VOCABAUDIT: AUDITING INFERENCE-VOCABULARY SENSITIVITY IN TRAINING-FREE OPEN-VOCABULARY SEGMENTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Open-vocabulary semantic segmentation (OVSS) is evaluated with a class-name list that directly controls test-time predictions, yet this list is often treated as fixed metadata. We study vocabulary sensitivity by freezing the images, model weights, prompt templates, and dense-inference pipeline while varying only the inference vocabulary. On a 300-image VOC-21 matrix, replacing an engineered vocabulary with plain dataset names reduces mIoU by 11.8–20.6 points. Filtered synonym substitutions cost up to 6 points and produce a 3.7-point spread across vocabulary draws. We then run VocabAudit V1.1 on 1,349 VOC images, three methods, six vocabulary-expansion conditions, and ten independent vocabulary draws per condition. Adding classes absent from the VOC-21 label set increases target-class mIoU by 4.49–12.75 points, with all 18 paired image-bootstrap intervals excluding zero, and reverses the cell-mean SCLIP–NACLIP ordering in every condition. At the same time, the expanded vocabulary substantially changes the predictions: with 200 added classes, pixel accuracy falls to 16.84–24.05%, and distractor labels receive 71.62–91.48% of benchmark-background pixels and 18.81–35.42% of foreground pixels. All-class mIoU also falls to 4.04–4.59, partly because the metric assigns zero IoU to label-set-absent classes. These results show that target-class mIoU alone can obscure changes in the prediction interface and in pixel-level errors under vocabulary expansion. Reliable OVSS evaluation should therefore release the exact inference vocabulary, report sensitivity across controlled vocabulary draws, distinguish target-class from all-class accounting, and include an explicit pixel-flow ledger.

1 INTRODUCTION

Open-vocabulary semantic segmentation (OVSS) promises dense predictions for class sets supplied at inference time. A growing family of *training-free* methods (Zhou et al., 2022; Wang et al., 2024; Lan et al., 2024a; Hajimiri et al., 2025; Lan et al., 2024b) adapts CLIP (Radford et al., 2021) to dense prediction without any segmentation training and reports zero-shot mIoU on standard benchmarks. Yet these evaluations typically fix the inference vocabulary, namely the class-name list presented to the text encoder. Standard leaderboards therefore do not test whether a method is robust to the input that makes the task open-vocabulary.

This omission confounds method quality with vocabulary engineering. Public implementations ship hand-tuned name files, including an expansion of the VOC "background" class into 26 scene subclasses (Wang et al., 2024), while prompt augmentation changes zero-shot classification accuracy (Roth et al., 2023; Pratt et al., 2023). If semantically equivalent or mildly expanded vocabularies also change dense predictions and method rankings, a reported OVSS score describes one name file rather than a stable model capability. A controlled audit must therefore vary the vocabulary while holding every other inference choice fixed.

We introduce VocabAudit, a protocol that freezes the backbone, prompt templates, resolution, sliding-window inference, and logit pooling, then perturbs only the inference vocabulary. Three axes probe complementary failure modes: filtered synonym substitution tests lexical equivalence, a frozen WordNet hypernym rule tests granularity, and the injection of 50–200 absent class names

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.

tests output-space expansion. The core audit covers three representative training-free methods and four benchmarks with 21–459 classes: VOC-21 (Everingham et al., 2010), COCO-171 (Lin et al., 2014; Caesar et al., 2018), ADE-150 (Zhou et al., 2017), and PC-459 (Mottaghi et al., 2014). VocabAudit V1.1 then confirms the distractor effect on the full dev-excluded VOC split with repeated vocabulary draws, paired uncertainty, two mIoU conventions, and a pixel-flow ledger.

Our main findings are:

1. **Vocabulary choice can outweigh method choice.** Replacing engineered name files with plain dataset names costs 11.8–20.6 mIoU on the core VOC-21 matrix, and synonym vocabulary draws span up to 3.7 points.
2. **Vocabulary expansion exposes a score paradox.** Across 180 method–draw evaluations, absent class names increase target-class mIoU while all-class mIoU and pixel accuracy collapse. Every distractor-cell mean also reverses the SCLIP–NACLIP ordering.
3. **Pixel flows explain the paradox.** Distractor labels act as background and foreground error sinks. Excluding them from mIoU can reward error absorption, whereas including every absent class mechanically dilutes the score.
4. **Simple repairs do not resolve the interface failure.** Pre-registered text-embedding whitening improves geometric diagnostics but reduces segmentation quality. Synthesized background negatives recover much of the macro effect, but none of the tested label-free mechanisms provides reliable per-class protection.

2 Related Work

Training-free OVSS. Open-vocabulary segmentation was first pursued with language-aligned decoders or mask classifiers trained on segmentation and image–text data (Li et al., 2022; Ghiasi et al., 2022; Xu et al., 2022; Liang et al., 2023; Xu et al., 2023; Cho et al., 2024), or learned from image–text pairs alone (Shin et al., 2022; Cha et al., 2023). Training-free methods instead reinterpret a frozen CLIP model (Radford et al., 2021). MaskCLIP (Zhou et al., 2022), SCLIP (Wang et al., 2024), ClearCLIP (Lan et al., 2024a), and NACLIP (Hajimiri et al., 2025) modify the final attention pathway; ProxyCLIP (Lan et al., 2024b), GEM (Bousselham et al., 2024), and CLIP-DINOiser (Wysoczańska et al., 2024) import or refine spatial affinities. Their reported evaluations use fixed official vocabularies. Our core matrix controls the common text–vision interface across representative final-block variants, and our claims are bounded to the audited methods (§6).

Text-side sensitivity and name quality. On the classification side, prompt template ensembles (Radford et al., 2021), learned soft prompts (Zhou et al., 2022), generated descriptions (Pratt et al., 2023), random descriptors (Roth et al., 2023), hierarchical names (Ge et al., 2023), concept-frequency analysis (Parashar et al., 2024), and synonym aggregation (Yin et al., 2026) all show that text formulation changes zero-shot accuracy; vocabulary-free classification (Conti et al., 2023) questions the assumption of a supplied class set. In segmentation, RENOVATE (Huang et al., 2024) aligns benchmark names with masks, FLOSS (Benigmim et al., 2025) selects per-class templates, and SynCLIP (Xie et al., 2026) reduces synonym inconsistency through re-pretraining. ReME (Xuan et al., 2025) uses name-mediated reference retrieval, while FreeCP (Chen et al., 2025) purifies class queries using precomputed per-name descriptions. These methods improve or adapt to a vocabulary. VocabAudit instead freezes the model and measures its response to controlled lexical, granularity, and output-space changes, including variation across independently sampled vocabularies.

Background modeling in OVSS. That an explicit background model matters is well established: CaR (Sun et al., 2024) adds hand-written background queries, OVDiff (Karazija et al., 2024) derives negative visual prototypes, and SCLIP's official VOC vocabulary expands "background" into 26 scene sub-classes (Wang et al., 2024). Our constructive experiment does not propose a new background module. It uses pre-registered text-side synthesis and random controls to separate recoverable macro background calibration from unresolved per-class protection. Relatedly, whitening can improve the geometry of anisotropic text embeddings (Su et al., 2021). We test whether that geometric improvement transfers to dense segmentation rather than assuming it does.

Evaluation audits and metrics. MESS (Blumenstiel et al., 2023) audits zero-shot segmentation across domains; VocabAudit varies the complementary input, the inference vocabulary. Open mIoU (Zhou et al., 2025) addresses synonymous predictions, and mask-wise evaluation (Huang et al., 2025) addresses ambiguous vocabulary expansion. BeyondBench (Jia et al., 2026) tests semantic variation in classification and retrieval, while OVEarth-Bench (Li et al., 2026) tests query diversity within a fixed remote-sensing benchmark. Our work instead asks whether OVSS conclusions survive controlled changes to the inference vocabulary, paralleling image-side test-set audits (Recht et al., 2019; Beyer et al., 2020). It complements static name-quality studies (Huang et al., 2024) and open-set-aware metrics by identifying how the vocabulary and metric interact at the dense output interface.

3 Audit Protocol

Frozen inference protocol. OpenCLIP (Ilharco et al., 2021) ViT-B/16 (OpenAI weights, QuickGELU), the 80 ImageNet prompt templates (Radford et al., 2021), a short side of 336 pixels, a 224×224 sliding window with stride 112, mean logit fusion, and temperature-40 softmax over sub-queries followed by per-class max pooling (Wang et al., 2024). Every method and condition uses this protocol. Legacy tables report GT-present mIoU, which averages classes with nonempty ground-truth support. For VocabAudit V1.1, let $\mathcal{T}$ be the 21 target classes, $\mathcal{V} = \mathcal{T} \cup \mathcal{D}$ the inference vocabulary, and I_c, U_c the dataset-level intersection and union. We report $mIoU_{\text{target}} = |\mathcal{T}|^{-1} \sum_{c \in \mathcal{T}} I_c/U_c$ and $mIoU_{\text{all}} = |\mathcal{V}|^{-1} \sum_{c \in \mathcal{V}} I_c/\max(U_c, 1)$; an absent class with empty union receives zero. Since distractors have no ground-truth support, $mIoU_{\text{all}} = 21/(21+|\mathcal{D}|)\, mIoU_{\text{target}}$. We use no PAMR (Araslanov and Roth, 2020) or other post-processing. Our SCLIP reproduction reaches 55.4 mIoU on VOC-21 with the official name file, compared with 59.1 reported with additional refinement. This gap does not affect controlled within-pipeline comparisons. Textual deltas use unrounded run records and can differ by 0.1 from displayed values.

Axis 1: synonym substitution. For each class, candidate synonyms are WordNet (Miller, 1995) noun-synset lemmas filtered by CLIP text cosine similarity to the original name within $[0.70, 0.95]$; the highest-similarity candidate is chosen. Classes without valid candidates, and non-semantic classes (`background`, `unknown`, `other`), are never substituted. We replace 25/50/100% of substitutable classes with three random seeds at 50%. Ground-truth labels are unchanged.

Axis 2: granularity shift. Each class name is replaced by the first lemma of the first-sense direct WordNet hypernym (frozen rule, no manual correction), with the same exclusion list. The mapping table is part of the released protocol (Appendix H); merged-credit scoring (crediting predictions of a coarse name shared by several classes) is left to future work.

Axis 3: distractor injection. Distractors are unique names from the union of ADE-847 and PC-459 after removing target names and candidates whose maximum CLIP text cosine to any target exceeds 0.92. We sort the remaining pool by this maximum similarity and freeze rank terciles as *far*, *mid*, and *near*. For each tercile and size ($+50$ or $+200$), we sample without replacement using vocabulary-draw seeds 0–9. VocabAudit V1.1 evaluates all 60 vocabularies on the dev-excluded VOC-21 split (1,349 images) for SCLIP, NACLIP, and ClearCLIP. We report target-class mIoU, all-class mIoU, pixel-flow counts, and 10,000-sample paired image-bootstrap intervals. The same bootstrap image multiplicities are used for the plain baseline and every perturbation arm. Alias candidates are excluded from V1.1 results until three independent blinded judges clear the pre-registered semantic-validity gate.

Pre-registered repair hypotheses. Before running segmentation experiments we froze two kill criteria: (KC1) vocabulary-conditioned ZCA whitening of class embeddings must gain > 0.5 mIoU on large-vocabulary benchmarks across three methods; (KC2) vocabulary-conditioned statistics must beat global statistics in at least one regime. The pre-registered verdict was taken at shrinkage 0.5; a post-verdict sensitivity sweep over $\{0.1, 0.3, 0.7, 0.9\}$ (PC-459, ClearCLIP) confirmed the conclusion is not a shrinkage artifact.

Table 1: VOC-21 results on a fixed 300-image subset using GT-present mIoU. The official SCLIP vocabulary expands background into 26 sub-classes and enriches `person` and `tv`; the plain vocabulary uses dataset names. All other inference settings are identical.

