SEER: SEMANTIC EVIDENCE FOR EXECUTABLE REGION PROPAGATION IN OPEN-VOCABULARY SEMANTIC SEGMENTATION

Anonymous authors
Paper under double-blind review

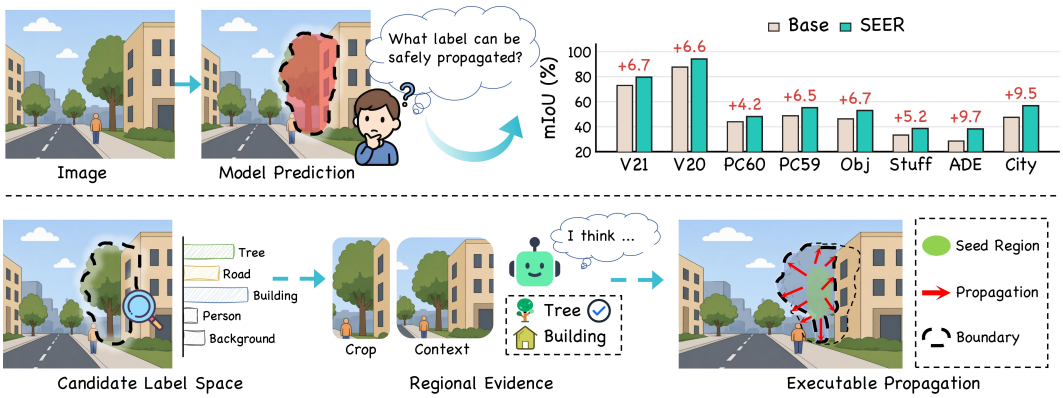

Figure 1: **SEER turns regional semantic evidence into executable OVSS label propagation.** Starting from a fixed OVSS prediction, SEER selects ambiguous proposals, derives compact candidate labels from the active score space, and asks a VLM to verify regional evidence. The OVSS system executes certified evidence: score-supported seeds guide label propagation, while boundary-conservative guards restrict unsafe spreading near object and structure boundaries. The bar chart shows consistent mIoU gains across diverse OVSS benchmarks.

ABSTRACT

Training-free open-vocabulary semantic segmentation (OVSS) has made strong progress by adapting CLIP-like representations to dense prediction, but its remaining errors are often local, structured, and tied to the active label space. A successful correction must be visually plausible, legal under the target vocabulary, and executable in the segmentation score field. We propose SEER, a training-free test-time repair layer for Semantic Evidence and Executable Region propagation. SEER casts post-prediction OVSS repair as evidence-conditioned label propagation. The OVSS score space proposes compact candidate labels, a frozen VLM verifies closed-set regional evidence, and an execution gate certifies whether the evidence is legally and semantically actionable. The accepted label is then propagated by an OVSS-side policy: coherent high-confidence regions allow broad propagation, while object- or boundary-risk cases use conservative score-grounded propagation. Across eight standard OVSS benchmarks, SEER reaches 58.0 average mIoU without updating OVSS weights. Matched-input controls isolate the repair mechanism, while resource measurements report the cost of the complete frozen system.

1 INTRODUCTION

Open-vocabulary semantic segmentation (OVSS) assigns dense labels from an open or user-specified vocabulary. Starting from early open-vocabulary scene parsing Zhao et al. (2017), the field has been accelerated by CLIP Radford et al. (2021) and OpenCLIP Cherti et al. (2023). Training-free methods further improve dense prediction by modifying CLIP attention, calibrating patch-text similarity, and injecting visual priors. MaskCLIP Zhou et al. (2022) exposes dense CLIP predictions, SCLIP Wang et al. (2024) adapts self-attention for segmentation, ProxyCLIP Lan et al. (2024b) introduces proxy

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.

features, Trident Shi et al. (2025) restructures feature aggregation, and CorrCLIP Zhang et al. (2025) corrects patch correlation. As these systems become stronger, the bottleneck shifts from global recognition to local residues such as boundary spillovers, object-part confusions, and stuff/object ambiguities.

Residual repair requires more than crop classification. OVSS predictions live under an active label space, where a visually meaningful region can be non-target when its label is absent, while a scene or stuff region becomes executable once the corresponding label enters the vocabulary. A repair action therefore has three coupled requirements. The label must be visually supported, legal under the active vocabulary, and propagated to the right pixels. Regional evidence can suggest a semantic direction, while pixel support must still be recovered from the dense prediction.

General-purpose vision-language models (VLMs) provide a natural source of regional semantic evidence. Qwen-VL Qwen Team (2026), LLaVA-OneVision-1.5 An et al. (2025), and GPT-4o Hurst et al. (2024) can compare candidate descriptions and answer localized visual questions. Direct pixel decisions remain unreliable because regional semantics, label-space validity, and pixel execution become entangled in one response. A VLM may name the highlighted region correctly without knowing whether the label is executable in the current OVSS score field or how far it should propagate inside the proposal. We therefore restrict the VLM to regional evidence and retain pixel authority in the segmentation model.

This paper studies how regional VLM evidence can be made executable in OVSS label propagation. Our answer is SEER, short for **S**emantic **E**vidence and **E**xecutable **R**egion propagation. SEER uses prediction-side signals from the baseline mask and score tensor to select auditable proposals, asks a VLM to verify a compact score-derived candidate set with `uncertain` as an abstention option, and accepts evidence only if it passes executable gates. After certification, the OVSS score field executes the decision through a propagation policy. Coherent high-confidence regions can use broad label propagation, while object-like or boundary-sensitive regions use conservative support from the score field.

This framing keeps the method tied to OVSS. Labels are proposed by open-vocabulary scores, action validity is checked against the active label space, and pixel changes are produced by score-space propagation. Broad propagation is used after closed-set evidence verification, while conservative propagation remains the fallback when boundaries or selector noise make broad support risky. Across eight standard benchmarks, SEER reaches 58.0 average mIoU with the ViT-L/14 base. Fixed-prediction controls hold the score tensor, proposals, candidates, parser, and audit budget constant while varying the repair signal or executor. We also report complete-system latency and memory, lighter verifiers, proposal-source transfer, sensitivity tests, and residual failures.

This paper makes three contributions.

- We formulate OVSS post-prediction repair as executable region propagation, separating regional semantic evidence, label-space validity, and pixel support.
- SEER is a training-free evidence-to-execution layer that uses frozen VLM verification while leaving pixel propagation to the OVSS score field.
- Systematic diagnostics evaluate compact candidates, executable gates, multi-view evidence, evidence-conditioned propagation, and complete-system cost.

