference. In *ECCV*, 2024a.

Mengcheng Lan, Chaofeng Chen, Yiping Ke, Xinjiang Wang, Litong Feng, and Wayne Zhang. ProxyCLIP: Proxy attention improves CLIP for open-vocabulary segmentation. In *ECCV*, 2024b.

Boyi Li, Kilian Q. Weinberger, Serge Belongie, Vladlen Koltun, and René Ranftl. Language-driven semantic segmentation. In *ICLR*, 2022.

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 *ECCV*, 2014.

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 *CVPR*, 2014.

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

Yuheng Shi, Minjing Dong, and Chang Xu. Harnessing vision foundation models for high-performance, training-free open vocabulary segmentation. In *ICCV*, 2025.

Feng Wang, Jieru Mei, and Alan Yuille. SCLIP: Rethinking self-attention for dense vision-language inference. In *ECCV*, 2024.

Jiarui Xu, Shalini De Mello, Sifei Liu, Wonmin Byeon, Thomas Breuel, Jan Kautz, and Xiaolong Wang. GroupViT: Semantic segmentation emerges from text supervision. In *CVPR*, 2022.

Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ADE20K dataset. In *CVPR*, 2017.

Chong Zhou, Chen Change Loy, and Bo Dai. Extract free dense labels from CLIP. In *ECCV*, 2022.

APPENDIX

The appendix collects experiments the main text refers to but does not have room for: the reproduction check, the cost of the ensemble, robustness of the fitted footprint, the scale-stratified view of the gain, the backbone sweep, a resolution-conditioned LoRA that did not help, cross-surgery transfer of a single adapter, and qualitative examples.

A REPRODUCTION CHECK, AND THE MASKCLIP VOC20 CELL

	MaskCLIP		SCLIP		ClearCLIP	
Dataset	pub.	ours	pub.	ours	pub.	ours
ADE20K	11.9	**13.65**	16.1	16.48	—	17.29
Context-59	26.4	**28.48**	34.2	34.02	—	35.77
COCO-Stuff	16.7	**19.30**	22.4	23.15	—	24.59
Cityscapes[†]	24.9	**28.41**	32.2	32.92	—	32.68
VOC20	74.9	64.11	80.4	81.52	80.9	82.44

Table 10: **Reproduction check** under the published protocol: short side 336, 224 crops at stride 112, ViT-B/16, no post-processing. Published MaskCLIP and SCLIP numbers are from the SCLIP paper. [†]Cityscapes uses short side 560, as SCLIP specifies for that dataset. Our SCLIP agrees within 0.8 on all five datasets. Our MaskCLIP is above the published row on four and below on VOC20; Appendix A examines that cell. For the three surgeries we add later, the independent reproduction in SC-CLIP (Bai et al., 2025) is the fairer reference than the authors' own configurations, and we match it closely: on ADE20K it reports 17.4 for NACLIP and 18.7 for ProxyCLIP against our 17.38 and 18.68, and on Context-59 and COCO-Stuff we agree within 0.6. Our port of SC-CLIP itself matches its own paper within 0.2 on ADE20K, Context-59, VOC20 and Cityscapes, and is 0.8 above it on COCO-Stuff (Table 9). On ViT-L/14 our ClearCLIP ensemble is within 0.6 of the published row on all five benchmarks (Table 15).

Table 10 checks our reproductions against published numbers under the published protocol: short side 336, ViT-B/16, no post-processing, and short side 560 on Cityscapes, which SCLIP singles out for its unusually large images. SCLIP agrees to within 0.8 mIoU on all five datasets. Our MaskCLIP is *above* the published row on ADE20K, Context-59, COCO-Stuff and Cityscapes, and below it only on VOC20 (64.1 against 74.9).

We looked into that one cell, because a ten-point deficit on a single dataset would ordinarily indicate a bug. It does not appear to be one. No VOC class collapses; MaskCLIP is uniformly 15–20 points below ClearCLIP on all twenty. We tried six formulations of the last block (the value path alone, with and without the residual, with and without the feed-forward network, and with the output projection removed). The four viable ones span 58.6 to 65.1 and none approaches 74.9; removing the output projection collapses the model entirely (1.9), which at least confirms that path is the one carrying the signal. Evaluating at MaskCLIP's own 520px whole-image setting is worse, not better (60.2). MaskCLIP's prompt-denoising refinement adds 3.8, but it adds a comparable amount to ClearCLIP as well, so it is a protocol change rather than a MaskCLIP-specific gain.