Vocabulary	SCLIP	ClearCLIP	MaskCLIP
Official (engineered)	**55.4**	53.6	43.6
Plain dataset names	34.8	**34.9**	31.8
Δ	-20.6	-18.7	-11.8

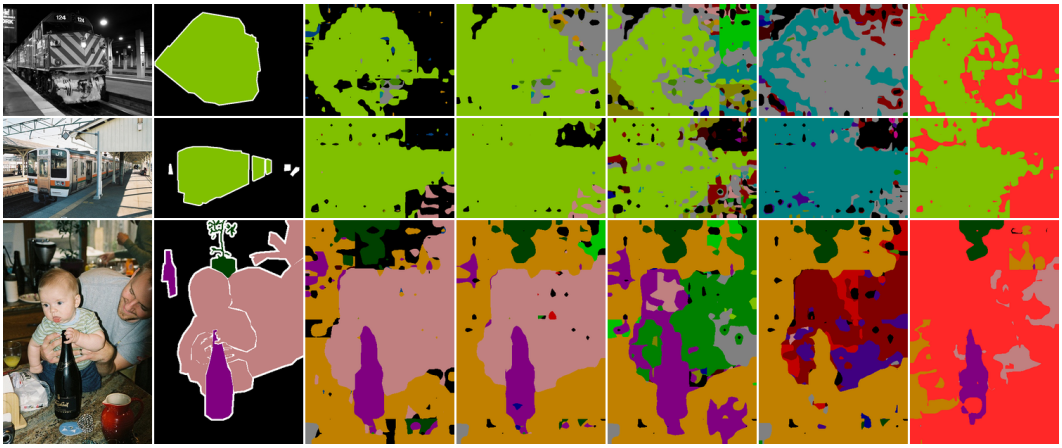


Figure 1: The same frozen SCLIP pipeline under five vocabularies (VOC-21 val). Columns: image, ground truth, official (engineered) names, plain dataset names, 100% synonym substitution, one-level WordNet hypernyms, plain $+200$ near distractors (distractor-class predictions in red). Only the list of class names changes between columns 3–7; every visible degradation is a vocabulary effect, not a model change.

4 Results

4.1 Naming conventions are worth more than method choice

Replacing the engineered name file with plain dataset names reduces mIoU by 11.8–20.6 points (Table 1). This vocabulary effect exceeds the gaps among the three methods under either naming regime. It complements RENOVATE (Huang et al., 2024): instead of renaming classes to improve mask alignment, we measure how much of a training-free result is attributable to the supplied name file under a frozen protocol. Model selection also depends on that file. SCLIP leads ClearCLIP by 1.8 points under official names, whereas their plain-name results differ by only 0.1 points, below the subset-resampling noise floor. Thus, a leaderboard entry without its exact inference vocabulary is not a reproducible method comparison.

4.2 Synonym and granularity axes

Full synonym substitution reduces mIoU by 4–6 points, and three vocabulary draws at 50% replacement span up to 3.7 points (Table 2). Because every draw uses the same images, the spread arises from vocabulary choice rather than image sampling (per-method mean$\pm$std across draws: SCLIP 32.1 ± 2.0, ClearCLIP 32.4 ± 1.1, MaskCLIP 28.8 ± 1.0); on ADE-150 the one-level hypernym rule collapses performance from 13.1/14.7/10.2 to 2.5/3.8/1.9. Both axes use frozen, released rules. They are automatic stress tests, not estimates over all valid human phrasings, but they show that plausible lexical and granularity changes can exceed several method gaps in the audited matrix.

Table 2: VOC-21 audit matrix (300 images, plain-name baseline, GT-present mIoU). Synonym substitution at rate r (seed 0; seeds 0/1/2 shown for 50%); one-level WordNet hypernym substitution (granularity). Labels are unchanged in all conditions. Single-subset values; the image-resampling noise floor is 3.1 mIoU (Fig. 2a), so differences below that are not interpreted.

Condition	SCLIP	ClearCLIP	MaskCLIP
plain (baseline)	34.8	34.9	31.8
syn 25%	34.1	32.9	29.6
syn 50% (s0/s1/s2)	34.4/31.3/30.7	33.3/32.8/31.2	29.4/29.3/27.7
syn 100%	30.6	29.0	26.3
hypernym (coarse)	17.0	17.0	16.9

Table 3: VocabAudit V1.1 distractor results on VOC-21 dev-excluded (1,349 images). Each perturbed entry is the mean over 10 fixed vocabulary draws. Entries report target-class mIoU, with all-class mIoU in parentheses. The former averages the 21 target classes; the latter also includes injected classes.

Condition	SCLIP	NACLIP	ClearCLIP
plain	35.60	36.93	34.75
far +50	42.04 (12.43)	41.42 (12.25)	39.27 (11.62)
mid +50	43.89 (12.98)	43.16 (12.76)	41.16 (12.17)
near +50	42.91 (12.69)	42.82 (12.66)	40.90 (12.10)
far +200	46.47 (4.42)	44.14 (4.19)	42.55 (4.04)
mid +200	48.34 (4.59)	46.10 (4.38)	44.88 (4.26)
near +200	46.53 (4.42)	44.88 (4.27)	43.97 (4.18)

4.3 Distractor injection exposes metric-interface incoherence

Across all 18 method–cell combinations, distractors increase target-class mIoU. The smallest cell mean is NACLIP far+50, $+4.49$ points with 95% paired CI $[4.08, 4.88]$; the largest is SCLIP mid+200, $+12.75$ $[11.80, 13.70]$. All 180 individual method–draw deltas are positive, so the direction is consistent across vocabulary draws. Cell means also change model selection: the plain order is NACLIP $>$ SCLIP $>$ ClearCLIP, whereas all six distractor-cell means yield SCLIP $>$ NACLIP $>$ ClearCLIP ($\tau_a = 1/3$ against the plain order). This is not a universal draw-level reversal: SCLIP beats NACLIP in 9/10 far+50, 8/10 mid+50, and 4/10 near+50 draws, but in 10/10 draws for every $+200$ cell.

The target-class gain masks worse segmentation. With 200 distractors, all-class mIoU falls to 4.04–4.59. In the near stratum, pixel accuracy falls by 34.38–41.83 points because distractors absorb most background pixels and 28.99–35.42% of foreground pixels (Table 4). Excluding distractor classes from the mIoU average can reward this error sink; including 200 absent classes instead dilutes the average by construction. We call the resulting mismatch between the expanded output interface and its score *metric-interface incoherence*. Neither mIoU convention is sufficient alone. The paired intervals and complete cell results are in Appendix G.

4.4 Pre-registered repairs fail; geometry is not a proxy

Raw class embeddings have the expected cone geometry: mean pairwise cosine ranges from 0.77 on A-847 to 0.82 on VOC-21, and the 10th-percentile nearest-neighbour margin shrinks from 0.078 to 0.037 as vocabularies grow. ZCA whitening improves every tested geometric diagnostic, including an A-847 margin increase from 0.09 to 0.55 and an effective-rank increase from 154 to 309. Nevertheless, it reduces mIoU in all nine method–dataset cells (Table 5); centering reduces mIoU in eight. Vocabulary-conditioned statistics outperform global statistics only on COCO-171, where both remain 2–4 points below the unwhitened baseline. A post-verdict shrinkage sweep preserves this conclusion. Thus, embedding separability does not certify dense transfer. This negative result is limited to the tabulated mid-to-large vocabularies, text-only whitening, and an unrecalibrated logit temperature.

Table 4: Pixel-flow mechanism for the near+200 cell (means over 10 draws). Accuracy is over all valid pixels. FG→D and BG→D are the fractions of foreground and background pixels assigned to injected distractor labels.

Method	plain acc.	near+200 acc.	FG→D	BG→D
SCLIP	53.63	18.27	29.37	89.72
NACLIP	58.67	16.84	35.42	91.48
ClearCLIP	52.62	18.24	28.99	86.77

Table 5: Text-side repairs under the frozen protocol (first 500 validation images; GT-present mIoU). ZCA uses shrinkage 0.5; global statistics from the ADE-847∪COCO-171 vocabulary. Whitening hurts in every cell (KC1 fails); vocabulary-conditioning edges out global statistics only on COCO-171, where both remain clearly below the unwhitened baseline.

Dataset	Method	none	center	ZCA (vocab)	ZCA (global)
PC-459	SCLIP	12.8	**13.7**	11.6	11.9
	ClearCLIP	**15.3**	15.2	13.2	13.7
	MaskCLIP	**9.6**	9.3	8.2	8.4
ADE-150	SCLIP	**13.1**	12.4	10.8	10.9
	ClearCLIP	**14.7**	13.9	12.7	12.9
	MaskCLIP	**10.2**	9.2	8.1	8.3
COCO-171	SCLIP	**22.9**	21.2	19.9	18.9
	ClearCLIP	**24.2**	23.1	21.7	21.4
	MaskCLIP	**17.4**	15.5	15.2	14.4

4.5 Can vocabulary engineering be automated? A constructive attempt

The 12–21 point value of engineered vocabularies motivates a constructive question: can a training-free, text-only procedure recover this value for an arbitrary user vocabulary? We pre-registered Vocabulary-Adaptive Background Synthesis (VABS): build a candidate lexicon of scene/stuff words (COCO-Stuff ∪ ADE-150 names), remove candidates within cosine τ_{sim} of any target class, greedily select M diverse negatives (facility-location objective), append them as sub-queries of the background class, and fold their predictions into background. Hyper-parameters were selected on a 100-image development set disjoint from the 300-image test set (M=64, τ_{sim}=0.9).

VABS raises VOC-21 plain-vocabulary mIoU by 12.7–18.8 points (macro $+16.1$), while matched random negatives recover $+12.6$ macro (Appendix Table 6). On the 300-image COCO-Object protocol, VABS gains 5.0–10.7 points (macro $+8.9$), but an ADE-only lexicon loses its advantage over random negatives. A labelled-development-set background-logit boost reproduces 73–85% of the VOC gain but harms every Context-60/ADE-150 transfer cell. These controls indicate that much of the macro gain comes from background calibration rather than semantic selection.

Per-class safety remains unresolved. Every registered text-, prediction-, or visual-side mechanism failed at least one frozen recovery or no-harm criterion. The tested negative vocabularies recover macro background calibration but not reliable class-specific protection. Appendix A reports the complete negative-result chain and pre-registered verdicts.

5 Discussion and Recommendations

The audit separates two failures that a single benchmark score conflates. First, model selection depends on the supplied names: every full-split distractor-cell mean swaps the SCLIP–NACLIP order, although draw-level stability is weaker at $+50$. Second, the score depends on how the expanded output space is accounted for. At fixed vocabulary size, target-class and all-class mIoU induce the same ranking because the latter is the former scaled by $21/(21 + |\mathcal{D}|)$, yet they imply opposite accounts of absolute quality. Pixel accuracy and the flow ledger reveal that neither account alone describes the segmentation failure.

The mechanism evidence narrows the interpretation. Synthesized negatives reproduce much of the macro background effect, whereas embedding whitening improves text geometry without improving

segmentation. The sensitivity therefore does not reduce to one geometric defect or one effective repair. Author-code and external-family anchors in Appendix B preserve the direction of the effect after changes to spatial processing, but they do not establish a universal effect size across OVSS families.

For benchmark users, the class-name file is part of the evaluated method and must be released. Robustness reports should use multiple frozen vocabulary draws, distribution summaries, and paired image-level uncertainty. **For method designers,** gains near the roughly 3-point subset noise floor should not be generalized from one official vocabulary, and background handling should be evaluated separately. **For metric designers,** target-class mIoU can reward error absorption, whereas all-class mIoU is mechanically sensitive to the number of absent classes. Both should be reported with pixel accuracy and foreground/background distractor flows until an open-set metric charges these errors without a vocabulary-size artifact.

6 LIMITATIONS

The core matrix covers three final-block CLIP variants under one ViT-B/16 interface. NACLIP, ViT-L/14, later training-free pipelines, author-code anchors, grounding-supervised detectors, and a Spanish probe appear in Appendix B. These extensions test direction, not a shared effect size for all OVSS systems. VocabAudit V1.1 confirms only the distractor axis on VOC-21 and three methods. Its alias ledger has not passed the pre-registered three-judge blinded review, so we report no V1.1 alias result. The synonym axis is an automatic lexical stress test rather than a human-validated equivalence study. A-847 segmentation was unavailable, and the fixed first-sense WordNet rule can select non-visual parents; the granularity axis is therefore a reproducible stress mapping, not validated semantic coarsening.

Most legacy conditions use fixed 300–500-image validation subsets. The V1.1 distractor matrix instead uses all 1,349 dev-excluded VOC images and paired bootstrap intervals; full-VOC calibrations and disjoint-subset checks in Appendix B characterize sampling variation for legacy cells. Effects near 3 mIoU remain uncertain on one subset. We did not ablate the frozen perturbation thresholds. Whitening was applied only to text queries and without temperature retuning. Finally, the novelty assessment covers the cited published and publicly accessible literature; unavailable in-review work cannot be ruled out.