2 Related Work

Open-vocabulary semantic segmentation. OVSS extends zero-shot and open-vocabulary recognition to dense prediction Zhao et al. (2017); Bucher et al. (2019); Xian et al. (2019). CLIP Radford et al. (2021) and OpenCLIP Cherti et al. (2023) provide the basic image-text alignment, but their image-level training objective does not directly produce spatially coherent patch predictions. Training-free OVSS methods therefore change how CLIP features are extracted, calibrated, or aggregated. MaskCLIP Zhou et al. (2022) exposes dense CLIP predictions, CLIP Surgery Li et al. (2025) modifies CLIP behavior for open-vocabulary localization, SCLIP Wang et al. (2024) rethinks self-attention for dense inference, CLIPtrase Shao et al. (2024) studies CLIP's segmentation ability, and ProxyCLIP Lan et al. (2024b) uses proxy attention from stronger visual features. More recent methods improve calibration, residual structure, correlations, and feature aggregation Lan et al. (2024a);

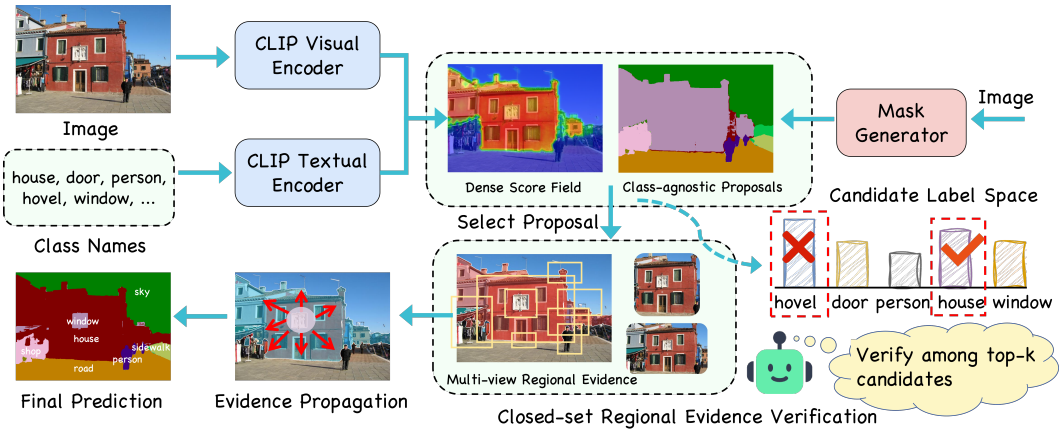

Figure 2: **Overview of SEER.** SEER converts regional VLM evidence into executable OVSS label propagation. From a fixed prediction, it selects ambiguous proposals, derives score-based candidates, and verifies them with multi-view closed-set reasoning. Score support and vocabulary legality certify the verified label, and the OVSS model controls its pixel propagation.

Hajimiri et al. (2025); Yang et al. (2025); Bai et al. (2025); Sun et al. (2025); Kim et al. (2025); Wang et al. (2026); Shi et al. (2025); Zhang et al. (2025). These works mainly seek a stronger initial OVSS prediction. SEER instead repairs residual local errors after that prediction is fixed.

Region priors and label propagation. Region and structure priors are widely used in segmentation. Training-based open-vocabulary systems introduce dense supervision, adapters, mask classification, or generative priors Rao et al. (2022); Xu et al. (2022); Liang et al. (2023); Xu et al. (2023b); Cho et al. (2024); Lee et al. (2025); Zou et al. (2023a;b); Xu et al. (2023a); Cha et al. (2023); Wu et al. (2024). Class-agnostic and grounding models such as SAM, SAM2, EntitySeg, RegionCLIP, and Grounding DINO provide masks, regions, or localization priors that are useful beyond their original tasks Kirillov et al. (2023); Ravi et al. (2024); Qi et al. (2023); Zhong et al. (2022); Liu et al. (2024). Classical CRFs, graph cuts, propagation methods, and modern segmentation architectures also show that spatial consistency and pixel-level confidence matter for reliable masks Krähenbühl & Koltun (2011); Rother et al. (2004); Chen et al. (2018); Xie et al. (2021); Cheng et al. (2021; 2022). SEER treats each proposal as a candidate domain over which verified evidence may propagate. The propagation distance is conditioned on evidence confidence, region form, score support, and risk.

Vision-language models for visual verification. Modern VLMs can answer region-level questions, compare candidate descriptions, and provide visual evidence that is not captured by local score margins Qwen Team (2026); An et al. (2025); Zhu et al. (2025); Clark et al. (2026); Hurst et al. (2024); Bousselham et al. (2024). This makes them appealing for segmentation repair, but free-form VLM outputs are poorly matched to a fixed OVSS label space. A generic region classifier may name the crop, yet a repair system must also know whether the label is legal and whether the segmentation system can execute the label at pixels. SEER uses the VLM only as a closed-set verifier over score-derived candidates. The executable decision remains inside the OVSS label and score space, which separates our setting from direct mask correction or ordinary region classification.

3 METHOD

Overview. SEER repairs a fixed OVSS prediction in four steps (Fig. 2). It (1) selects a small proposal subset, (2) builds legal candidate labels from the frozen score tensor, (3) asks a frozen VLM to select one candidate or abstain, and (4) lets the OVSS scores decide which pixels may change. High-confidence coherent regions may use the full proposal; object-like or boundary-sensitive cases use a conservative score-supported subset. The VLM therefore supplies only a regional label decision and never predicts or edits a mask.

3.1 PROBLEM SETUP: A LABEL AND ITS PIXEL SUPPORT

Let $I \in \mathbb{R}^{H\times W\times 3}$ be an image and $\Omega = \{1,\dots,H\} \times \{1,\dots,W\}$ be the pixel lattice. The active OVSS label space is $\mathcal{C} = \{1,\dots,C\}$. A fixed OVSS model outputs a prediction map $Y^0 \in \mathcal{C}^{H\times W}$ and a dense score tensor $S \in \mathbb{R}^{H\times W\times C}$. We use $S_{p,c}$ for the score of class c at pixel p, and m_p for the top-1/top-2 score margin. A fixed SAM2-based class-agnostic proposal bank provides regions $\mathcal{M} = \{M_i\}_{i=1}^{N}$, where $M_i \subseteq \Omega$. The proposal bank is not a semantic predictor in our method; it only defines candidate spatial domains for later evidence verification and score-side execution.

For a proposal M_i, a regional label is not yet a repair. It becomes executable only after the system determines how the label should propagate. We write an executable propagation action as

$$r_i = (\ell_i, W_i), \qquad \ell_i \in \mathcal{C}, \quad W_i \subseteq M_i, \tag{1}$$

where ℓ_i is the verified target label and W_i is the propagation support. Applying r_i updates the current prediction by

$$Y_p^{t+1} = \begin{cases} \ell_i, & p \in W_i, \\ Y_p^t, & p \notin W_i, \end{cases} \tag{2}$$

where t denotes the execution order. Regional evidence determines the target label, and the OVSS system determines the propagation support W_i.

For clarity, we write the three stages of SEER as

$$\mathcal{M} \xrightarrow{\text{Prioritize}} \mathcal{R}(I), \qquad (I, M_i, \mathcal{A}_i) \xrightarrow{\text{Verify}} a_i, \qquad (S, Y^0, M_i, a_i) \xrightarrow{\text{Propagate}} (\ell_i, W_i). \tag{3}$$

Here $\mathcal{R}(I) \subseteq \mathcal{M}$ is the selected auditable subset, $\mathcal{A}_i$ is the candidate label set constructed below, and a_i is the VLM evidence tuple. This factorization is the design principle: region selection limits the VLM budget, verification extracts closed-set evidence, and propagation decides whether the evidence should spread broadly or conservatively.

The pair (ℓ_i, W_i) makes the division of labor explicit: the VLM helps choose a regional label, while the OVSS scores choose the pixels that can receive it. A correct label can still increase error if W_i crosses an already-correct boundary, and a safe support can miss useful corrections if it is too small. Appendix B gives the exact benefit–harm decomposition and a short non-identifiability result showing why regional evidence alone cannot determine pixel support.