The likelier explanation is in the provenance of the published row. SCLIP states that for MaskCLIP it reports "the higher numbers between our re-implementation based on their official code bases and results from the existing work", so the row is a maximum over sources and protocols rather than a single measurement. Consistently, a third-party comparison (Barsellotti et al., 2024) that tabulates parameter counts attributes 291.0M parameters to MaskCLIP, whereas CLIP ViT-B/16, the backbone SCLIP and ClearCLIP use, and the one we use, has 149.6M. Whatever produced 74.9 on VOC20 was not a ViT-B/16 run of the protocol we and the other two surgeries are evaluated under.

B COST OF THE WINDOW ENSEMBLE AGAINST ONE PASS

Table 12 gives the window counts, FLOP ratios and measured wall-clock ratios behind the cost figures quoted in the main text. The single pass's cost grows faster than the ensemble's because its

	all	tiny	small	medium	large
ADE20K, short side 448					
frozen, 1 pass	15.90	6.76	13.79	16.51	12.37
frozen, slide	17.81	**10.58**	**17.71**	17.94	11.77
WindowFold	**18.15**	10.00	17.63	**18.91**	**12.05**
VOC20, short side 336 (frozen only)					
frozen, 1 pass	**84.63**	60.03	74.94	**84.03**	**86.70**
frozen, slide	82.44	**64.63**	**77.21**	83.59	80.80
Cityscapes, short side 448					
frozen, 1 pass	17.11	6.97	16.27	17.55	8.73
frozen, slide	**30.46**	**19.02**	**28.78**	27.63	12.41
WindowFold	29.43	16.49	26.28	**28.37**	**12.99**
WindowFold @672	32.05	17.94	28.73	31.17	13.25

Table 11: **Pixels stratified by the area of their ground-truth region** (tiny $< 1\%$, small 1–5%, medium 5–25%, large $> 25\%$ of the image). Sliding windows win on small regions and *lose* on large ones: a 224 crop is a zoom on the former and a blindfold on the latter. WindowFold takes the small-region gain without paying the large-region cost, which is why it ends up ahead overall. On VOC20, where large objects dominate, the same mechanism makes the frozen single pass beat sliding windows outright. On Cityscapes the remaining deficit is entirely in the two smallest bins, and raising the resolution addresses it.

	4:3					2:1				
short		ViT-B/16		ViT-L/14			ViT-B/16		ViT-L/14	
side	wins	FLOP×	time×	FLOP×	time×	wins	FLOP×	time×	FLOP×	time×
336	6	1.93	1.28	1.92	1.74	10	1.95	1.45	1.96	1.69
448	15	2.43	1.82	2.45	2.06	21	2.05	1.65	2.06	1.62
560	24	2.27	1.76	2.25	1.93	36	1.97	1.76	1.98	1.67
672	35	2.07	1.78	2.08	1.90	55	1.81	1.48	1.83	1.44
896	70	1.83	1.50	1.83	1.53	105	1.46	1.19	1.47	1.16
1344	165	1.20	1.03	1.21	0.95	253	0.91	0.75	0.92	0.69

Table 12: **Cost of the window ensemble against one pass** on one RTX 4090. "wins" is the number of 224 windows at stride 112; FLOP× and time× are sliding-window cost divided by single-pass cost, so values above 1 favour us. FLOPs are counted analytically for the encoder; time is the mean of 50 timed forwards after warm-up, windows batched 32 at a time, one image at a time for the single pass. The saving holds at every resolution used in the paper, on both backbones and both aspect ratios, and closes only around short side 1344, where the single pass's quadratic attention has caught up with the ensemble's fourfold redundancy in FLOPs as well.

attention is quadratic in the token count where the ensemble's is linear (fixed 196 keys per window), so the two curves must cross; on these backbones they cross around short side 1344, about $2\times$ the largest resolution at which training-free dense CLIP is evaluated.