7 CONCLUSION

The inference vocabulary is an experimental variable, not passive benchmark metadata. In the audited training-free OVSS methods, vocabulary engineering can outweigh method differences, plausible substitutions can exceed several method gaps, and output-space expansion can raise target-class mIoU while pixel agreement collapses. Pixel flows identify this paradox as an interface-and-accounting failure rather than improved robustness. Text-geometry repair does not resolve it, and synthesized negatives recover macro background calibration without reliable per-class protection. Claims about user-supplied open-vocabulary behavior should therefore release the exact vocabulary and report sensitivity across vocabulary draws, complementary metrics, and open-set pixel flows. The conclusion is limited to the audited methods, datasets, and validated perturbation axes.

AI USE STATEMENT

Generative AI tools assisted research brainstorming, code development, experiment analysis, language editing, and LaTeX organization. AI tools did not serve as authors. The human authors retain responsibility for the study design, source verification, reported results, and final manuscript.

ETHICS STATEMENT

This audit uses established public vision benchmarks and does not involve human subjects or the collection of personal data. Its purpose is to expose evaluation sensitivity that could otherwise lead to misleading comparisons of open-vocabulary systems; dataset licenses and original dataset documentation remain applicable.

Reproducibility statement

The supplementary artifact contains the frozen perturbation generators, vocabulary files, preregistrations, evaluation harness, fixed subset definitions, and run-level JSON records used by the tables. V1.1 also releases cache signatures, blinded alias review sheets, per-image sufficient statistics, pixel-flow ledgers, and the frozen 10,000-sample paired-bootstrap output. Appendix H indexes these materials, while the protocol and split conventions are specified in the main text and appendix.

References

C. Zhou, C. C. Loy, B. Dai. Extract free dense labels from CLIP. In *ECCV*, 2022.

F. Wang, J. Mei, A. Yuille. SCLIP: Rethinking self-attention for dense vision-language inference. In *ECCV*, 2024.

M. Lan, C. Chen, Y. Ke, X. Wang, L. Feng, W. Zhang. ClearCLIP: Decomposing CLIP representations for dense vision-language inference. In *ECCV*, 2024.

M. Lan, C. Chen, Y. Ke, X. Wang, L. Feng, W. Zhang. ProxyCLIP: Proxy attention improves CLIP for open-vocabulary segmentation. In *ECCV*, 2024.

S. Hajimiri, I. Ben Ayed, J. Dolz. Pay attention to your neighbours: Training-free open-vocabulary semantic segmentation. In *WACV*, 2025.

S. Sun, R. Li, P. Torr, X. Gu, S. Li. CLIP as RNN: Segment countless visual concepts without training endeavor. In *CVPR*, 2024.

L. Karazija, I. Laina, A. Vedaldi, C. Rupprecht. Diffusion models for open-vocabulary segmentation. In *ECCV*, 2024.

A. Radford et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

K. Roth, J. M. Kim, A. S. Koepke, O. Vinyals, C. Schmid, Z. Akata. Waffling around for performance: Visual classification with random words and broad concepts. In *ICCV*, 2023.

S. Pratt, I. Covert, R. Liu, A. Farhadi. What does a platypus look like? Generating customized prompts for zero-shot image classification. In *ICCV*, 2023.

J. Su, J. Cao, W. Liu, Y. Ou. Whitening sentence representations for better semantics and faster retrieval. *arXiv:2103.15316*, 2021.

H. Huang, S. Peng, D. Zhang, A. Geiger. Renovating names in open-vocabulary segmentation benchmarks. In *NeurIPS*, 2024.

X. Yin, Q. Wang, B. Cao, Q. Hu. S^3: Synonymous semantic space for improving zero-shot generalization of vision-language models. In *ICASSP*, 2026.

Y. Benigmim, M. Fahes, T.-H. Vu, A. Bursuc, R. de Charette. FLOSS: Free lunch in open-vocabulary semantic segmentation. In *ICCV*, 2025.

Q. Huang, H. Hu, J. Jiao. Revisit the open nature of open vocabulary semantic segmentation. In *ICLR*, 2025.

V. Stojnić, Y. Kalantidis, G. Tolias. LPOSS: Label propagation over patches and pixels for open-vocabulary semantic segmentation. In *CVPR*, 2025.

S. Bai, Y. Liu, Y. Han, H. Zhang, Y. Tang. Self-calibrated CLIP for training-free open-vocabulary segmentation. arXiv:2411.15869, 2024.

H. Zhou, T. Shen, X. Yang, H. Huang, X. Li, L. Qi, M.-H. Yang. Rethinking evaluation metrics of open-vocabulary segmentation. *TPAMI*, 2025 (arXiv:2311.03352).

Q. Chen, L. Yang, Y. Chen, N. Zhao, J. Lai, J. Shao, X. Xie. FreeCP: Training-free class purification for open-vocabulary semantic segmentation. In *ICCV*, 2025.

X. Xuan, Z. Deng, K.-L. Ma. ReME: A data-centric framework for training-free open-vocabulary segmentation. In *ICCV*, 2025.

M. Xie, G. He, D. Xu, Y. Lin, H. Li, P. Feng, J. Guan. SynCLIP: Synonym-coherent language-image pretraining for robust open-vocabulary dense perception. arXiv:2607.11008, 2026.

C. Jia, A. MaungMaung, H. H. Nguyen, J. Chen, I. Echizen. Beyond standard benchmarks: A systematic audit of VLM robustness to natural semantic variation. In *ICPR*, 2026.

B. Blumenstiel, J. Jakubik, H. Kühne, M. Vössing. What a MESS: Multi-domain evaluation of zero-shot semantic segmentation. In *NeurIPS Datasets and Benchmarks*, 2023.

Y. Ge, J. Ren, A. Gallagher, Y. Wang, M.-H. Yang, H. Adam, L. Itti, B. Lakshminarayanan, J. Zhao. Improving zero-shot generalization and robustness of multi-modal models. In *CVPR*, 2023.

S. Parashar, Z. Lin, T. Liu, X. Dong, Y. Li, D. Ramanan, J. Caverlee, S. Kong. The neglected tails in vision-language models. In *CVPR*, 2024.

A. Conti, E. Fini, M. Mancini, P. Rota, Y. Wang, E. Ricci. Vocabulary-free image classification. In *NeurIPS*, 2023.

W. Bousselham, F. Petersen, V. Ferrari, H. Kuehne. Grounding everything: Emerging localization properties in vision-language transformers. In *CVPR*, 2024.

M. Wysoczańska, O. Siméoni, M. Ramamonjisoa, A. Bursuc, T. Trzciński, P. Pérez. CLIP-DINOiser: Teaching CLIP a few DINO tricks for open-vocabulary semantic segmentation. In *ECCV*, 2024.

Y. Shi, M. Dong, C. Xu. Harnessing vision foundation models for high-performance, training-free open vocabulary segmentation. In *ICCV*, 2025.

D. Zhang, F. Liu, Q. Tang. CorrCLIP: Reconstructing patch correlations in CLIP for open-vocabulary semantic segmentation. In *ICCV*, 2025.

M. Zamini, D. Shukla. SCI-CLIP: Segment-centric inference with reference memory for training-free open-vocabulary segmentation. arXiv:2608.05627, 2026.

J. Li, Y. Lu, Y. Zhang, Y. Xie, F. Wang, Y. Xie, Y. Qu. Target refocusing via attention redistribution for open-vocabulary semantic segmentation: An explainability perspective. In *AAAI*, 2026.

K. Li, Z. Xin, Z. Jiang, J. Fu, L. Xue, L. Zhang, X. Cao. OVEarth-Bench: Evaluating category breadth and query diversity for open-vocabulary earth observation. arXiv:2607.27278, 2026.

M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL visual object classes (VOC) challenge. *IJCV*, 88(2):303–338, 2010.

R. Mottaghi, X. Chen, X. Liu, N.-G. Cho, S.-W. Lee, S. Fidler, R. Urtasun, and A. Yuille. The role of context for object detection and semantic segmentation in the wild. In *CVPR*, 2014.

T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft COCO: Common objects in context. In *ECCV*, 2014.

H. Caesar, J. Uijlings, and V. Ferrari. COCO-Stuff: Thing and stuff classes in context. In *CVPR*, 2018.

B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba. Scene parsing through ADE20K dataset. In *CVPR*, 2017.

G. Ilharco, M. Wortsman, R. Wightman, C. Gordon, N. Carlini, R. Taori, A. Dave, V. Shankar, H. Namkoong, J. Miller, H. Hajishirzi, A. Farhadi, and L. Schmidt. OpenCLIP. Zenodo, 2021. `https://doi.org/10.5281/zenodo.5143773`.

G. A. Miller. WordNet: A lexical database for English. *Communications of the ACM*, 38(11):39–41, 1995.

N. Araslanov and S. Roth. Single-stage semantic segmentation from image labels. In *CVPR*, 2020.

A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollár, and R. Girshick. Segment anything. In *ICCV*, 2023.

N. Ravi, V. Gabeur, Y.-T. Hu, R. Hu, C. Ryali, T. Ma, H. Khedr, R. Rädle, C. Rolland, L. Gustafson, et al. SAM 2: Segment anything in images and videos. *arXiv:2408.00714*, 2024.

M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, 2021.

H. Xu, S. Xie, X. Tan, P.-Y. Huang, R. Howes, V. Sharma, S.-W. Li, G. Ghosh, L. Zettlemoyer, and C. Feichtenhofer. Demystifying CLIP data. In *ICLR*, 2024.

M. Minderer, A. Gritsenko, A. Stone, M. Neumann, D. Weissenborn, A. Dosovitskiy, A. Mahendran, A. Arnab, M. Dehghani, Z. Shen, et al. Simple open-vocabulary object detection with vision transformers. In *ECCV*, 2022.

M. Minderer, A. Gritsenko, and N. Houlsby. Scaling open-vocabulary object detection. In *NeurIPS*, 2023.

S. Liu, Z. Zeng, T. Ren, F. Li, H. Zhang, J. Yang, Q. Jiang, C. Li, J. Yang, H. Su, J. Zhu, and L. Zhang. Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection. In *ECCV*, 2024.

T. Ren, S. Liu, A. Zeng, J. Lin, K. Li, H. Cao, J. Chen, X. Huang, Y. Chen, F. Yan, et al. Grounded SAM: Assembling open-world models for diverse visual tasks. *arXiv:2401.14159*, 2024.

J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, 2019.

K. Zhou, J. Yang, C. C. Loy, and Z. Liu. Learning to prompt for vision-language models. *IJCV*, 130(9):2337–2348, 2022.

B. Li, K. Q. Weinberger, S. Belongie, V. Koltun, and R. Ranftl. Language-driven semantic segmentation. In *ICLR*, 2022.

G. Ghiasi, X. Gu, Y. Cui, and T.-Y. Lin. Scaling open-vocabulary image segmentation with image-level labels. In *ECCV*, 2022.

J. Xu, S. De Mello, S. Liu, W. Byeon, T. Breuel, J. Kautz, and X. Wang. GroupViT: Semantic segmentation emerges from text supervision. In *CVPR*, 2022.

F. Liang, B. Wu, X. Dai, K. Li, Y. Zhao, H. Zhang, P. Zhang, P. Vajda, and D. Marculescu. Open-vocabulary semantic segmentation with mask-adapted CLIP. In *CVPR*, 2023.

M. Xu, Z. Zhang, F. Wei, H. Hu, and X. Bai. Side adapter network for open-vocabulary semantic segmentation. In *CVPR*, 2023.

S. Cho, H. Shin, S. Hong, A. Arnab, P. H. Seo, and S. Kim. CAT-Seg: Cost aggregation for open-vocabulary semantic segmentation. In *CVPR*, 2024.

G. Shin, W. Xie, and S. Albanie. ReCo: Retrieve and co-segment for zero-shot transfer. In *NeurIPS*, 2022.

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

B. Recht, R. Roelofs, L. Schmidt, and V. Shankar. Do ImageNet classifiers generalize to ImageNet? In *ICML*, 2019.

L. Beyer, O. J. Hénaff, A. Kolesnikov, X. Zhai, and A. van den Oord. Are we done with ImageNet? *arXiv:2006.07159*, 2020.

Table 6: VABS on VOC-21 (test 300 images, GT-present mIoU). Automatic negatives recover most of the engineered-vocabulary gap and beat matched random negatives by $+3.5$ macro-average, without harming the official vocabulary. Right block: transfer to a COCO-Object-style protocol (80 things + background, stuff pixels remapped to background; the same fixed 300-image subset for every arm).