3.2 CHOOSING PROPOSALS UNDER AN AUDIT BUDGET

Querying every proposal is unnecessary and expensive. SEER selects a budgeted subset using information available before VLM inference. For each proposal, the dominant baseline label is

$$b_i = \mathrm{mode}\{Y_p^0 \mid p \in M_i\}. \tag{4}$$

The mean score of class c inside the proposal is

$$\bar{s}_{i,c} = \frac{1}{|M_i|} \sum_{p\in M_i} S_{p,c}, \tag{5}$$

and the strongest non-baseline competitor is

$$c_i = \arg\max_{c\in\mathcal{C}\setminus\{b_i\}} \bar{s}_{i,c}. \tag{6}$$

The priority function is a fixed utility proxy over four prediction-side signals: regional coherence, potential impact, score-space competition, and risk exposure. We write

$$\pi_i = \pi(\mathrm{coh}_i, \mathrm{area}_i, \mathrm{comp}_i, \mathrm{risk}_i)\,, \tag{7}$$

where coh_i measures the consistency of Y^0 inside M_i, $\mathrm{area}_i = |M_i|/|\Omega|$, comp_i measures the score competition between c_i and b_i, and risk_i measures exposure to object or boundary-sensitive predictions. All terms are computed from Y^0, S, and M_i. The exact constants and routing predicates are fixed once and listed in Appendix C. The priority score is a budget controller that exposes a small auditable set. The selected set under budget K is

$$\mathcal{R}(I) = \mathrm{TopK}_{M_i\in\mathcal{M}}\, \pi_i. \tag{8}$$

3.3 CLOSED-SET REGIONAL VERIFICATION

For each selected proposal $M_i \in \mathcal{R}(I)$, SEER constructs a compact candidate set in the OVSS label space:

$$\mathcal{A}_i = \{b_i\} \cup \mathcal{Q}_i \cup \{\texttt{uncertain}\}, \tag{9}$$

where $\mathcal{Q}_i \subseteq \mathcal{C} \setminus \{b_i\}$ contains up to three score-derived alternatives from proposal-averaged scores and local score support. In the default setting, this gives a closed candidate set consisting of the current base label, the top-3 non-base alternatives, and `uncertain`. The VLM is not asked to search the full vocabulary or invent a label; labels outside $\mathcal{A}_i$ are not executable.

The VLM sees multiple views of the same proposal, including the full image with M_i highlighted and cropped views around the proposal. It returns an evidence tuple

$$a_i = (\hat{\ell}_i, \mathrm{conf}_i, \mathrm{form}_i), \tag{10}$$

where $\hat{\ell}_i \in \mathcal{A}_i$ is the selected label, $\mathrm{conf}_i \in \{\texttt{high}, \texttt{medium}, \texttt{low}\}$ is confidence, and form_i indicates whether the region is coherent, boundary-like, mixed, or unclear. This tuple specifies a label direction and a region form; pixel support is still decided by the OVSS score field.

3.4 FILTERING VLM ANSWERS BEFORE WRITEBACK

The VLM answer is filtered before it can affect any pixel. The executable indicator is

$$e_i = \mathbf{1}\Big[\hat{\ell}_i \in \mathcal{C},\ \hat{\ell}_i \neq b_i,\ \hat{\ell}_i \neq \texttt{uncertain},\ \mathrm{conf}_i = \texttt{high},\ \mathrm{form}_i \notin \{\mathrm{mixed}, \mathrm{unclear}\}\Big]. \tag{11}$$

If $e_i = 0$, the proposal leaves the prediction unchanged. If $e_i = 1$, the target label is $\ell_i = \hat{\ell}_i$. Thus every accepted answer must be legal, high-confidence, different from the current label, and not mixed or unclear. Boundary-like answers may still be used, but only through conservative pixel support. Section 4.4 and Appendix G test the filter and propagation choices under matched inputs.

3.5 SCORE-CONTROLLED PIXEL UPDATES

Once the label is certified, the remaining question is how far the evidence should propagate. SEER uses a policy family Γ that maps certified evidence and OVSS scores to a support set:

$$W_i = \Gamma(S, Y^0, M_i, \ell_i, a_i), \qquad W_i \subseteq M_i. \tag{12}$$

We use three propagation regimes. *Coherent propagation* sets $W_i = M_i$ when a region is coherent, high-confidence, and low-risk; in this case, the proposal is treated as a semantic unit. *Score-grounded propagation* propagates only to pixels where the target remains supported by the OVSS score field. *Boundary-conservative propagation* restricts propagation to low-margin pixels for object-like or risk-sensitive cases.

For score-grounded propagation, let

$$y_p^{\mathrm{top}} = \arg\max_{c \in \mathcal{C}} S_{p,c}. \tag{13}$$

We use the support predicate

$$\rho(p, \ell_i, b_i) = \mathbf{1}\big[y_p^{\mathrm{top}} = \ell_i \ \vee\ S_{p,\ell_i} \geq S_{p,b_i} - \tau_s \ \vee\ \big(m_p \leq \tau_m \wedge S_{p,\ell_i} \geq S_{p,b_i} - \tau_l\big)\big], \tag{14}$$

where τ_s, τ_m, τ_l are fixed thresholds. The score-grounded support is $W_i = \{p \in M_i \mid e_i = 1, \rho(p, \ell_i, b_i) = 1\}$; the boundary-conservative support further restricts it to low-margin pixels. In practice, SEER uses the same deterministic contract for all datasets: broad propagation is reserved for reliable context/stuff evidence, while object, small-region, boundary, or high-risk cases use conservative support after closed-set certification.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Evaluation protocol. Following common open-vocabulary semantic segmentation (OVSS) protocols Zhou et al. (2022); Wang et al. (2024); Lan et al. (2024b); Shi et al. (2025); Zhang et al. (2025),

Table 1: **Complete-system accuracy on eight OVSS benchmarks.** Published baseline numbers follow the CorrCLIP-style evaluation configuration and are grouped by CLIP size. SEER is an accuracy–cost point, not an equal-resource comparison: it adds fixed SAM2 proposals and sparse local Qwen3.5-9B verification to CorrCLIP. Paired end-to-end cost and matched-input attribution are reported in Tab. 7 and Tab. 4.