C Robustness of the fitted footprint

Table 13 repeats the SC-CLIP fit of Section 5.5 with the footprint initialised at $\sigma_0 = 4$ and 20 patches instead of 10, with a second random seed, and with the unlabelled training images drawn from Cityscapes rather than ADE20K. The fitted widths are the finding of Section 5.5; this table is what lets us call them a property of the encoder rather than of the run.

D Scale structure of the gain

Table 11 separates the pixels by region area. The reason the adapted single pass ends up ahead of sliding windows overall is not that it wins on small objects (it very nearly matches the windows there and does not exceed them) but that it does not pay their price on large ones. On ADE20K the adapted pass reaches 10.00 on the tiny bin against 10.58 for windows and 17.63 against 17.71 on the

	σ at 28×28			published protocol					@672		
Run	attn	pre	post	V20	PC59	Stuff	ADE	City†	ADE	City	avg
$\sigma_0 = 10$, seed 0 (main)	7.4	1.4	4.1	85.77	41.09	27.86	20.14	38.55	19.36	39.14	42.68
$\sigma_0 = 4$	7.2	1.9	3.9	85.79	41.09	27.87	20.15	38.39	19.35	38.99	42.66
$\sigma_0 = 20$	7.5	1.4	4.2	85.78	41.09	27.84	20.16	38.58	19.34	39.18	42.69
$\sigma_0 = 10$, seed 1	7.4	1.4	4.1	85.60	40.97	27.81	20.10	38.62	19.29	39.37	42.62
trained on unlabelled Cityscapes	6.9	1.1	4.3	85.14	40.96	27.73	20.14	40.06	19.40	40.70	42.81
frozen, slide (reference)	—	—	—	84.51	40.19	27.40	20.05	40.96	18.83	41.74	42.62

Table 13: **The footprint does not depend on where it starts, on the seed, or on the training images.** SC-CLIP, LoRA + footprint, mixed recipe. Widths are on the 28×28 training grid; †Cityscapes at short side 560. Across three initialisations and two seeds the attention footprint stays within 7.2–7.5 and the second aggregation within 3.9–4.2, and the five-benchmark average within 0.07. Distilled on the 2975 unlabelled Cityscapes training images instead of ADE20K, the widths are the same to within half a patch and the Cityscapes deficit against the ensemble falls to 0.9 mIoU at both resolutions, with the other four benchmarks unchanged.

	pretraining	native	ADE20K @448, mIoU				footprint σ	
Backbone	crop	grid	frozen 1	slide	LoRA	LoRA + foot.	patches	pixels
ViT-B/32	224	7×7	11.97	15.74	15.57	**15.80**	3.6	115
ViT-B/16	224	14×14	15.90	17.81	18.15	**18.27**	7.0	112
ViT-L/14	224	16×16	14.26	17.87	18.43	**19.09**	8.4	118
ViT-L/14@336	336	24×24	13.47	15.15	—	**15.69**	15.1	211
NACLIP	336	24×24	14.47	15.21	—	**15.83**	11.1	155

Table 14: **Four backbones.** ClearCLIP (and, on the last row, NACLIP) on ADE20K at short side 448; the LoRA alone, distilled on COCO, and the full method, distilled on unlabelled ADE20K training images (Section 4.4). The full method is at or above the ensemble on all four, including B/32, where the LoRA alone fell 0.17 short. For the three 224-crop models the fitted attention footprint is half the native grid, 112–118 pixels; for L/14@336, pretrained on 336 crops with a 24-patch grid and evaluated with 336 windows at stride 168, it is 15.1 patches on ClearCLIP and 11.1 on NACLIP, 211 and 155 pixels, against 8.4 and 6.5 for the same surgeries on L/14 at 224. The width follows the pretraining crop, not a fixed pixel extent.

small bin, while on the medium bin it scores 18.91 against 17.94 and on the large bin 12.05 against 11.77. It recovers the crop-level detail without cutting the image apart.

The same decomposition explains the Cityscapes deficit exactly: the shortfall sits in the tiny (16.25 against 19.02) and small (25.36 against 28.78) bins, while the medium and large bins are already ahead.