	VOC-21					COCO-Object		
Method	plain	+rand. 64	+VABS	official	off.+VABS	plain	+rand. 64	+VABS
SCLIP	34.8	50.3	53.6	55.4	55.7	22.6	30.6	31.8
ClearCLIP	34.9	47.8	51.4	53.6	54.2	22.4	32.3	33.0
MaskCLIP	31.8	40.6	**44.5**	43.6	43.8	19.8	23.9	24.8
NACLIP	36.5	49.5	52.7	55.0	55.1	22.1	32.2	32.8

A Additional Repair and Mechanism Evidence

A.1 Full VABS Results

Macro gains are cheap and robust. Table 6: automatic negatives lift plain vocabularies by $+12.7$ to $+18.8$ mIoU (macro $+16.1$), reaching $\approx$96–102% of the official engineered vocabulary itself (on MaskCLIP, exceeding it); the official vocabulary is never harmed ($+0.1$ to $+0.6$). These gains are *not* the metric artifact of §4.3: there, injected classes compete with ground-truth classes and their predictions are errors under the all-class convention, whereas VABS folds every negative prediction back into the background class, so the improvement is realised background IoU under the unchanged 21-class metric. The effect partially transfers: on a COCO-Object-style protocol all four methods gain $+5.0$ to $+10.7$ (macro $+8.9$), but on PASCAL-Context-60, where the 59 classes already cover most stuff, the mechanism has no room ($+0.5$/$+0.7$ for SCLIP/NACLIP, indistinguishable from random negatives). One disclosure bounds the COCO-Object transfer: the candidate lexicon contains COCO-Stuff names, which that protocol remaps to background by construction; a leak-free ADE-only lexicon still gains $+8.0$/$+9.8$ (SCLIP/NACLIP) but no longer beats random words. Most of the gain is thus a generic *background-sink* effect: matched random negative words already recover $+12.6$ macro-average on VOC-21, and the selection advantage over random ($+3.5$ on VOC-21) shrinks to $+0.6$–$+1.2$ on COCO-Object. Figure 4 summarises the VOC-21 arms; NACLIP is included via the replication protocol of §B.1. VABS was also benchmarked against §4.3's one-parameter control, as that section demands of any absorption claim: a background-logit boost tuned on the labelled dev split reproduces 73–85% of VABS's text-side gain on VOC-21 (e.g. SCLIP 50.7 vs 53.5, matched folding, heldout-200), but the tuned scalar is a dataset-specific calibration constant — transplanted to Context-60 or ADE-150 it harms every cell (down to -9.6), and on protocols without a background channel it has nothing to attach to, whereas VABS transfers without labels or tuning. The frozen protocol and verdict are archived as `preregs/prereg_w13_l5_vabsboost.md` and the W13-L5/L5t entries of `results/RESULTS_stage12.md`.

Two further pre-registered failures. Two later probes sharpen the dissociation. (i) A frozen text-only vulnerability predictor (z-scored nearest-neighbour margin plus semantic drift to the plain names) was tested for ranking *vocabularies* against the archived benchmark mIoU values: median per-method Spearman is 0.03 on the full VOC suite (kill threshold 0.5) — the same class of geometry signal that ranks surgery *configurations* at a fixed vocabulary with $\rho = 0.89$ (Appendix F) carries no information across vocabularies at a fixed configuration. (ii) A transductive prediction-space repair (RECAL: name-conditioned additive logit debiasing toward a mass-flattening prior, estimated label-free from the evaluation stream, frozen before the run) reduces mIoU in every cell (SCLIP plain $34.8 \rightarrow 27.6$, syn100 $30.6 \rightarrow 23.3$; NACLIP analogous): predicted class-mass imbalance under a perturbed vocabulary is dominated, in this setup, by legitimate scene statistics rather than name-conditional bias, so re-balancing it injects more error than it removes (verdict scoped to the single frozen prior strength α=0.5; no post-verdict sweep was run).

Per-class safety is not. A pre-registered per-class criterion (no target class may lose >3 IoU on ≥ 2 methods) failed on every variant we registered. The text-side filter cannot block negatives that are textually distant but visually overlapping: clothing words steal person pixels (-4.8 to -8.2

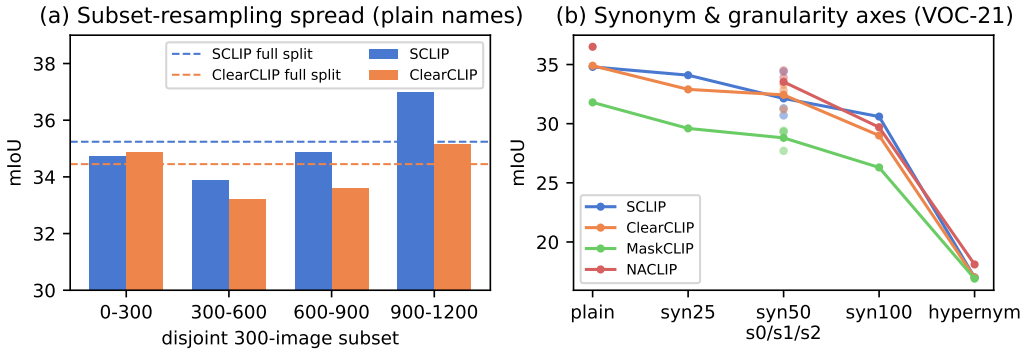


Figure 2: (a) Plain-name mIoU on four disjoint 300-image VOC-21 subsets vs. the full-split value (dashed): subset resampling alone spans up to 3.1 mIoU, our noise floor for method comparisons. (b) Synonym and granularity axes on VOC-21 for all four methods; dots show the three 50%-substitution seeds evaluated on the identical image subset.

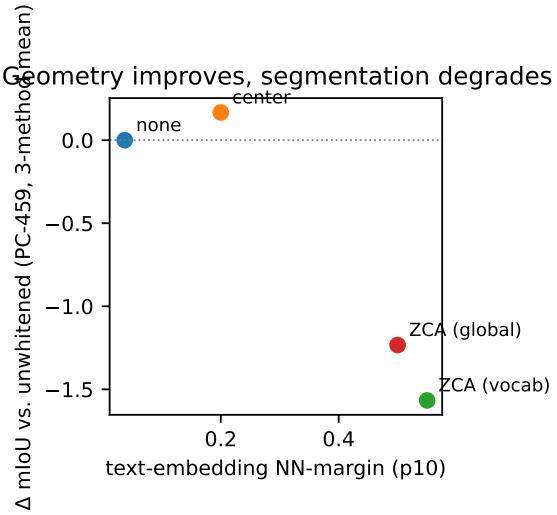


Figure 3: Text-side repairs on PC-459: each repair improves the embedding-space nearest-neighbour margin (x-axis) yet changes segmentation mIoU (y-axis, three-method mean) by at most $+0.2$ (centering) and down to -1.6 (ZCA, vocabulary-conditioned; largest single cell -2.1, ClearCLIP)—embedding geometry is not a proxy for dense segmentation quality.

IoU on the three main-matrix methods, and up to -15.4 on an LPOSS-style base in the archived extension runs) and "screen" steals tvmonitor (-5.6 to -14.7 on four; near-destroyed at 2.1 IoU on the LPOSS-style base)—precisely the classes the official vocabulary protects with *positive*-side expansions ("person in shirt, . . . "). Two label-free visual repairs also failed their pre-registered criteria. A steal-rate guard (drop negatives claiming high-confidence target patches on unlabeled calibration images) is blind at high confidence—dense CLIP probabilities are so diffuse that *zero* person patches exceed a 0.5 softmax—and at low confidence the safe operating point removes so many candidates that the macro gain collapses to $+2.5$. Reassigning high-steal negatives as sub-queries of the stolen class (automating the official vocabulary's positive expansion) is worse: without labels, the targets-only argmax on background pixels is itself the pathology being repaired, so the diagnostic attaches "sky" to aeroplane and "water" to bottle (-28 IoU).

Why this is an audit finding. The gap between engineered and plain vocabularies thus has two parts: a large generic component (any background sink, automatable for free when the vocabulary under-covers the scene) and a per-class protective component that, in our attempts, required exactly

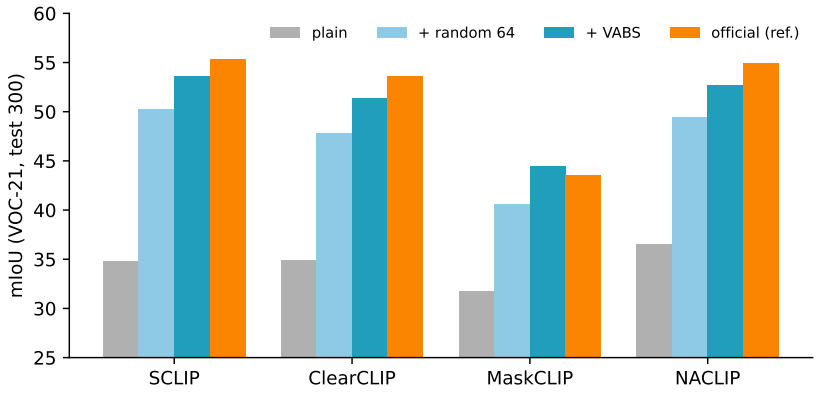


Figure 4: Automated background synthesis on VOC-21 (test 300 images). Synthesised negatives (VABS) close most of the plain-to-official gap for all four methods (official-vocabulary bars shown as reference); a matched random-word control already recovers the bulk of it, isolating the generic background-sink component.

the supervision signal (which pixels truly belong to class c) that a training-free segmenter cannot provide about itself. Engineered vocabularies appear to encode label-derived knowledge that our text-side and CLIP-confidence-based attempts do not recover; mechanisms with independent visual evidence (e.g. diffusion-derived negative prototypes (Karazija et al., 2024)) remain untested here, but within the text-only regime, automating vocabulary engineering is not a hyper-parameter search.

A.2 Learned adaptation replays the failure on both sides

A final pre-registered attempt asked whether the missing robustness can be *learned* on the text side. We trained a lightweight residual adapter over the frozen text embeddings (vision tower untouched) plus a set of learnable background queries initialised from VABS negatives, on 8,000 unlabelled ADE20K training images (disjoint from every evaluation set), using region-pooled pseudo-labels from the strongest training-free configuration as teacher, a naming-invariance consistency loss over a WordNet variant pool (variants appearing in any evaluation vocabulary were excluded from training), an anchoring regulariser, and an anti-collapse margin term. Losses converged and the embedding effective rank stayed within 2% of frozen, yet every pre-registered criterion failed. (i) Learned *universal* background queries recover at most half of the vocabulary-adaptive VABS gain, for a structural reason: the best negatives for one vocabulary (sky, grass, water for VOC-21) are target classes of the training vocabulary, so no fixed query set can contain them. (ii) On held-out synonym vocabularies the invariance-trained adapter is *worse* than the frozen encoder (23.9 vs. 26.9 mIoU), and a direct variant→canonical normalisation objective is worse still (20.7); both variants also degrade the engineered-vocabulary regime (52.8→46.5–48.5 over VABS). Together with §4.4's linear repairs and §4.5's selection filters, three text-side mechanism families—linear transforms, learned adapters, and learned negative queries—now fail in the same direction: objectives that improve text-space geometry do not transfer to, and often damage, dense segmentation. A final pre-registered probe moved the same consistency objective to the *visual* side—a residual convolutional adapter over frozen dense patch features, trained on the same unlabelled images so that patch-level class distributions become invariant to synonym re-samplings of the vocabulary. Training was unstable (two runs hit a pre-registered feature-drift abort), and the last in-budget checkpoint degraded every regime: plain 36.5→30.2, held-out synonyms 29.7→25.3 (the same signature as the text-side failure), and official -0.9. With this, the naming-invariance objective has now failed on both sides of the embedding space, and five mechanism families—linear transforms, selection filters, learned text adapters, learned negative queries, and learned visual adapters—fail in the same direction. Two further pre-registered probes close the remaining repair spaces: a graded corpus-frequency predictor of per-name damage fails (median Spearman -0.22 over 757 per-name observations across nine models — within the synonym pool, rarer names are *not* more damaging, and the reversal is robust to swapping the wordfreq estimator for a CLIP-BPE merge-rank proxy, so the effect of §B.4

is binary non-canonical-name sensitivity rather than a frequency law), and a training-free discrete word-space repair (rewriting each incoming name to its highest-frequency WordNet alias under a CLIP-cosine guard) recovers 35–44% of non-canonical-control and searched (ANS) damage but violates its frozen no-harm criterion, costing 3–4 mIoU on already-canonical vocabularies: without a reliable canonicality signal — which neither text geometry nor corpus frequency provides — the rewriter cannot tell which names to leave alone. A third signal family, an LLM world-knowledge canonicality judge (frozen concept-blind prompt), passes no-harm but *worsens* every damage arm (-0.6 to -4.5 mIoU; -3.3 to -4.5 on the synonym and search arms): without the intended visual concept it resolves each name to its dominant word sense and rewrites legitimate synonyms across senses (verdict scoped to this judge configuration: a single LLM with a single frozen prompt, judgments archived before any run). Word-space repair thus fails under the three signal families we pre-registered; the missing ingredient is the intended concept, which a vocabulary-only repair by definition does not have. A pre-registered follow-up asked the prediction question directly: over 47 substituted COCO-Object classes (per-class IoU deltas, three hosts), neither the original-to-synonym cosine (rho $+0.32$ to $+0.38$, below the frozen $+0.4$ bar) nor the induced change in maximum inter-class cosine (rho ≈ 0) predicts per-class synonym damage — the text-space analogue of visual-side inter-class-correlation accounts does not carry first-order per-class signal, so which classes break under renaming cannot be read off first-order text geometry. Pre-registrations, checkpoints and run records are released with the artifact.