Method	Size	VOC21	VOC20	PC60	PC59	Object	Stuff	ADE	City	Avg.
Training-based										
TCL Cha et al. (2023)		55.0	83.2	30.4	33.9	31.6	22.4	17.1	24.0	37.2
CLIP-DINOiser Wysoczańska et al. (2024)		62.1	80.9	32.4	35.9	34.8	24.6	20.0	31.7	40.3
CoDe Wu et al. (2024)	ViT-B/16	57.7	–	30.5	–	32.3	23.9	17.7	28.9	–
LoGoSeg Chen et al. (2026)		78.6	95.4	–	58.2	–	–	32.6	–	–
ESC-Net Lee et al. (2025)		80.1	97.3	–	59.0	–	–	35.6	–	–
Training-free										
CLIP Radford et al. (2021)		11.5	41.9	4.4	9.2	1.6	4.4	2.9	5.0	10.1
MaskCLIP Zhou et al. (2022)		38.8	74.9	23.6	26.4	20.6	16.4	9.8	12.6	27.9
ClearCLIP Lan et al. (2024a)		51.8	80.9	32.6	35.9	33.0	23.9	16.7	30.0	38.1
SCLIP Wang et al. (2024)		59.1	80.4	30.4	34.2	30.5	22.4	16.1	32.2	38.2
ProxyCLIP Lan et al. (2024b)		61.3	80.3	35.3	39.1	37.5	26.5	20.2	38.1	42.3
LaVG Kang & Cho (2024)		62.1	82.5	31.6	34.7	34.2	23.2	15.8	26.2	38.8
CLIPtrase Shao et al. (2024)		53.0	81.2	30.8	34.9	44.8	24.1	17.0	–	–
NACLIP Hajimiri et al. (2025)	ViT-B/16	64.1	83.0	35.0	38.4	36.2	25.7	19.1	38.3	42.5
Trident Shi et al. (2025)		67.1	84.5	38.6	42.2	41.1	28.3	21.9	42.9	45.8
ResCLIP Yang et al. (2025)		61.1	86.0	33.5	36.8	35.0	24.7	18.0	35.9	41.4
SC-CLIP Bai et al. (2025)		64.6	84.3	36.8	40.1	37.7	26.6	20.1	41.0	43.9
CLIPer Sun et al. (2025)		65.9	85.2	37.6	41.7	39.0	27.5	21.4	–	–
CASS Kim et al. (2025)		65.8	87.8	36.7	40.2	37.8	26.7	20.4	39.4	44.4
CorrCLIP Zhang et al. (2025)		74.8	88.8	44.2	48.8	43.7	31.6	26.9	49.4	51.0
SEER (Ours)		**79.3** (+4.5)	**93.8** (+5.0)	**47.7** (+3.5)	**53.9** (+5.1)	**48.5** (+3.7)	**36.2** (+4.6)	**35.2** (+8.3)	**55.0** (+5.6)	**56.2** (+5.2)
FreeDA Barsellotti et al. (2024)		55.4	87.9	38.3	43.5	37.4	28.8	23.2	36.7	43.9
CaR Sun et al. (2024)		67.6	91.4	30.5	39.5	36.6	–	17.7	–	–
ProxyCLIP Lan et al. (2024b)		60.6	83.2	34.5	37.7	39.2	25.6	22.6	40.1	43.0
ResCLIP Yang et al. (2025)	ViT-L/14	54.1	85.5	30.9	34.5	32.5	23.4	18.2	33.7	39.1
SC-CLIP Bai et al. (2025)		65.0	88.3	36.9	40.6	40.5	26.9	21.7	41.3	45.2
CLIPer Sun et al. (2025)		69.8	90.0	38.0	43.6	43.3	28.7	24.4	–	–
CorrCLIP Zhang et al. (2025)		76.7	91.5	44.9	50.8	49.4	34.0	30.7	51.1	53.6
SEER (Ours)		**79.6** (+2.9)	**94.2** (+2.7)	**48.2** (+3.3)	**55.2** (+4.4)	**52.9** (+3.5)	**38.6** (+4.6)	**38.3** (+7.6)	**56.9** (+5.8)	**58.0** (+4.4)

we evaluate on eight benchmarks from six validation sets: PASCAL VOC with and without background (VOC21/VOC20) Everingham et al. (2015), PASCAL Context with and without background (PC60/PC59) Mottaghi et al. (2014), COCO Object and COCO-Stuff164K Caesar et al. (2018), ADE20K Zhou et al. (2019), and Cityscapes Cordts et al. (2016). We report mean Intersection over Union (mIoU). Comparisons with prior methods follow the CorrCLIP-style evaluation configuration: published numbers are grouped by CLIP backbone size, and our method is evaluated with both CLIP ViT-B/16 and ViT-L/14 settings.

Implementation details. We follow the CorrCLIP evaluation configuration, while the repair input itself is a fixed CLIP score field S and its induced prediction Y^0. Region domains come from a fixed SAM2 (Ravi et al., 2024) proposal bank and are never treated as semantic predictions. The default verifier is the locally deployable Qwen3.5-9B (Qwen Team, 2026), chosen for strong regional recognition, open weights, and deterministic decoding; it is not fine-tuned on any evaluation benchmark. Its public documentation does not provide a sample-level pretraining manifest, so we do not assume zero overlap with segmentation or grounding data and instead test replacement by three other VLM families and CLIP. The VLM receives proposal views and legal label names. Ground truth, target masks, and pixel-write supervision are excluded.

For mechanism studies, we fix Y^0, S, the proposal bank, candidates, parser, gate, propagation policy, audit inputs, and evaluation split; only the tested repair signal or executor changes. This *fixed-prediction protocol* uses MetaCLIP FullCC ViT-B/16 QuickGELU, frozen DINO ViT-B/8, benchmark-native mappings, and aspect-ratio-preserving resolution $(2048, 336)$ for VOC/PC and $(2048, 448)$ for City/ADE. Its initial scores are a separate locally frozen diagnostic state, not a reproduction of the published CorrCLIP row in Tab. 1. The constants were selected once from aggregate diagnostics on the evaluated natural-image validation domains, then frozen across all eight benchmarks, two backbones, four OVSS bases, and every verifier; we do not claim domain-blind development or medical/remote-sensing validation. Appendix C lists the complete contract and sensitivity tests.

4.2 Comparison with State-of-the-Art Methods

Complete-system comparison. Table 1 compares complete-system accuracy using published baseline results; it does not imply matched hardware or equal external resources. With both ViT-B/16 and ViT-L/14, SEER improves its corresponding CorrCLIP base on every benchmark, with the

Table 2: **Component diagnostics.** Each row is an alternative repair setting, not a cumulative build-up. The initial values are an independently frozen diagnostic S, Y^0, not the published CorrCLIP row in Tab. 1; every delta uses this identical state.

Setting	VLM	Closed	Multi-view	Coherent	VOC21	PC60	City	ADE	Avg.
Initial prediction					72.88	44.01	47.41	28.63	48.23
Score-only rule				✓	+2.10	+3.80	+2.53	+2.13	+2.64
Free-form VLM	✓		✓	✓	+2.82	+0.00	+0.00	+0.00	+0.70
Single view	✓	✓		✓	+2.43	+0.30	+1.68	+0.67	+1.27
Conservative exec.	✓	✓	✓		+3.73	+2.47	+4.05	+7.06	+4.33
SEER	✓	✓	✓	✓	**+6.72**	**+4.19**	**+9.49**	**+9.67**	**+7.52**

largest gains on ADE20K and Cityscapes. The remaining experiments use one frozen prediction state and matched proposal/audit inputs to attribute the gain, while Tab. 7 reports the paired cost on our hardware.

4.3 Fixed-Prediction Repair Analysis

Fixed-prediction repair. The fixed-prediction diagnostic in Fig. 1 reports the controlled setting used in the mechanism analysis. The prediction map and score tensor are fixed, and only the evidence-to-execution layer is applied. Gains across eight benchmarks rule out changes in backbone, resolution, or training recipe, supporting the premise that residual OVSS errors are repairable once regional evidence becomes executable.

Plug-in transfer across OVSS bases. We next ask whether the repair layer is tied to one score geometry. The same SEER pipeline is applied to ProxyCLIP, Trident, ResCLIP, and CorrCLIP: proposals, candidates, VLM auditing, filters, propagation rules, coefficients, and thresholds stay fixed across bases and benchmarks. Only the OVSS prediction and score cache change. Every reported base–benchmark pair improves, with mean gains from 4.53 to 5.93 mIoU; full results and confidence intervals appear in Appendix I.

4.4 Mechanism Analysis

Component chain. The next experiments follow the method decomposition. The full-chain diagnostic in Table 2 compares SEER with simpler repair configurations. Each alternative breaks one link in the deterministic execution contract: visual verification, label-space restriction, multi-view context, or high-recall coherent propagation. Within this diagnostic, all variants reduce the average gain, supporting SEER as a coupled design whose gain depends on both regional evidence and score-side execution.