E BACKBONES

Table 14 repeats the ClearCLIP column of the main experiment for ViT-B/32 and ViT-L/14. With the LoRA alone the adapted single pass exceeds the window ensemble on B/16 and L/14 ($+0.34$ and $+0.56$) but not on B/32 (-0.17). All three see the same $2\times$ increase in grid side at short side 448, since all were pretrained on 224 crops; what is different about B/32 is that its tokens are $4\times$ coarser in pixels, so the 448 pass has only about 14×18 of them and the crop-level detail the ensemble recovers is, in patch units, that much finer than anything a single pass can represent. Adding the footprint closes that case ($+0.06$) and widens the other two ($+0.46$, $+1.22$); under the published protocol (Table 15) the full method on B/32 is above the ensemble on four of five benchmarks and on average (35.91 against 35.20), and 0.4–0.9 below on Cityscapes. The footprint the fit chooses is 3.6 patches on B/32's 7-patch grid, 7.0 on B/16's 14 and 8.4 on L/14's 16: half the native grid on each of three backbones whose grids span a factor of 2.3, with b between -0.11 and -0.20 on all three.

Table 15 runs the full method, LoRA and footprint together, on ViT-L/14 under the published protocol, for ClearCLIP, NACLIP and SCLIP. Training is the recipe of Section 4.4 with one change

	published protocol						higher resolution	
	V20	PC59	Stuff	ADE	City†	avg	ADE@448	City@672
ClearCLIP, ViT-L/14								
published (Lan et al., 2024a)	84.8	31.5	21.2	16.9	32.1	37.3	—	—
frozen, 1 pass	82.86	28.94	19.76	14.36	19.56	33.10	14.26	18.28
frozen, slide	85.04	31.26	21.76	16.71	32.70	37.49	17.87	34.28
WindowFold, 1 pass	**86.20**	**33.34**	**23.26**	**17.51**	**34.42**	**38.95**	**19.09**	**36.11**
NACLIP, ViT-L/14								
frozen, 1 pass	78.85	30.52	20.87	16.10	25.48	34.36	15.91	23.93
frozen, slide	78.45	31.87	21.98	17.15	31.89	36.27	17.64	33.10
WindowFold, 1 pass	**81.85**	**34.65**	**23.63**	**18.58**	**38.13**	**39.37**	**19.51**	**39.24**
SCLIP, ViT-L/14								
frozen, 1 pass	58.85	19.60	12.80	6.53	4.96	20.55	6.28	4.88
frozen, slide	70.32	24.75	16.72	10.77	21.02	28.72	12.38	23.43
WindowFold, 1 pass	**75.45**	**27.61**	**18.73**	**11.95**	**28.20**	**32.39**	**14.34**	**31.16**
ClearCLIP, ViT-B/32								
frozen, 1 pass	78.31	30.81	20.45	13.41	12.82	31.16	11.97	11.07
frozen, slide	80.07	31.46	21.29	14.74	**28.45**	35.20	15.74	**29.84**
WindowFold, 1 pass	**81.60**	**32.86**	**22.13**	**14.94**	28.02	**35.91**	**15.80**	28.98

Table 15: **The full method on other backbones.** LoRA + footprint, mixed recipe, distilled once per surgery on unlabelled ADE20K training images (Section 4.4); the ADE column is in-domain in that sense. †Cityscapes at short side 560. Our ClearCLIP L/14 sliding-window port is within 0.6 mIoU of the published row on every benchmark. On L/14 one forward is above the ensemble on all five benchmarks for all three surgeries, by +1.46 (ClearCLIP), +3.10 (NACLIP) and +3.67 (SCLIP) on average, and on Cityscapes at both resolutions, by up to +7.7. On B/32 it is above on four of five and on average (+0.71) and 0.4–0.9 below on Cityscapes. The fitted attention footprints on L/14 are $\sigma = 8.4$ ($b = -0.15$) on ClearCLIP, 6.5 ($b = -0.10$) on NACLIP and 8.0 ($b = -0.30$) on SCLIP, on a 32×32 training grid whose native side is 16; on B/32, native side 7, the fit is 3.6; on B/16, native side 14, the five surgeries chose 5–7.