B Extended Replications and Stress Tests

Unless a seed sweep is stated explicitly, the extension cells below are deterministic inference evaluations of one frozen checkpoint, image subset, and vocabulary. The three synonym "seeds" are frozen vocabulary draws rather than model-training seeds; learned-adaptation results are single training runs and are not presented with multi-seed uncertainty. This convention also covers the author-code, detector, and cross-lingual checks.

B.1 Replication on a fourth method: NACLIP

As a generalization check beyond the initial trio, we re-ran the key audit cells on NACLIP (Hajimiri et al., 2025), which augments k–k attention with a Gaussian neighborhood prior (our re-implementation under the identical frozen protocol, without its extra refinements; same 300-image subsets as the main matrix). Every effect replicates: naming, official 55.0 vs. plain 36.5 ($\Delta 18.5$); synonym, 50% seeds 33.9/34.5/32.2 and 100% 29.7; granularity, VOC-21 36.5$\rightarrow$18.1 and ADE-150 15.4$\rightarrow$4.1. The full-split V1.1 distractor confirmation supersedes the earlier single-draw near+200 distractor cell: near+200 reaches target/all-class mIoU 44.88/4.27 (plain 36.93), while the legacy official+200 control drops to 52.3 (-2.7). Repairs on PC-459, ZCA 14.3$\rightarrow$12.8 and centering 14.3$\rightarrow$14.5 (a second small positive centering cell, matching the SCLIP exception). The audit's conclusions are thus not specific to the three methods of the main matrix, though all four remain final-block variants over the same text encoder.

B.2 Scale generalization: ViT-L/14

A pre-registered replication on OpenCLIP ViT-L/14 (OpenAI weights, identical protocol, VOC-21 test-300, SCLIP and NACLIP) shows that every audited effect *direction* transfers to the larger backbone, while magnitudes become method-heterogeneous. Synonym substitution at 100% costs 7.0 (SCLIP, 37.2$\rightarrow$30.2) and 5.7 (NACLIP, 36.3$\rightarrow$30.7); near-distractor injection leaves GT-present mIoU flat or rising (SCLIP -1.1, NACLIP $+5.6$) while all-class mIoU collapses to 3–4, reproducing the background-sink/metric decomposition. The naming effect, however, splits: NACLIP retains a large official-vs-plain gap ($+14.3$) whereas SCLIP's shrinks to $+3.5$ — on ViT-L the correlative-attention variant extracts far less from vocabulary engineering. Dense quality overall is lower than ViT-B for all cells, consistent with prior reports on dense CLIP at larger scales. We therefore state the audit's conclusions as direction-stable but magnitude-dependent on backbone–method pairing.

Table 7: Robust-mIoU benchmark (VOC-21, full dev-excluded split, 1349 images). The first five numeric columns report GT-present mIoU. *robust* is the mean and *worst* the minimum over {plain, syn50×3 seeds, syn100}; NEG = official−plain. The last column reports all-class mIoU under 200 near distractors, with target-class mIoU in parentheses. The SCLIP, NACLIP, and ClearCLIP distractor entries are V1.1 means over 10 draws; other distractor entries are the earlier deterministic draw. All rows are our-protocol (re)implementations. Separate author-code anchor experiments (§B.3; ProxyCLIP, LPOSS, SC-CLIP, and Trident — not rows of this table) show smaller NEG under the ProxyCLIP and LPOSS pipelines but a matching NEG for SC-CLIP and Trident, so NEG *sizes* here are protocol-conditional. NEG is computed from unrounded run records and may differ from the difference of the displayed official and plain columns by up to 0.1.

Method	official	plain	robust	worst	NEG	distract.
MaskCLIP	44.6	32.8	29.7	26.8	11.8	3.5 (37.1)
SCLIP	57.2	35.6	33.5	31.8	21.6	4.4 (46.5)
ClearCLIP	55.7	34.8	32.5	29.8	20.9	4.2 (44.0)
NACLIP	56.9	36.9	34.0	30.7	20.0	4.3 (44.9)
ProxyCLIP-style	60.0	38.0	35.0	31.7	22.1	4.3 (44.8)
LPOSS-style	**60.5**	**40.8**	**37.3**	**32.7**	19.6	4.3 (45.5)
SC-CLIP-style	58.4	38.3	35.2	31.6	20.1	4.2 (44.1)

B.3 Generational generalization: 2024–25 visually-guided methods

A pre-registered extension asks whether the newest training-free generation — methods that inject an external self-supervised visual prior rather than editing CLIP's attention — has implicitly fixed naming sensitivity. We add three style-reimplementations to the unified protocol (disclosed as such, not reproductions): ProxyCLIP-style DINO (Caron et al., 2021) proxy attention, LPOSS-style DINO-affinity label propagation, and SC-CLIP-style anomaly-token restoration. On VOC-21 (full dev-excluded split, 1349 images) their official-vs-plain gaps are $+22.1$, $+19.6$ and $+20.1$ respectively — as large as the attention-surgery family — and all three collapse to all-class mIoU $\approx$4 under 200 near-distractors while GT-present mIoU stays at 44–46, reproducing the background-sink/metric decomposition. Vocabulary sensitivity is therefore a property of the shared CLIP text-matching interface, not of any particular dense-feature mechanism. One robustness-relevant difference does emerge: label propagation is the most vocabulary-robust mechanism on every perturbed aggregate (plain, mean-robust and worst-case); two pre-registered mechanism probes show this is a higher-plain-baseline effect rather than noise filtering — naming noise is graph-low-frequency, and transplanting the propagation operator onto the other six methods raises absolute mIoU but does not reduce the synonym drop. A cautionary methodological finding accompanies this: on a 300-image subset the *official*-vocabulary top spot went to proxy attention, but on the full dev-excluded split (1349 images) label propagation tops both rankings — the identity of the official-vocabulary leader is subset-sensitive at sub-0.5 mIoU margins, whereas the robust-aggregate ranking was stable across both evaluations, which itself argues for reporting perturbed aggregates.

Searched worst case (ANS). The random-synonym suite understates worst-case fragility. A pre-registered greedy search over per-class naming choices (WordNet candidates filtered to CLIP cosine $[0.70, 0.95]$; one coordinate pass; search and evaluation subsets disjoint) finds *lexicon-legal* vocabularies at held-out mIoU 13.4 (ClearCLIP) and 13.7 (LPOSS-style) — 15–17 mIoU *below* the worst random-suite member (28.3 / 31.0; random-suite figures are test-300 values, ANS figures heldout-200 — disjoint subsets, both disjoint from the search set) and with no search–held-out gap. The search is run independently per method, but a transfer matrix shows the found fragility is shared, not search adaptation: each method's searched vocabulary is nearly as damaging on the other method (cross-transfer 13.3 / 15.2 vs own 13.4 / 13.7). The bound is also robust to the search procedure itself: a pre-registered control re-runs the greedy pass on SCLIP under two random coordinate-visit orders alongside the alphabetical one (all three SCLIP searches run fresh for this control; the original searches used ClearCLIP and LPOSS-style); the three searches agree on 16/21 class choices and land within 1.4 held-out mIoU of each other (13.2–14.5 vs plain 34.7), so the reported worst case is not an artefact of the visit order. The most vocabulary-robust method in Table 7 (the LPOSS-style reimplementation) collapses to the same level as the least robust one: the label-propagation advantage vanishes under search. The full searched vocabularies, with a per-class plausibility an-

notation, are listed in Appendix E. We frame this as an upper-bound stress instrument rather than a typical-user scenario — the selected names are dictionary-sanctioned but often low-frequency senses (bicycle→wheel, person→soul, dog→frump) — but it shows that random perturbation suites measure average-case, not worst-case, naming robustness, and the searched axis is therefore reported separately in the benchmark. A pre-registered control decomposes what the search finds: replacing each ANS name with a *different* synonym from the same cosine pool, matched on CLIP token count (3 seeds), already costs 15.3 of the 21.0-point ANS drop on ClearCLIP — roughly $2\times$ the random-suite drop. Most of the searched damage is therefore *non-canonical-name sensitivity*: any dictionary synonym that departs from the canonical class name hurts (random suites under-sample such names), with a genuine but modest residual adversarial component of roughly 5–7 points (control seeds span 18.5–19.7) on top. We deliberately do not phrase this as a graded frequency law: a pre-registered follow-up over 757 per-name damage observations across nine models found the correlation between corpus rarity and damage to be absent and slightly sign-reversed (median Spearman -0.22 with a wordfreq Zipf estimator, -0.12 with a CLIP-BPE merge-rank proxy closer to the training corpus), so the effect is binary — canonical vs. not — rather than monotone in frequency; the matched control above remains valid under this reading (matching on the candidate pool and token count matches non-canonicality, whether or not frequency grades it). The practical lesson for suite reuse: random synonym suites under-represent non-canonical names and therefore understate the tail by 2–3$\times$; suites should either report a non-canonical stratum explicitly or, since no text-side signal we tested (geometry, frequency) can rank names, accompany average-case suites with an ANS-style search bound.

External anchor on unmodified author code. To rule out reimplementation artefacts, we ran the *unmodified* official ProxyCLIP release under its own protocol (mmseg pipeline, 2048×336 sliding inference, the authors' background thresholding, full VOC val), changing only the class-name file. The authors' name file scores 61.2 mIoU (matching the published number and within 1.2 of our style-reimplementation's official row); plain names score 47.3 (NEG $+14.0$); syn100 names score 41.2 (synonym drop 6.1). Naming engineering and synonym fragility therefore replicate on author code without any of our infrastructure — the smaller NEG relative to our protocol ($+22.1$) is partly explained by their background threshold absorbing part of the plain-vocabulary penalty, itself an instance of the background-modeling coupling documented in Section 4.3. A second author-code anchor, the unmodified official LPOSS release (Stojnić et al., 2025) (its own DINO-affinity label-propagation pipeline, 2048×448, full VOC val, again changing only the class-name list): authors' names 61.1 (published 61.2; our style-reimplementation's official row 60.5), plain 55.1 (NEG $+6.0$), syn100 47.7 (synonym drop 7.4). Direction replicates on both axes — and the synonym drop is *larger* on author code than our style-reimplementation's 4.3 — but the author-code NEG is far below our-protocol's 19.6: label propagation plus their pipeline absorbs much of the plain-vocabulary penalty, so leaderboard statements that depend on the *size* of the LPOSS naming-engineering effect are anchored to $+6.0$, not 19.6. A third anchor, the unmodified official SC-CLIP release (Bai et al., 2024) (its mmseg pipeline, 2048×336, full VOC val, again changing only the class-name file): authors' names 64.6 (matching the published 64.6 exactly), plain 43.0 (NEG $+21.6$), syn100 38.0 (synonym drop 5.0) — here the author-code NEG essentially *matches* our-protocol's $+20.1$ for the SC-CLIP-style row, because the official SC-CLIP name file carries the heaviest naming engineering in our set (a 26-item background expansion plus multi-alias foreground entries) and its pipeline, unlike ProxyCLIP's thresholding or LPOSS's propagation, has no mechanism absorbing the plain-name penalty. A fourth anchor extends the evidence to the SAM-coupled family, which none of our re-implementations cover: the unmodified official Trident (Shi et al., 2025) release (CLIP+DINO feature splicing with SAM (Kirillov et al., 2023) encoder aggregation and SAM refinement, its own protocol, full VOC val, again changing only the class-name file): authors' names 67.1 (matching the published 67.1), plain 47.4 (NEG $+19.7$), syn100 41.7 (synonym drop 5.8) — both axes replicate on a method whose visual pathway is qualitatively different from the dense-CLIP families, at a NEG size comparable to our-protocol values. Trident's pipeline includes SAM-based refinement, yet its NEG is undiminished: spatial refinement corrects boundaries, not naming — visual add-ons do not by themselves confer vocabulary robustness. The worst-case (ANS) axis also replicates on author code: the frozen search protocol run end-to-end on the official Trident release (query features recomputed exactly as its own initialization; disjoint frozen search/held-out subsets) drives held-out mIoU to 18.2 against plain 47.8 and syn100 40.7 — 22.5 below the random-synonym level, with no search–held-out shrinkage (search-100 was 20.4), so the searched fragility tail replicates end-to-end in an unmodified official pipeline: the worst-case regime is not specific to our protocol