Table 3: **Evidence interface.** Free-form asks the VLM to name a label, full label list exposes the whole vocabulary, and closed top-k contains the current/base label, up to k non-base score-derived alternatives, and `uncertain`.

Interface	VOC21	PC60	City	ADE	Avg.
Free-form	+2.82	+0.00	+0.00	+0.00	+0.70
Full label list	+5.97	+0.55	+6.02	+6.56	+4.78
Closed top-1	+5.68	+1.38	+7.95	+8.43	+5.86
Closed top-2	**+6.94**	+2.22	+6.79	+8.93	+6.22
Closed top-3	+6.72	**+4.19**	**+9.49**	**+9.67**	**+7.52**

Figure 3: **Execution frontier.** Bars show Δ mIoU on VOC, PC, and ADE; the dashed line reports certified support. SEER reaches the strongest gains by turning broader verified evidence into executable propagation.

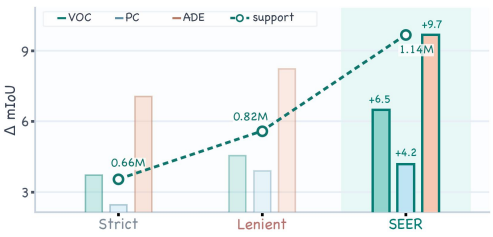

Table 4: **Matched-input controls.** Every row uses the same frozen prediction, proposals, candidates, audit pool, and 2,000-audit cap.

Control	VOC21	PC60	City	ADE	Avg.
OVSS score rule	+2.10	+3.80	+2.53	+2.13	+2.64
CLIP crop classifier	+3.88	+2.38	+2.70	+5.80	+3.69
VLM + score support	+2.09	+0.87	+1.74	+2.11	+1.70
Dense-CRF smoothing	+0.01	−0.01	+0.02	+0.01	+0.01
SEER	**+6.72**	**+4.19**	**+9.49**	**+9.67**	**+7.52**

Table 5: **Multi-view evidence.** Results use the default audit pool, verifier, candidates, gate, and propagation policy.

View	VOC21	PC60	City	ADE	Avg.
Full image	+4.26	+2.18	+4.43	+4.71	+3.90
Local crop	+4.91	+2.76	+5.88	+4.64	+4.55
Tight crop	+3.48	+2.07	+3.96	+3.52	+3.26
Three views	**+6.72**	**+4.19**	**+9.49**	**+9.67**	**+7.52**

Evidence interface and execution. The interface study in Table 3 shows why the verifier should operate in a bounded label space. Free-form answers may miss the active vocabulary, while the full vocabulary introduces legal but weakly supported choices. With the default top-3 alternatives, the analysis-only target-in-set coverage is 99.9% across the four mechanism benchmarks. Filtering raises decision precision from 57.9% before execution to 82.3% among executed answers; 36.2% of audits leave the prediction unchanged and only 0.2% are invalid or out of set (Appendix D). Figure 3 gives the matching pixel-support story: strict routing is safe but low-recall, whereas SEER expands only high-confidence, score-compatible regions.

Controls and visual context. Under the matched inputs in Tab. 4, a score rule, ordinary CLIP crop classifier, VLM answer with only score support, and Dense-CRF smoothing all remain below full SEER. The gain therefore is not reproduced by proposals plus a heuristic, a stronger regional classifier, an isolated VLM answer, or generic smoothing alone. The view ablation in Tab. 5 further shows that a tight crop is insufficient: global scene context, local boundaries, and region appearance provide complementary evidence.

Table 6: **VLM auditor transfer.** All rows use the same replay protocol and differ only in the regional verifier.

Family	Scale	VOC21	PC60	City	ADE	Avg.
Qwen	0.8B	+2.23	+0.34	+0.70	+2.64	+1.48
Qwen	4B	+5.86	+3.26	+7.49	+8.45	+6.26
Qwen	9B	**+6.72**	**+4.19**	**+9.49**	**+9.67**	**+7.52**
LLaVA-OV-1.5	4B	+5.44	+2.30	+6.46	+6.21	+5.10
LLaVA-OV-1.5	8B	+6.01	+3.72	+8.35	+8.52	+6.65
InternVL3	1B	+1.38	+1.41	+1.43	+1.25	+1.37
InternVL3	9B	+5.38	+3.91	+7.43	+7.93	+6.16
Molmo2	4B	+5.17	+2.65	+6.28	+6.63	+5.18
Molmo2	9B	+5.64	+3.84	+8.77	+8.95	+6.80

Figure 4: **Audit budget.** More proposals improve repair quality; SEER stays above score-gated execution.



4.5 Cost and Deployment Scope

Table 7: **End-to-end cost (s/image).** Validation-weighted measurements on the same hardware. Cached SEER reuses fixed SAM2 proposals; cold SEER includes proposal generation. Audit load averages 0.71 calls per image.

Configuration	Proposal gen.	OVSS	Audit	Writeback	Total	Overhead
Frozen OVSS	–	5.042	–	–	5.042	–
Cached SEER	–	5.042	0.424	0.043	5.509	+9.3%
Cold SEER	0.508	5.042	0.424	0.043	6.018	+19.3%

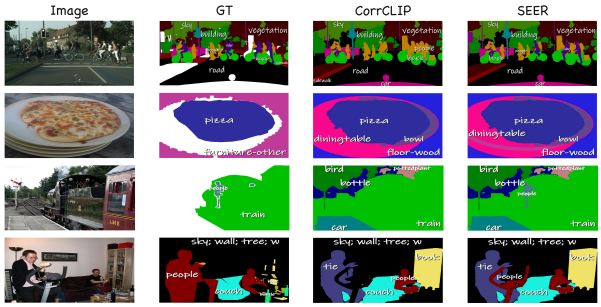

Figure 5: **Residual failure cases.** Mixed proposals, thin structures, and weak score support can leave errors near boundaries. Additional examples appear in Appendix K.

Robustness checks. Joint threshold scaling changes the gain by at most 0.14 mIoU. One-factor sweeps give a largest per-dataset range of 0.083 mIoU. Replaying 8,000 decisions with three shuffled candidate orders gives a largest per-dataset decline of 0.078 mIoU. These checks use cached verifier outputs and leave the remaining pipeline fixed.
Failure boundary. Residual errors concentrate in mixed proposals, thin structures, candidate misses, and fragmented score support. The gate and conservative route reduce harmful writes, while some difficult regions remain unchanged.

Proposal and execution audit. We further audit 2,000 selected SAM2 proposals, with 500 from each mechanism benchmark. Ground truth is used only for this post-hoc analysis after proposal selection, VLM decisions, and execution have been fixed. Table 8 separates proposal quality from the behavior of the executable gate.

Table 8: **Proposal quality and execution coverage.** Purity and semantic correctness are analysis-only measurements over the fixed audit pool.

Diagnostic	Observation	Execution outcome
GT purity ≥ 0.95	1,732 / 2,000 (86.6%)	+0.466 mean mIoU
GT purity < 0.70	71 / 2,000 (3.55%)	43 / 71 filtered
Low-purity bin effect	-0.0055 / $+0.0018$	Near-zero aggregate effect
Correct high-confidence decisions	1,397	1,061 base confirmations
Correct non-base directions	336	327 executed (97.32%)
Empty supported execution	0	9 safely retain Y^0

The useful gain is concentrated in high-purity proposals, while most low-purity proposals are suppressed and have negligible aggregate effect. Among correct non-base directions, the score-side executor applies 327 of 336 decisions; the remaining nine preserve the initial prediction, and none fails through an empty support. This analysis addresses proposal contamination and overly conservative execution separately.