for memory: the two largest shapes, 448×896 and 672×672, are drawn one image per step instead of two, so a step holds 8+6+1+1 images. The mismatch is larger on the larger model: the frozen ClearCLIP single pass trails the ensemble by 2.4 mIoU on ADE20K at short side 336 and by 3.6 at 448, against 1.1 and 1.9 for B/16 (Table 4), and the frozen SCLIP single pass collapses outright, to 4.96 on Cityscapes against the ensemble's 21.02. The repair is also larger: one forward beats the ensemble on all five benchmarks for all three surgeries, by +1.46 on ClearCLIP, +3.10 on NACLIP and +3.67 on SCLIP, against +0.92, +1.30 and +1.07 for the same surgeries on B/16 (Table 9), and, unlike on B/16, on Cityscapes at every resolution we measured (ClearCLIP 34.42 against 32.70 at 560 and 36.11 against 34.28 at 672; NACLIP 38.13 against 31.89 and 39.24 against 33.10; SCLIP 28.20 against 21.02 and 31.16 against 23.43). NACLIP's own Gaussian, $\sigma = 5$, was chosen on B/16; on L/14 the fit replaces it with 6.5, and ClearCLIP and SCLIP, which have no prior of their own, are given 8.4 and 8.0. All three are on a grid whose native side is 16; on B/16, whose native side is 14, the five surgeries chose 5–7. The footprint has moved with the pretraining grid, on a backbone with a different patch size, depth and width, and again lands near half of it, as it does on B/32 (3.6 of 7). Two accounts fit these three: the width could be half the token grid attention was trained over, or a fixed pixel extent, since all three were pretrained on 224-pixel crops and half the grid is the same 112 pixels on each. What they do rule out is a width tied to the image, which would need $b = 1$: the attention-site b is -0.06 to -0.51 on B/16, -0.10 to -0.30 on L/14 and -0.20 on B/32.

A backbone pretrained at another resolution separates the two. ViT-L/14@336 is the same architecture as ViT-L/14 with one further epoch of pretraining at 336 pixels, so its native grid is 24×24 and the sliding window the protocol implies is a 336 crop at stride 168. A fixed pixel extent predicts that each surgery keeps the width it had on L/14 at 224, 8.4 patches for ClearCLIP and 6.5 for NACLIP; half the grid predicts 12 for both. The fits (Table 14, last two rows; shapes 448^2, 560^2, 672^2 at batch 6, 2, 1) land at 15.1 patches (211 pixels, $b = -0.17$) on ClearCLIP and 11.1 (155 pixels, $b = -0.02$) on NACLIP: both have moved with the pretraining crop, by 1.8× and 1.7× for a 1.5× larger crop, and neither is a pixel constant. As fractions of the native grid they are 0.63

Surgery at test	frozen	transferred	matched	closed
SCLIP	13.03	15.06	17.29	48%
MaskCLIP	12.86	13.71	14.25	61%

Table 16: **Cross-surgery transfer.** A ClearCLIP-trained adapter applied, without retraining, on top of a different surgery. About half of what the adapter learns is a surgery-independent correction for the resolution mismatch. ADE20K, short side 448.

and 0.46, against 0.53 and 0.41 for the same surgeries at 224, so each surgery also keeps its own relative width across the two pretraining resolutions. The method's result on this backbone is the usual one: one forward beats the 336-window ensemble on every setting we ran for both surgeries: for ClearCLIP, ADE20K 15.69 against 15.15 at 448 and 17.38 against 16.24 at 672, Cityscapes 29.44 against 26.72 at 672, Context-59 29.08 against 26.86 and VOC20 79.11 against 76.40 at 448; for NACLIP, 15.83 against 15.21, 16.28 against 15.15, 29.41 against 24.03, 26.37 against 24.72 and 65.99 against 63.61 in the same order.

F Should the correction scale with resolution? (A negative result)

The diagnosis says the mismatch grows with the grid, roughly with $s = \log(N/N_{\text{native}})$, so a fixed gain must compromise across the resolutions it serves. We let the gain follow the quantity it corrects, giving every rank direction of every LoRA site an affine schedule in s: the update BAx becomes $B\,\mathrm{diag}(\gamma_0 + \gamma_1 s)Ax$, centred so that $\gamma_0 = 1, \gamma_1 = 0$ recovers the fixed-gain adapter exactly. The two models are nested and trained identically, 1536 extra parameters on 1.78M.