(demonstrated on one official codebase; single alphabetical search — the search-order control was not re-run on author code). The Trident ANS level (18.2) and our in-protocol ANS levels (13.4/13.7) come from different pipelines and subsets and are not directly comparable. The LPOSS ANS figures remain our-protocol-only. A fifth anchor adds venue freshness and a 2026 method: the official RF-CLIP (Li et al., 2026) release (AAAI 2026, attention-redistribution family, CLIP-only ViT-B/16), changing only the class-name files (plus one memory workaround: the release's one-hot alias post-processing exhausts 24 GB with 145 alias queries, so we substituted a mathematically equivalent per-class max, regression-verified to reproduce the official VOC score bit-identically), reproduces its published scores on VOC (64.7 vs 64.8) and COCO-Object (36.8 vs 37.9, within tolerance) and shows the largest naming-engineering gap we measure on any author code: VOC NEG $+23.8$ (plain 41.0), with syn100 at 36.8 (drop 4.1); on COCO-Object NEG $+4.9$ and a synonym drop of 6.0. On VOC, the newest attention-surgery method is therefore the *most* dependent on engineered names among our anchors, not the least — 2026-era gains do not come with naming robustness. (On Context-60 the RF-CLIP official arm scored 1.57 above the published value, just outside our frozen 1.5 reproduction gate — likely our converted ground truth, as for Trident — so per protocol we draw no conclusions from that dataset for this method.) A sixth anchor covers the mask-generator family and the current training-free performance ceiling: the official CorrCLIP (Zhang et al., 2025) release (ICCV 2025 oral; MetaCLIP (Xu et al., 2024) + DINO + the authors' pre-generated SAM2 (Ravi et al., 2024) region masks) reproduces its published scores to within 0.1 (VOC 74.79 vs 74.8, COCO-Object 43.67 vs 43.7) and shows NEG $+20.5$ on VOC (plain 54.3) against $+1.4$ on COCO-Object, with synonym drops running opposite (5.8 vs 9.8) — the NEG/synonym dissociation observed on Trident replicates on a second SAM-family method. Because the region proposals are pre-computed and never see class names, every delta on this anchor is attributable to the text-side interface by construction. (Its Context-60 official arm lands 1.51 above the published value, marginally outside the same frozen gate — again likely our converted ground truth — so per protocol we draw no conclusions from that dataset for this method either; the authors' pre-generated Context masks cover 5104 of the 5105 validation images, so all Context arms for this anchor use the matching 5104-image list.) A seventh anchor covers the retrieval family that ReME opened and that our re-implementations do not reach: the official SCI-CLIP release (Zamini and Shukla, 2026) (training-free retrieval over a reference memory of region-pooled segment embeddings; in the released configuration we run, the bank is built from each benchmark's training split and regions come from SAM2 masks — a code-level description of the official pipeline as shipped, not of every configuration the paper reports). Its published CLIP ViT-B/16 scores reproduce closely on three of our four datasets — VOC 75.91 vs 75.9, COCO-Object 45.20 vs 45.2 (exact), ADE-150 29.83 vs 29.8 — while its Context-60 arm lands 1.56 above the published value, marginally outside the frozen 1.5 gate (again consistent with our converted Context ground truth), so per protocol we draw no conclusions from that dataset for this method. Changing only the class-name file (the frozen reference bank is reused across arms; retrieval is keyed on class indices, though the bank's stored label features derive from the official names — a residual coupling that favours the official arm and may inflate the perturbed-arm penalties, so the deltas on this anchor read as upper bounds): VOC plain 57.54 (NEG $+18.4$), syn100 52.23 (drop 5.3); COCO-Object plain 41.75 (NEG $+3.5$), syn100 32.37 (drop 9.4); ADE-150 plain 29.49 (NEG $+0.3$), syn100 15.29 (drop 14.2). This anchor thus reproduces every pattern in our set at once: a large VOC naming-engineering gap, the NEG/synonym dissociation across datasets ($+18.4/5.3$, $+3.5/9.4$, $+0.3/14.2$ — the same NEG-down/synonym-up ordering as Trident), and the largest synonym drop we measure on any author code (14.2 on ADE-150, an upper-bound reading given the bank coupling above): grounding class scores in retrieved *visual* references does not immunize the text-side interface, because the queries that address the bank are still CLIP name embeddings (one official implementation; family-level statements here are anchored to this single codebase). The spread between author-code and our-protocol NEG (ProxyCLIP $+14.0$ vs $+22.1$; LPOSS $+6.0$ vs 19.6; SC-CLIP $+21.6$ vs $+20.1$) shows that implementation and protocol details modulate effect *sizes* even where the direction replicates — and that, across the anchored methods, the modulation is method-dependent rather than a uniform reimplementation inflation.

A second round on Context-60 and COCO-Object (same suite minus official, which does not exist for these datasets *in our protocol*) confirms that synonym drops of 2–9 mIoU persist across all seven methods, though rank swaps there are minor — the label-propagation robustness advantage is largest on VOC. The SC-CLIP author-code anchor extends to COCO-Object, where the official release does ship an engineered name file: authors' names 37.7 (published 37.7, exact), plain 34.4 (NEG $+3.3$), syn100 27.9 (synonym drop 6.5, larger than its VOC drop of 5.0) — both axes repli-

cate on author code on a second dataset. The NEG size, however, is $6\times$ smaller than on VOC: naming-engineering gaps are strongly dataset-dependent, and VOC-scale NEG values must not be extrapolated to other benchmarks. The Trident anchor extends to COCO-Object with the same pattern: authors' names 41.1 (published 41.1, exact), plain 39.1 (NEG $+2.0$, $10\times$ smaller than Trident's VOC $+19.7$), syn100 29.0 (synonym drop 10.1, far *larger* than its VOC drop of 5.8). The two axes therefore decouple on author code: a benchmark can have little naming-engineering headroom yet high synonym fragility — background headroom and naming sensitivity are different quantities and must be reported separately. A third Trident dataset sharpens this: on Context-60 (our 60-class conversion of the full annotations; the official arm scores 40.1 vs published 38.6, a gap of exactly 1.5 — at, but not beyond, the reproduction gate frozen before any Context run as "within 1.5 mIoU", so this dataset is retained while the 1.51–1.57 Context gaps of the other anchors fall outside it; label-conversion differences may contribute; all five arms share the same converted ground truth, so the NEG and synonym deltas are internally consistent — the conversion can shift absolute levels only) the naming-engineering gap is essentially zero (plain 39.9, NEG $+0.2$) while the synonym drop is the largest we measure on author code (12.4, to 27.5). The two axes therefore dissociate across the three datasets (n=3, an observation rather than a law): ordered by NEG size ($+19.7$, $+2.0$, $+0.2$), the synonym drop runs opposite (5.8, 10.1, 12.4) — consistent with the two axes being driven by different mechanisms.

B.4 Cross-family transfer: grounding-supervised detectors

All methods above share CLIP's contrastive text tower, so a natural objection is that naming fragility is an artefact of that one interface. A pre-registered transfer applies the same three axes to a different training paradigm: box-supervised open-vocabulary detectors converted to segmenters by prompting SAM with their boxes (Grounded-SAM-style harness (Ren et al., 2024); detection scores arbitrate overlapping masks; unclaimed pixels are background). We audit Grounding DINO (Liu et al., 2024) (BERT (Devlin et al., 2019) based phrase grounding) and OWLv2 (Minderer et al., 2023) (CLIP-initialised text tower, detection-tuned) on VOC-21 and COCO-Object. Because the harness differs from the dense pipelines, all comparisons below are *within-detector deltas*; absolute mIoU values are not comparable across the two families. Three findings:

(1) Synonym fragility differs sharply within the detection family, and is not organised by text-encoder lineage. Grounding DINO collapses under legal synonyms: VOC 77.6→40.0 (-37.6; COCO-Object -28.3) — roughly $5\times$ the CLIP-family drop — whereas OWLv2 degrades like the CLIP-surgery family (VOC -7.1, COCO-Object -12.9). Both detectors saw VOC/COCO categories in supervised pretraining, so the plain rows are quasi-in-distribution and these drops measure how tightly masks are bound to category *surface forms*, not zero-shot failure (the in-distribution advantage is shared, so it does not confound the contrast). An obvious reading is a text-encoder-lineage effect (BERT phrase grounding vs CLIP contrastive tower), and one prediction of that reading does hold: a worst-case vocabulary searched in CLIP text space (below) transfers strongly to the CLIP-tower detector while adding nothing on the BERT-grounded one (OWLv2 -31.8; Grounding DINO lands at its already-collapsed random-synonym level, -38.9 vs -37.6). A pre-registered third-model check, however, falsifies lineage as the organising variable: OWL-ViT v1 (Minderer et al., 2022), which shares OWLv2's CLIP text tower, drops 19.0 under the same synonyms (64.0→45.0) — outside the frozen CLIP-tower band ($\leq$15) and nearly in the BERT-grounded band ($\geq$20). Since OWLv2 and OWL-ViT v1 differ mainly in training recipe (both share the CLIP tower; v2 adds self-training at scale), the contrast is consistent with a training-recipe effect rather than text-encoder pedigree; the Grounding DINO vs OWLv2 contrast stands as a model-level observation. A same-architecture natural experiment supports the recipe reading observationally: across the released MM-Grounding-DINO Swin-T ladder (identical architecture, training mixture varying from O365+GoldG to +GRIT to +GRIT+V3Det), the synonym drop moves by 10 mIoU (32.2 / 22.2 / 24.7) — the largest same-architecture spread we observe — with the web-grounding mixture the most robust; the pattern is not monotone in data volume in this three-point sweep (and each tier entangles composition with scale), so it demonstrates that the recipe matters without identifying which ingredient or licensing a data-scale story. A second attempted third model triggered the pre-registered infeasibility clause (plain GT-present < 35, frozen before any synonym result was seen) and is itself an interface finding: MDETR (RoBERTa grounding, no presence head) grounds *every* caption somewhere in every image, so per-class querying floods the image with boxes (3.4 mIoU

regardless of vocabulary) — grounding without a presence decision cannot support vocabulary-conditional segmentation in this harness.

(2) The searched worst case transfers across paradigms. The ANS vocabulary searched on ClearCLIP (§B.3) costs OWLv2 31.8 mIoU (73.4→41.6 held-out) — $4.5\times$ its random-synonym drop — despite OWLv2 never being touched by the search. Random suites therefore understate worst-case fragility across training paradigms, not just within the searched family. A pre-registered token-matched control bounds what transfers: matched non-canonical synonyms drawn from the same candidate pool cost OWLv2 25.3 of the 31.8-point ANS drop (80%; 3 seeds, 44.1–50.3 vs plain 73.4). The cross-paradigm transfer is therefore mostly shared *non-canonical-name sensitivity* — itself a cross-family phenomenon random suites underestimate by $\sim 3\times$ — with a modest search-specific increment ($\sim$6.5 mIoU).