5 Conclusion and Limitations

SEER frames OVSS repair as executable region propagation after a fixed prediction has been produced. It couples closed-set VLM evidence with OVSS score-space execution: the verifier identifies actionable regional labels, and the segmenter decides their pixel support. This evidence-to-execution design supports high-recall propagation on coherent regions while using conservative support near uncertain boundaries. Across eight benchmarks, cross-base plug-in tests, and protocol-matched controls, the experiments show consistent repair gains with controlled proposal-level intervention cost. The controls attribute the gain to the coupling of closed-set regional evidence and score-controlled execution; proposals, crop classification, and smoothing each remain below the full system. The complete system is intended for offline or selective repair. It adds 9.3% cached or 19.3% cold latency and uses 18.7GB peak verifier memory in the measured configuration. Proposal misses, absent candidates, verifier errors, and weak score support remain important limitations, especially for thin or mixed regions. Future work will study lighter auditing, stronger proposal coverage, and more robust boundary execution.

Reproducibility statement

Appendix C documents the execution contract, prompts, decoding settings, preprocessing, dataset mappings, and evaluation protocol. It also reports matched-input controls, threshold and candidate-order sensitivity, verifier and proposal-source transfer, and complete resource measurements. All benchmarks and frozen models are public, and the same parameter profile is used throughout. An anonymized implementation and run commands will accompany the submission.

Ethics statement

This work uses public vision benchmarks and frozen pretrained models, with no human-subject study or personal-data collection. Deployments may inherit biases, privacy risks, and licensing constraints from the component models and data sources. Each audit exposes image crops to the verifier, so sensitive deployments should use an approved local verifier and follow all component licenses. SEER is intended for offline or selective repair and should not be used as a safety-critical decision system. Abstention and score-constrained writeback reduce harmful edits, while semantic and boundary errors can remain.

AI use statement

Generative AI tools assisted with language editing and ICLR formatting. The reported method, experiments, results, mathematical claims, and code were produced and verified by the authors. The authors reviewed the assisted text and take responsibility for the submission.

References

Xiang An, Yin Xie, Kaicheng Yang, Wenkang Zhang, Xiuwei Zhao, Zheng Cheng, Yirui Wang, Songcen Xu, Changrui Chen, Chunsheng Wu, Huajie Tan, Chunyuan Li, Jing Yang, Jie Yu, Xiyao Wang, Bin Qin, Yumeng Wang, Zizhen Yan, Ziyong Feng, Ziwei Liu, Bo Li, and Jiankang Deng. Llava-onevision-1.5: Fully open framework for democratized multimodal training, 2025. URL https://arxiv.org/abs/2509.23661.

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

Luca Barsellotti, Roberto Amoroso, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. Training-free open-vocabulary segmentation with offline diffusion-augmented prototype generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3689–3698, 2024.

Walid Bousselham, Felix Petersen, Vittorio Ferrari, and Hilde Kuehne. Grounding everything: Emerging localization properties in vision-language transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3828–3837, 2024.

Maxime Bucher, Tuan-Hung Vu, Matthieu Cord, and Patrick Pérez. Zero-shot semantic segmentation. *Advances in Neural Information Processing Systems*, 32, 2019.

Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Coco-stuff: Thing and stuff classes in context. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1209–1218, 2018.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, et al. Sam 3: Segment anything with concepts. *arXiv preprint arXiv:2511.16719*, 2025.

Junbum Cha, Jonghwan Mun, and Byungseok Roh. Learning to generate text-grounded mask for open-world semantic segmentation from only image-text pairs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11165–11174, 2023.

Junyang Chen, Xiangbo Lv, Zhiqiang Kou, Xingdong Sheng, Ning Xu, and Yiguo Qiao. Logoseg: Integrating local and global features for open-vocabulary semantic segmentation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(4):2886–2894, 2026. doi: 10.1609/aaai.v40i4.37279. URL https://ojs.aaai.org/index.php/AAAI/article/view/37279.

Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 801–818, 2018.

Bowen Cheng, Alex Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. *Advances in neural information processing systems*, 34:17864–17875, 2021.

Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1290–1299, 2022.

Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2818–2829, 2023.

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

Christopher Clark, Jieyu Zhang, Zixian Ma, Jae Sung Park, Mohammadreza Salehi, Rohun Tripathi, Sangho Lee, Zhongzheng Ren, Chris Dongjoo Kim, Yinuo Yang, et al. Molmo2: Open weights and data for vision-language models with video understanding and grounding. *arXiv preprint arXiv:2601.10611*, 2026.

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

Mark Everingham, SM Ali Eslami, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes challenge: A retrospective. *International journal of computer vision*, 111(1):98–136, 2015.

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

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

Dahyun Kang and Minsu Cho. In defense of lazy visual grounding for open-vocabulary semantic segmentation. In *European Conference on Computer Vision*, pp. 143–164. Springer, 2024.

Chanyoung Kim, Dayun Ju, Woojung Han, Ming-Hsuan Yang, and Seong Jae Hwang. Distilling spectral graph for object-context aware open-vocabulary semantic segmentation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 15033–15042, 2025.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4015–4026, 2023.

Philipp Krähenbühl and Vladlen Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. *Advances in neural information processing systems*, 24, 2011.

Mengcheng Lan, Chaofeng Chen, Yiping Ke, Xinjiang Wang, Litong Feng, and Wayne Zhang. Clearclip: Decomposing clip representations for dense vision-language inference. In *European Conference on Computer Vision*, pp. 143–160. Springer, 2024a.

Mengcheng Lan, Chaofeng Chen, Yiping Ke, Xinjiang Wang, Litong Feng, and Wayne Zhang. Proxyclip: Proxy attention improves clip for open-vocabulary segmentation. In *European Conference on Computer Vision*, pp. 70–88. Springer, 2024b.

Minhyeok Lee, Suhwan Cho, Jungho Lee, Sunghun Yang, Heeseung Choi, Ig-Jae Kim, and Sangyoun Lee. Effective sam combination for open-vocabulary semantic segmentation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 26081–26090, 2025.

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

Feng Liang, Bichen Wu, Xiaoliang Dai, Kunpeng Li, Yinan Zhao, Hang Zhang, Peizhao Zhang, Peter Vajda, and Diana Marculescu. Open-vocabulary semantic segmentation with mask-adapted clip. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 7061–7070, 2023.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European conference on computer vision*, pp. 38–55. Springer, 2024.

Roozbeh Mottaghi, Xianjie Chen, Xiaobai Liu, Nam-Gyu Cho, Seong-Whan Lee, Sanja Fidler, Raquel Urtasun, and Alan Yuille. The role of context for object detection and semantic segmentation in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 891–898, 2014.

Lu Qi, Jason Kuen, Tiancheng Shen, Jiuxiang Gu, Wenbo Li, Weidong Guo, Jiaya Jia, Zhe Lin, and Ming-Hsuan Yang. High quality entity segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL `https://qwen.ai/blog?id=qwen3.5`.

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

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

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024.

Carsten Rother, Vladimir Kolmogorov, and Andrew Blake. " grabcut" interactive foreground extraction using iterated graph cuts. *ACM transactions on graphics (TOG)*, 23(3):309–314, 2004.