It buys nothing: 17.87, 18.11 and 17.75 on ADE20K at short sides 336, 448 and 672, against 17.92, 18.10 and 17.69 fixed, a mean difference of $+0.01$ mIoU. The coefficients are more informative than the scores. The slope is positive at 47 of 48 sites (mean $\gamma_1 = 0.088$), so the optimizer does use the freedom, in the direction the diagnosis predicts; but the gain only travels from 0.96 at 336 to 1.08 at 672. How hard to correct is evidently not the question, which agrees with the position-grid control: a larger grid changes where attention puts its mass, and rescaling a fixed update cannot express that.

G Is the adapter specific to one surgery?

The adapter is trained against one surgery's teacher, but the LoRA touches all twelve blocks while the surgery touches only the last. Table 16 moves a ClearCLIP-trained adapter onto SCLIP and MaskCLIP without retraining. It recovers 48% and 61% of the gap that a matched adapter recovers. Roughly half of what is learned is a surgery-independent correction for the resolution mismatch, and the remainder is specific. Matched training is worth doing and is cheap, but a single adapter is a usable default.

H Qualitative examples

Figure 3 shows the three arms on SC-CLIP. The selection is mechanical: among the first 100 validation images of each dataset (landscape orientation), we show the image at the median and the image at the 90th percentile of the pixel-accuracy gap between the frozen single pass and the window ensemble, so each dataset contributes a typical case and a severe one. The frozen single pass fails in the way the diagnosis predicts: whole regions are assigned one wrong label (a street scene almost entirely as sidewalk; a gallery wall and its paintings merged), which is what attention spreading its mass over too many keys looks like at the pixel level. The adapted pass restores the ensemble's structure with one forward, and on the gallery image its regions are cleaner than the ensemble's, whose crop seams are visible as vertical bands on the wall.

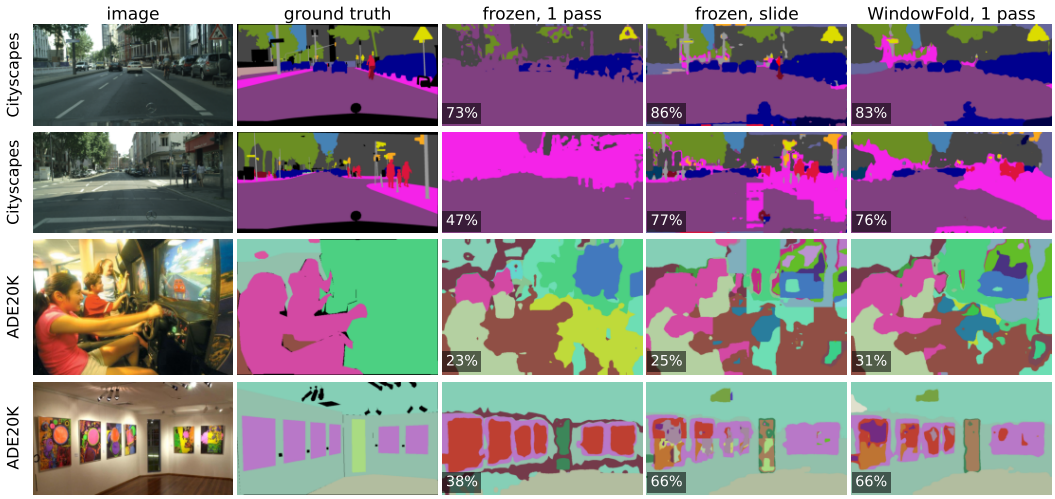


Figure 3: **Qualitative examples, SC-CLIP.** Cityscapes at short side 672 (top two rows) and ADE20K at 448 (bottom two). Columns: image, ground truth, frozen single pass, frozen sliding window, and one forward through WindowFold (LoRA + footprint, mixed recipe). The number in each panel is that image's pixel accuracy. Rows are the median and 90th-percentile images by single-pass deficit among the first 100 validation images, not a curated selection.