(3) No foreground absorption in the detection family. Under 200 injected near-distractors, GT-present mIoU stays essentially flat (Grounding DINO -4.3, OWLv2 -1.3, versus a $+7.95$ to $+10.93$ rise for the V1.1 CLIP-family trio; the legacy seven-method single-draw range was $+4.6$ to $+9.8$). This is not mere dilution: the detectors do fire on the injected names (17.8% / 23.6% of pixels are claimed by distractor classes), but $\sim$92% of those pixels lie on ground-truth background, and only 4.5% / 8.5% of foreground mass is stolen (V1.1 CLIP-family near+200: 28.99–35.42%). The explicit per-query presence decision produces genuine open-set false positives on background regions while protecting foreground pixels — corroborating the two-ledger decomposition of Appendix G from a second architecture family, where the all-class collapse ($\approx$7) is again the metric-convention term.

Scope: a box→SAM harness rather than native segmentation (synonym drops could partly reflect boxes falling below the fixed detection threshold rather than degraded matching; a threshold sweep was not run); one seed; test-300/held-out-200 subsets vs the dev-excluded full split of Table 7; Grounding DINO's 256-token text limit requires 40-class query chunks on the distractor vocabulary, OWLv2 queries classes independently.

A pre-registered native-metric check removes the harness confound entirely: evaluating OWLv2 and YOLO-World as *detectors* (official COCO `pycocotools` mAP, val2017 first 1000 images, only the class-name strings changed) reproduces both signatures. Full-synonym vocabularies cost $\sim$30% of mAP on both models (OWLv2 48.4→36.2/31.1 under two frozen synonym vocabularies; YOLO-World-L 46.8→35.3/30.0), so synonym fragility is a property of the vision–language interface, not of our box→SAM conversion or of mIoU conventions. Injecting 200 near-distractor queries, by contrast, barely moves native AP (-1.2 / -2.6 mAP): per-query decoding has no shared softmax or background ledger to absorb. The mechanically small all-class mIoU therefore has no detection analogue, while the V1.1 pixel ledger shows that segmentation also suffers real foreground and background reassignment under shared competition. The contrast is an interface effect, not evidence that the segmentation failure is purely a metric convention or purely lost recognition. A decomposition at IoU 0.5 attributes the synonym drop chiefly to outright recall loss (GT recall 78.5%→58.1%; 21.5% of objects are no longer detected at any score) rather than score reordering. Two further pre-registered detection-side couplings came back *null* and we report them as such: letting original names and synonyms coexist as 160 queries neither degrades mAP (-0.7) nor produces duplicate boxes (0.2% at IoU$\geq$0.9), and a cross-name NMS merge changes nothing (-0.2) — the per-query interface is robust to vocabulary redundancy, which is precisely why these findings live here as an audit extension rather than as a separate detection study (frozen go/no-go criteria: the phenomenon-existence condition passed at $2\times$ margin, the detection-specific-coupling condition failed both thresholds).

B.5 A CROSS-LINGUAL PROBE: A SPANISH CASE STUDY

Real vocabularies are not always English. A pre-registered quick test (fixed dictionary translations of the VOC-21 names, frozen before any run) found Spanish structured and Chinese a floor collapse, so the Spanish axis was promoted to the full matrix (VOC test-300, 9 models); we scope this as a single-language probe rather than a general cross-lingual axis — broader typology is untested. Translations are fixed common dictionary renderings frozen before any run (provenance in the released vocabulary files). The profile is a family split, not a method ranking: all seven dense CLIP-family methods — across three method generations — retain a tightly clustered 52–62% of their plain mIoU (16.5–

23.4 absolute), OWLv2 retains 79% (72.5→57.3; test-300 baseline), and Grounding DINO retains 18% (77.6→14.1), under default translations; the per-class best-of-3 oracle raises Grounding DINO only to 25% retention (19.7 vs plain 77.6), so the family split is not a translation-choice artifact. The axis decouples from the synonym axis in an informative way: the dense family's synonym-robustness spread collapses to a uniform band under translation, because the cross-lingual response is set by the shared CLIP text tower's multilingual capability rather than by any dense-inference design choice; the detector contrast (OWLv2 most robust on both axes, Grounding DINO most fragile on both) survives (one model per detector family; the contrast is a model-level observation). Note this does not resurrect the lineage hypothesis falsified in §B.4: multilingual capability tracks the text tower almost by construction (CLIP towers saw multilingual alt-text; BERT-grounding saw English grounding data), whereas *synonym* robustness demonstrably does not track lineage — the two statements concern different mechanisms and are compatible. Chinese collapses everywhere (<10 mIoU; for OWLv2 the translated queries do not even fit the 16-token CLIP context and require truncation — the boundary is partly tokenizer representability, disclosed as such), so it is reported as a boundary note rather than an axis; pre-registered artifact controls (bare-name queries without the English template, and shortened 2-character forms) recover nothing ($<$ 1 mIoU gain, mostly negative), so the collapse is not explained by the two tested prompt-interface controls. For Spanish, by contrast, a pre-registered best-of-3 translation control (two frozen dictionary alternates per class, per-class oracle — an upper bound biased *against* this reading) recovers 86% of plain-English mIoU on SCLIP (30.0 vs 34.8; single translations 14.7–20.7), crossing the frozen translation-choice band, and 84% on OWLv2 (62.0 vs 73.4; heldout-200 baseline, hence the difference from the 72.5 test-300 figure above), same direction but just under the 85% line: the 52–62% single-translation figures largely measure sensitivity to *which* legal translation is chosen — the naming-choice fragility of §4.2 re-appearing in another language — rather than a Spanish capability deficit (the family split across translations survives). Chinese, per the artifact controls above, remains a genuine capability boundary. A pre-registered two-language extension (German, Russian; k=3 frozen translations each) finds the same structure at language-dependent magnitude: oracle-minus-default retention gaps of 14.0 (German) and 10.9 (Russian) points on SCLIP and 12.7 (German) on OWLv2, while Russian on OWLv2 collapses to a floor (5.9% retention) — a second non-Latin-script data point alongside the Chinese boundary (for Russian we ran no artifact controls analogous to H4's, so tokenizer-representability artifact and genuine incapacity are not separated there). In every language we tested (three languages, k=3 frozen translations — a lower bound on the oracle gap), translation choice is a substantial free variable, and single-translation cross-lingual scores should be read with an oracle-gap control; broader typology is future work.

C Extended Limitations and Calibration Details

Our audit covers three final-block CLIP variants under one backbone (with a two-method ViT-L/14 direction replication, §B.2); because they share the text encoder, templates, and pooling, their correlated response to vocabulary perturbation is partly by construction, and all method-general statements are scoped to this family (§B.1 extends it to NACLIP, a fourth member). The generational extension (§B.3) adds ProxyCLIP-, LPOSS- and SC-CLIP-style reimplementations plus author-code ProxyCLIP, LPOSS, and SC-CLIP anchors, and §B.4 transfers the protocol to two box-supervised grounding detectors through a box→SAM harness; SAM-based pipelines (e.g. Trident-class refiners) as *audit subjects* remain open (the protocol applies unchanged). The discrete word-space rewriter's partial success against searched and non-canonical attacks (35–44% recovery) indicates a repair direction that a reliable canonicality detector would unlock, though none of the signals we tested provides one. A-847 was unavailable at audit time; PC-459 serves as the large-vocabulary regime. Evaluations use fixed, released 300–500-image validation subsets per condition for tractability of the ∼150-run matrix. Two calibrations bound the subset error: on the full VOC-21 validation split (all 1449 images; the dev-excluded 1349-image variant used elsewhere differs by at most 0.4) the naming effect persists at full strength (plain 35.2/34.5/32.5 vs. official 56.9/55.4/44.4 for SCLIP/ClearCLIP/MaskCLIP, Δ 11.9–21.7), and re-evaluating the plain condition on four disjoint 300-image subsets gives a subset-resampling spread of up to 3.1 mIoU (SCLIP) and 1.9 (ClearCLIP). Effects of this order or smaller (including the 0.1-point VOC-21 plain-name difference between SCLIP and ClearCLIP) are treated as noise throughout. The frozen perturbation thresholds (synonym similarity window, distractor guard and strata boundaries) were fixed before segmentation runs but not themselves ablated; first-sense WordNet hypernym mapping can pro-

duce non-visual parents, so the granularity numbers are lower bounds under a fixed automatic rule rather than human-validated coarsenings. The remaining gap of our official-vocabulary reproduction to published numbers (full-split 56.9 vs. 59.1 for SCLIP) is consistent with our protocol omitting PAMR post-processing. A pre-registered check bounds this axis: adding PAMR refinement to SCLIP/NACLIP (VOC test-300, three vocabulary conditions) leaves the synonym drop unchanged (5.5 vs 5.4 mean) and *amplifies* the naming-engineering gap by 15% (19.6 $\rightarrow$ 22.4) — spatial post-processing helps the engineered vocabulary more than the plain one, so the audited magnitudes are not artifacts of omitting refinement (multi-scale and full published stacks remain unbounded). Scoop-check against in-review venues (OpenReview) was not possible; novelty claims are made against the published literature.

D Synthesis of the Extended Audit

Treating the inference vocabulary as a controlled experimental variable reveals that reported numbers for final-block training-free OVSS methods are strongly shaped by vocabulary artifacts: name engineering comparable to or larger than method differences, synonym noise larger than several method gaps in our matrix, a mIoU protocol whose response to vocabulary expansion is incoherent with pixel accuracy and can reverse the selected method, and text-side geometric repairs that help every diagnostic except actual segmentation. A constructive attempt to automate vocabulary engineering (§4.5) recovers most of its macro-mIoU value with synthesized background negatives, but its per-class protective component resisted every label-free mechanism we pre-registered, indicating that engineered vocabularies encode protection not reconstructed by the tested label-free machinery—linear, selected, or learned, on either the text or the visual side (§A.2). We release the protocol and harness so that vocabulary robustness can become a standard reported quantity for open-vocabulary segmentation.

E ANS Searched Vocabularies

Table 8 lists every class whose name the greedy search replaced, for both searched methods, with a coarse plausibility annotation: *plausible* = a name an ordinary user might type for the class; *legal-but-odd* = a WordNet-sanctioned synonym of a rarer sense that a user is unlikely to intend. Most of the damage comes from legal-but-odd choices, supporting the "upper-bound stress instrument" framing; a handful (e.g. *lounge* for sofa, *soul/mortal* for person) are plausible user inputs, so the searched axis is not purely adversarial.

Table 8: ANS replacements (VOC-21). Unlisted classes kept their plain name.

class	ClearCLIP choice	LPOSS-style choice	plausibility
bicycle	wheel	wheel	legal-but-odd
bird	skirt	skirt	legal-but-odd
boat	gravy holder	gravy holder	legal-but-odd
bottle	bottleful	nursing bottle	legal-but-odd
bus	charabanc	charabanc	legal-but-odd
car	railroad car	railroad car	legal-but-odd
cat	guy	guy	legal-but-odd
chair	chairwoman	chairwoman	legal-but-odd
dog	frump	frump	legal-but-odd
horse	sawhorse	sawhorse	legal-but-odd
person	soul	mortal	plausible
sofa	lounge	lounge	plausible
train	caravan	caravan	legal-but-odd

F Head-Level Decomposition of Final-Block Surgery

The audited methods differ only in the final attention block, inviting a mechanistic question: *where* inside that block does dense quality live? We decomposed the final-block output into per-head contributions (via the output projection's column blocks) for four attention flavours (vanilla q–k, q–q,

k–k, identity/value-only) and evaluated each single head and head subset under the unified protocol (VOC-21, 100 development images, official vocabulary). Two descriptive findings and one negative result. (i) The head spectrum is extremely uneven (single-head spans of 13.6–21.6 mIoU) and *flavour-invariant*: head rankings correlate at Spearman 0.92–0.96 across all four flavours, with heads $\{3, 9, 11\}$ carrying dense semantics in every case—surgery methods appear to amplify one fixed head structure rather than create new ones. (ii) A label-free alignment margin (mean top1$-$top2 text similarity over patches) predicts the segmentation quality of 20 surgery configurations (4 flavours $\times$ exit layers 8–12) at Spearman 0.89 without any annotation, though restricting the margin to the top-3 heads weakens it to 0.63, so the predictive signal lives at the whole-output level, not specifically in the head structure. (iii) Head *selection* is not a method: oracle subsets beat all-heads by only $+0.4$ to $+0.6$ on surgery flavours ($+2.8$ on vanilla), and label-free selection loses up to -2.6; the surgery has already harvested the fixable component. All 20 configurations and per-head records are in the artifact release; sample sizes (100 images, 20 configurations, one backbone) bound these as within-distribution observations.