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

Yuheng Shi, Minjing Dong, and Chang Xu. Harnessing vision foundation models for high-performance, training-free open vocabulary segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 23487–23497, 2025.

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

Shuyang Sun, Runjia Li, Philip Torr, Xiuye Gu, and Siyang Li. Clip as rnn: Segment countless visual concepts without training endeavor. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13171–13182, 2024.

Feng Wang, Jieru Mei, and Alan Yuille. Sclip: Rethinking self-attention for dense vision-language inference. In *European conference on computer vision*, pp. 315–332. Springer, 2024.

Jing Wang, Huimin Shi, Quan Zhou, Qibo Liu, Suofei Zhang, and Huimin Lu. Dcp-clip: A coarse-to-fine framework for open-vocabulary semantic segmentation with dual interaction. *arXiv preprint arXiv:2603.13951*, 2026.

Ji-Jia Wu, Andy Chia-Hao Chang, Chieh-Yu Chuang, Chun-Pei Chen, Yu-Lun Liu, Min-Hung Chen, Hou-Ning Hu, Yung-Yu Chuang, and Yen-Yu Lin. Image-text co-decomposition for text-supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26794–26803, 2024.

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

Yongqin Xian, Subhabrata Choudhury, Yang He, Bernt Schiele, and Zeynep Akata. Semantic projection network for zero-and few-label semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8256–8265, 2019.

Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in neural information processing systems*, 34:12077–12090, 2021.

Jiarui Xu, Shalini De Mello, Sifei Liu, Wonmin Byeon, Thomas Breuel, Jan Kautz, and Xiaolong Wang. Groupvit: Semantic segmentation emerges from text supervision. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 18134–18144, 2022.

Jiarui Xu, Sifei Liu, Arash Vahdat, Wonmin Byeon, Xiaolong Wang, and Shalini De Mello. Open-vocabulary panoptic segmentation with text-to-image diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2955–2966, 2023a.

Mengde Xu, Zheng Zhang, Fangyun Wei, Han Hu, and Xiang Bai. Side adapter network for open-vocabulary semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2945–2954, 2023b.

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

Dengke Zhang, Fagui Liu, and Quan Tang. Corrclip: Reconstructing patch correlations in clip for open-vocabulary semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 24677–24687, 2025.

Hang Zhao, Xavier Puig, Bolei Zhou, Sanja Fidler, and Antonio Torralba. Open vocabulary scene parsing. In *Proceedings of the IEEE international conference on computer vision*, pp. 2002–2010, 2017.

Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel Codella, Liunian Harold Li, Luowei Zhou, Xiyang Dai, Lu Yuan, Yin Li, et al. Regionclip: Region-based language-image pre-training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16793–16803, 2022.

Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ade20k dataset. *International journal of computer vision*, 127(3):302–321, 2019.

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

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

Xueyan Zou, Zi-Yi Dou, Jianwei Yang, Zhe Gan, Linjie Li, Chunyuan Li, Xiyang Dai, Harkirat Behl, Jianfeng Wang, Lu Yuan, et al. Generalized decoding for pixel, image, and language. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 15116–15127, 2023a.

Xueyan Zou, Jianwei Yang, Hao Zhang, Feng Li, Linjie Li, Jianfeng Wang, Lijuan Wang, Jianfeng Gao, and Yong Jae Lee. Segment everything everywhere all at once. *Advances in neural information processing systems*, 36:19769–19782, 2023b.

SEER: Semantic Evidence for Executable Region Propagation in Open-Vocabulary Semantic Segmentation Supplementary Materials

A Supplementary Overview

The main paper reports the compact experimental chain supporting the central claim that SEER converts closed-set regional evidence into executable OVSS label propagation. The supplementary material gives the extended diagnostics behind that chain and uses the same terminology as the main text. *Coherent region propagation* applies a certified label to the full proposal support. *Boundary-conservative propagation* restricts the support to score- and margin-supported pixels. *Adaptive propagation* chooses between these regimes based on evidence form and predicted risk.

B Theory and Algorithmic Summary

This section gives the formal reading behind the main text. It does not claim a universal performance guarantee; instead, it shows why a VLM-assisted OVSS repair action naturally decomposes into a label-space direction and a pixel-space projector.

Notation. Let Ω be the pixel lattice, $n = |\Omega|$, and $\mathcal{C} = \{1, \dots, C\}$ be the active OVSS label space. At repair step t, the prediction map is $Y^t \in \mathcal{C}^n$ and its one-hot matrix is $Z^t \in \{0,1\}^{n \times C}$, where the p-th row is $z_p^t = e_{Y_p^t}$. For analysis only, let $Y^\star$ be the ground-truth map, $g_p = Y_p^\star$, and $Z^\star$ its one-hot matrix. A local repair action is

$$r = (\ell, W), \qquad \ell \in \mathcal{C}, \quad W \subseteq \Omega, \tag{15}$$

where ℓ is the target label and W is the set of pixels to be changed. Let $\mathbf{1} \in \mathbb{R}^n$ be the all-one vector and let $P_W = \mathrm{diag}(\mathbf{1}_W)$ be the diagonal projector onto pixels in W. Applying r is the linear update

$$Z^{t+1} = T_{\ell,W}(Z^t) = Z^t + P_W(\mathbf{1}e_\ell^\top - Z^t). \tag{16}$$

Thus e_ℓ is a direction in label space, while P_W is a projector in pixel space. This separation is the algebraic object that SEER operationalizes.

Proposition B.1 (Exact benefit–risk decomposition)**.** *Define the pixel-error loss $L(Z) = n - \langle Z, Z^\star \rangle_F$. For any action $r = (\ell, W)$,*

$$\Delta L(r) := L(T_{\ell,W}(Z^t)) - L(Z^t) = \mathcal{H}(\ell, W) - \mathcal{B}(\ell, W), \tag{17}$$

where

$$\mathcal{H}(\ell, W) = \sum_{p \in W} \mathbf{1}[Y_p^t = g_p]\mathbf{1}[\ell \neq g_p], \qquad \mathcal{B}(\ell, W) = \sum_{p \in W} \mathbf{1}[Y_p^t \neq g_p]\mathbf{1}[\ell = g_p]. \tag{18}$$

Here $\mathcal{H}$ counts pixels harmed by the action and $\mathcal{B}$ counts pixels repaired by the action.

Proof. Starting from the definition of L, the change in error is

$$\Delta L(r) = \big(n - \langle Z^{t+1}, Z^\star \rangle_F\big) - \big(n - \langle Z^t, Z^\star \rangle_F\big) \tag{19}$$

$$= -\big\langle Z^{t+1} - Z^t, Z^\star \big\rangle_F \tag{20}$$

$$= -\big\langle P_W(\mathbf{1}e_\ell^\top - Z^t), Z^\star \big\rangle_F \tag{21}$$

$$= \big\langle P_W(Z^t - \mathbf{1}e_\ell^\top), Z^\star \big\rangle_F . \tag{22}$$

Because P_W is diagonal, only rows in W contribute:

$$\Delta L(r) = \sum_{p \in W} \big\langle z_p^t - e_\ell, e_{g_p} \big\rangle \tag{23}$$

$$= \sum_{p \in W} \big(\mathbf{1}[Y_p^t = g_p] - \mathbf{1}[\ell = g_p]\big). \tag{24}$$

For a selected pixel p, the summand is $+1$ exactly when $Y_p^t = g_p$ and $\ell \neq g_p$, i.e., a correct pixel is overwritten by a wrong label; it is -1 exactly when $Y_p^t \neq g_p$ and $\ell = g_p$, i.e., an incorrect pixel is repaired; otherwise it is 0. Grouping the positive and negative terms gives Eqs. 17–18. □

Proposition B.1 explains why regional recognition alone cannot determine the value of a repair. Even if the label direction e_ℓ is semantically correct for part of a proposal, the action can still increase error when the projector P_W covers already-correct pixels. The central quantity is the net benefit produced by the pair (e_ℓ, P_W).