Two pre-registered follow-up probes sharpen the picture with oracle upper bounds. (iv) Decomposing the similarity itself into per-head subspace terms $\sum_h w_h \langle P_h t, P_h v \rangle$ (with P_h the projector onto head h's output subspace, linearised through the final layer norm) does not help even with GT-fitted weights: the held-out gain is $+1.1$ mIoU on vanilla attention (below the $+2.8$ oracle head-selection bound) and -5.4 on q–q; a uniform-weight decomposition loses 2–6 mIoU against the plain inner product, indicating that cross-subspace interference terms carry signal that any per-head decomposition discards. (v) Routing *configurations by region* has a large oracle ceiling but no label-free access to it: choosing the best of the 20 configurations per SAM region with GT reaches 68.2 mIoU versus 53.2 for the best single configuration ($+14.9$; an image-level oracle gets only $+4.3$), yet every label-free gate we pre-registered (region margin, confidence, cross-configuration agreement over the margin-selected top-4 pool) lands *below* the best single configuration (38.2–46.7)—consistent with (ii): the margin signal predicts quality globally but not at region granularity. Both probes were killed by their pre-registered criteria and are reported as mechanism-level negative results with their oracle gaps. A follow-up asked whether a *small amount of supervision* unlocks the routing ceiling: a ridge router over region margin, confidence, entropy, size and cross-configuration agreement features, fit on 50 labelled images, transfers negatively to held-out images (53.4 vs. 54.0 for the best member) and degenerates on a disjoint 300-image split to always selecting the best single configuration ($+0.0$). The routing signal is not present in this family of region-level confidence statistics even with labels; accessing the oracle gap appears to require qualitatively different evidence. Two further pre-registered probes closed the remaining signal families we could test. (vi) *Cross-configuration consistency* routing (route each region to the configuration agreeing with the 8-pool majority vote, optionally gated by DINO boundary crispness) lands below the best single configuration on both VOC (46.8 vs. 54.2) and Context-60 (21.9 vs. 26.4), and a margin control performs equally — the third signal family (after margin and few-label supervision) that fails to access the routing oracle; an 8-configuration pool restricted to exit layers 11–12 has an oracle ceiling of only $+2.4$, showing most of the $+14.9$ headroom lives in early-exit configurations that are individually too weak for any realistic gate to select. (vii) A *conformal prediction-set* wrapper over region scores (split conformal on 500 COCO-calibrated regions, nonconformity from vocabulary-ensemble mean and variance) fails to transfer across vocabulary sizes: at a 90% target, coverage drops to 59.7% on VOC-21 with prediction sets covering 72% of the vocabulary, because softmax-based scores are not exchangeable across vocabularies of different cardinality (within-distribution calibration on Context-60 is valid at 92.3%, but that weaker per-dataset form was outside the pre-registered claim). (viii) Two pre-registered probes of the *propagation-smoothing* hypothesis — that LPOSS-style label propagation is robust because it low-pass-filters text-side naming noise over the DINO affinity graph — both failed. A graph-spectral decomposition shows that vocabulary-perturbation logit noise is 97% *low*-frequency on the affinity graph (high-band energy fraction 0.029), i.e. naming noise is spatially class-coherent, leaving spatial or graph smoothing nothing to filter; and transplanting the propagation operator onto the six other methods raises absolute mIoU on every one of them (plain $+1.7$ to $+12.8$) while *increasing* the synonym drop on all six. The robustness advantage of label propagation in Table 7 is therefore a higher-plain-baseline effect, not noise filtering: the tested propagation operator did not repair text-side naming noise. (ix) A pre-registered *presence-gating* wrapper (rank-based gate on top-3 SAM-region-pooled class probabilities, frozen threshold) fails to repair distractor collapse: under 200 near-distractors the gate retains half the vocabulary (precision 0.24 at recall 0.99) and leaves all-class mIoU at $\approx$4 on both ClearCLIP and NACLIP — region-pooled softmax mass

Hypothesis	Signal family	Frozen criterion	Outcome
ZCA/shrinkage whitening, centering	text geometry	no regime harmed	harms almost every regime
Learned text adapter (invariance / normalisation)	learned text	held-out synonym gain	worse than frozen encoder
Learned universal background queries	learned text	match VABS gain	$\leq$half; structurally capped
Learned visual adapter	learned visual	no regime harmed	degrades every regime
Region routing (label-free / few-label / consistency)	CLIP confidence	approach oracle $+14.9$	none approach
Conformal VocabUQ	CLIP confidence	cross-dataset coverage	fails transfer
Presence gating (rank; margin+consistency)	CLIP region stats	precision at recall $\geq$0.98	0.24–0.35; retired†
Propagation smoothing transplant	spatial graph	reduce synonym drop	increases it on all six
Graded frequency predictor	corpus frequency	median $\rho \geq 0.5$	-0.22 (sign reversed)
Text-geometry vocabulary ranking	text geometry	median $\rho \geq 0.5$	0.03
Discrete canonicalisation rewrite	word space	no-harm on plain	-3 to -4
LLM canonicality judge (single config.)	LLM world knowledge	no-harm + recovery	passes no-harm, negative recovery
Transductive logit debiasing (α=0.5)	prediction space	no regime harmed	large negative everywhere
OWLv2 box shield (distractor axis)	external detector	all-class recovery	4.23$\rightarrow$4.27
Text-encoder-lineage banding	model family	third model in band	falsified (OWL-ViT v1)
Alias-diversity class-level law	training corpus	class-level $\rho \geq 0.25$	0.107
Absent-leak class-level predictor	pixel-flow accounting	class-level $\rho \geq 0.25$	-0.20 (pooled contrast real)
GT-presence oracle repair	ground-truth presence	restore all-class mIoU	capped at $\sim$4 (metric, not presence)
Contamination-curve protocol	effective vocabulary	GT-present drop ≥ 3	rises $+6$ to $+9$ from 10 names
Metric-convention re-ranking	accounting convention	ranking change on any axis	Spearman 1.0 everywhere
Pseudo-word free-lunch expansion	non-semantic fillers	$\geq$80% of real-noun gain	52%; beaten by one scalar
Detector-granularity recall law	WordNet depth	$\rho \leq -0.3$	$+0.17$ (sign reversed)

Table 9: Pre-registered negative results (kill table). †: the all-class mIoU bar in the presence-gating pre-registrations was later shown structurally unreachable (App. F); the retirement rests on the precision evidence.

does not separate present from absent vocabulary items at this granularity. A second, pre-registered signal family — raw-cosine margins against VABS-selected negatives combined with a top-region winner-consistency score, explicitly distinct from the killed rank gate — mildly *helps* on official and plain vocabularies ($+0.2$ to $+1.5$) but equally fails the distractor repair criterion (all-class mIoU 3.8/4.2 vs. a pre-registered bar of 20; precision 0.31–0.35 at recall 0.98): image-level presence of a vocabulary item is not recoverable from either softmax-rank or raw-margin region statistics. Scope of this verdict: two signal families, two methods (ClearCLIP/NACLIP), and the high-recall operating regime that protecting ground-truth classes requires (recall $\geq$ 0.98 under the frozen gates); we did not sweep the precision–recall curve for a post-hoc operating point, as both pre-registrations froze their thresholds. Retrospective note: the GT-presence oracle probe (§4.3) later showed that the all-class mIoU bar of 20 is structurally unreachable under our fixed-denominator convention — even a perfect presence gate is capped near 4 — so the mIoU criterion in these two pre-registrations was invalid as frozen; the retirement of presence gating rests on the precision evidence alone (0.24–0.35 at the recall a gate must operate at to protect ground-truth classes — at that precision roughly half the injected vocabulary survives the gate regardless of the metric convention). Within that scope presence gating was retired for this project.

F.1 Index of pre-registered negative results

Table 9 indexes every pre-registered repair or predictor hypothesis retired during this project, with its signal family, frozen criterion, and outcome; full records accompany the artifact.

G Confusion-Flow Decomposition of Perturbation Damage

VocabAudit V1.1 records per-image intersections, unions, and pixel-flow counts for every arm on the 1,349-image split. We resample images with replacement 10,000 times using the same multiplicity vector for the plain baseline and every distractor arm. Table 10 reports paired intervals for each 10-draw cell mean. All 18 intervals exclude zero; the smallest lower endpoint is 4.08 mIoU points. The plain ranking NACLIP $>$ SCLIP $>$ ClearCLIP becomes SCLIP $>$ NACLIP $>$ ClearCLIP for all six cell means, so Kendall $\tau_a = 1/3$ against the plain ranking. The statistic summarizes mean ordering

Table 10: Target-class mIoU change from plain, in points, with 95% paired image-bootstrap intervals. Each row averages 10 vocabulary draws; 10,000 bootstrap samples, seed 0, float64 accumulation.

Cell	SCLIP	NACLIP	ClearCLIP
far +50	6.44 [5.97, 6.91]	4.49 [4.08, 4.88]	4.52 [4.12, 4.92]
mid +50	8.29 [7.72, 8.87]	6.22 [5.67, 6.79]	6.41 [5.87, 6.95]
near +50	7.31 [6.73, 7.89]	5.88 [5.26, 6.50]	6.15 [5.53, 6.77]
far +200	10.87 [10.00, 11.75]	7.21 [6.46, 7.97]	7.80 [7.07, 8.54]
mid +200	12.75 [11.80, 13.70]	9.16 [8.29, 10.07]	10.13 [9.25, 11.00]
near +200	10.93 [9.85, 12.05]	7.95 [6.95, 9.03]	9.22 [8.15, 10.31]

rather than uncertainty in cross-method gaps; draw-level SCLIP/NACLIP order is unanimous for $+200$ but mixed for $+50$.

The pixel ledger prevents a denominator-only explanation. For far/mid/near+200, the fraction of all pixels assigned to distractors is 62.25/67.19/73.47% for SCLIP, 63.52/70.00/76.39% for NACLIP, and 57.40/64.21/71.21% for ClearCLIP. Near+200 absorbs 86.77–91.48% of background pixels and 28.99–35.42% of foreground pixels. Target-class mIoU rises because injected labels remove some false positives from target-class unions; all-class mIoU falls partly because 200 absent classes enter its denominator. Pixel accuracy shows that the predictions themselves also deteriorate. The three measurements therefore answer different questions, and their divergence is the empirical basis for metric-interface incoherence.

H Released Artifacts and Run Provenance

The release accompanying this paper (a public code repository, linked from the arXiv page once posted) contains: (i) all perturbed vocabulary files (synonym 25/50/100% $\times$ 3 seeds, hypernym mappings for VOC-21 and ADE-150, distractor pools and 10 injected vocabularies per rank-tercile/size cell, and the exclusion list {`background`, `unknown`, `other`}); (ii) the perturbation scripts with frozen thresholds; (iii) the evaluation harness implementing the uniform inference protocol; (iv) the fixed image-subset lists; and (v) one JSON record per evaluation run (method, dataset, vocabulary file, whitening configuration, subset, per-class IoU, both mIoU conventions), from which every table cell in this paper is reproduced by a lookup script. The V1.1 package additionally contains signed cache manifests, three independently shuffled score-blinded alias sheets per dataset, pixel-flow ledgers, and the paired-bootstrap JSON. Alias result files are deliberately absent until the three-human gate passes. An early internal run series contained a protocol bug in which the non-semantic `background` class could be synonym-substituted; those runs were archived, the exclusion rule was added to the frozen protocol, and all perturbed vocabulary files were regenerated from the corrected pipeline. Because pool construction and sampling share one code path, the regenerated distractor sets are not byte-identical to the archived ones, which shifts distractor cells slightly (e.g. SCLIP $+200$ near: archived 45.4 vs. corrected 44.5); all legacy synonym and distractor conditions were re-run on the regenerated files. The V1.1 distractor matrix is a separate full-split rerun with rank-tercile strata and 10 draws. Every number reported in this paper, including the full-split and subset-resampling calibrations, comes from a completed, on-disk run record of the corrected series. The constructive experiments of §4.5 follow the same discipline: the VABS generator (lexicon, filter, selection, and guard code), the pre-registration documents with kill criteria for each mechanism variant, the generated vocabularies (including random-negative controls), and one JSON record per run (development-set grid, VOC-21 test arms for all four methods, COCO-Object and PASCAL-Context-60 transfer, per-class IoU for the safety criterion) are all included; the failed guard variants are released alongside the reported ones. During the RF-CLIP anchor runs, one class-name-file swap contaminated a concurrently launched official arm (exposed by the official and syn100 arms returning identical mIoU); the official file was restored from the upstream repository and both affected arms were re-run cleanly — all RF-CLIP numbers reported here come from the clean re-run, and the incident log is preserved in the release.