Lemma B.2 (Regional evidence alone cannot identify the support projector)**.** *Let an evidence-only executor be any deterministic rule G that receives only a proposal M and a regional evidence tuple a, and returns a support $W = G(M, a) \subseteq M$. Suppose a certifies the same target label ℓ for two local configurations but does not include dense OVSS scores, margins, or pixel ownership information. Then the support is not identifiable from (M, a): there exist two configurations consistent with the same evidence for which no single W minimizes ΔL in both configurations.*

Proof. Assume, for contradiction, that the support projector is identifiable from (M, a), i.e., there exists a rule G whose support is optimal for every configuration consistent with the same regional evidence. Choose a boundary pixel $p \in M$ whose regional evidence supports label ℓ, and keep all pixels except p identical across two configurations. Construct configuration A so that p is already correct and should not be edited:

$$Y_p^{t,A} = g_p^A \neq \ell. \tag{25}$$

Construct configuration B so that the same pixel is a residual error and should be changed to ℓ:

$$Y_p^{t,B} \neq g_p^B = \ell. \tag{26}$$

Both configurations share the same proposal M and the same regional evidence tuple a, because a only states that the region provides evidence for label ℓ; it does not encode whether the boundary pixel p is already correct or belongs to the repairable part. Therefore G must make the same inclusion decision for p in both configurations.

If $p \in G(M, a)$, then by Eq. 18 this pixel contributes $+1$ to ΔL in configuration A, while excluding it would contribute 0 and strictly improve the action. Hence G is not optimal in A. If $p \notin G(M, a)$, then this pixel contributes 0 in configuration B, while including it would contribute -1 and strictly improve the action. Hence G is not optimal in B. Both cases contradict uniform optimality. Thus regional evidence alone cannot identify the support projector. □

Corollary B.3 (Evidence-to-execution factorization)**.** *A repair interface that aims to reduce ΔL should use different information for the two factors of an action:*

$$\underbrace{e_\ell}_{\textit{label direction}} \leftarrow f_{\mathrm{vlm}}(I, M_i, \mathcal{A}_i), \qquad \underbrace{P_{W_i}}_{\textit{support projector}} \leftarrow \Gamma(S, Y^0, M_i, \ell, a_i). \tag{27}$$

This is the formal motivation for SEER. The VLM verifies the semantic direction, while the OVSS score field and propagation policy determine the executable projector.

Closed candidates as a label-space projection. For proposal M_i, let b_i be the dominant current label and let $\mathcal{Q}_i$ be the score-derived alternatives. The token `uncertain` represents abstention and is excluded from the class set. We define the executable candidate set

$$\mathcal{A}_i^+ = \{b_i\} \cup \mathcal{Q}_i \subseteq \mathcal{C}. \tag{28}$$

It induces a coordinate subspace and projection operator

$$\mathcal{U}_i = \mathrm{span}\{e_c : c \in \mathcal{A}_i^+\} \subseteq \mathbb{R}^C, \qquad \Pi_{\mathcal{U}_i} v = \sum_{c \in \mathcal{A}_i^+} \langle v, e_c \rangle e_c. \tag{29}$$

If $v_i \in \mathbb{R}^C$ denotes the verifier's latent regional compatibility over the active vocabulary, a full-vocabulary verifier searches $\arg\max_{c \in \mathcal{C}} v_{i,c}$, while SEER searches only within the projected subspace:

$$\hat{\ell}_i \in \arg\max_{c \in \mathcal{A}_i^+} \langle \Pi_{\mathcal{U}_i} v_i, e_c \rangle, \quad \text{or abstains with } \texttt{uncertain}. \tag{30}$$

This makes the closed candidate interface more than prompt shortening. It reduces the executable action directions from C to $|\mathcal{A}_i^+|$, and every non-abstained direction is both legal in the active label space and supported by the OVSS score tensor. The evidence-interface ablations in Tables 3 and 12 empirically support this projection view.

Propagation as constrained projector selection. For a certified label ℓ, define the score-feasible support family

$$\mathfrak{P}_i(\ell; S, Y^0) = \{P_W : W \subseteq M_i, \rho(p, \ell, b_i) = 1 \text{ for all } p \in W\}, \tag{31}$$

where ρ is the score-support predicate in Eq. 14. If $Z^\star$ were known, the best support within this family would solve

$$P_i^\star(\ell) \in \arg\min_{P_W \in \mathfrak{P}_i(\ell; S, Y^0)} \left\langle P_W(Z^t - \mathbf{1}e_\ell^\top), Z^\star \right\rangle_F . \tag{32}$$

Since $Z^\star$ is unavailable at inference time, SEER uses S, margins, evidence form, and risk predicates as an executable surrogate for selecting P_W. Coherent propagation chooses a larger projector after strong certification, while score-grounded and boundary-conservative propagation choose smaller projectors when score support is ambiguous. The resulting action is therefore a constrained selector in the OVSS score space.

Algorithm 1 SEER inference. Regional evidence is certified before pixel-level propagation.

Require: image I, fixed prediction Y^0, score tensor S, label space $\mathcal{C}$, SAM2 proposal bank $\mathcal{M}$, audit budget K
Ensure: repaired prediction Y

$Y \leftarrow Y^0$
$\mathcal{R}(I) \leftarrow \mathrm{TopK}_{M_i \in \mathcal{M}}\, \pi(Y^0, S, M_i)$
for each $M_i \in \mathcal{R}(I)$ **do**
  $b_i \leftarrow \mathrm{mode}_{p \in M_i}\, Y_p$
  $\mathcal{A}_i \leftarrow \{b_i\} \cup \mathcal{Q}_i(S, Y, M_i) \cup \{\texttt{uncertain}\}$
  $a_i = (\hat{\ell}_i, \mathrm{conf}_i, \mathrm{form}_i) \leftarrow f_{\mathrm{vlm}}(I, M_i, \mathcal{A}_i)$
  if a_i is not executable under Eq. 11 **then**
    continue
  end if
  $\ell_i \leftarrow \hat{\ell}_i$
  $W_i \leftarrow \Gamma(S, Y, M_i, \ell_i, a_i)$
  $Y_p \leftarrow \ell_i$ for all $p \in W_i$
end for
return Y

C Implementation Details

Unified protocol. The protocol matrix in Table 9 records which factor changes in each study. The default state uses a fixed SAM2 proposal bank, closed top-3 candidates with `uncertain`, local Qwen3.5-9B deterministic decoding, the three-view prompt, and the default SEER propagation policy. The same coefficients and thresholds are used for all benchmarks without dataset-specific tuning. Closed top-k denotes the current/base label plus up to k non-base score-derived alternatives; `uncertain` is an abstention output and is not counted as a semantic candidate.

Table 9: **Unified protocol.** Each row lists the experimental scope, the factor varied, and the quantities kept identical.

Study	Scope	Varied factor	Held fixed
Main	Full validation	Backbone; benchmark	Default verifier, prompt, candidates, policy
Components	Four benchmarks	Module ablation	Prediction, proposals, candidates, parser, split
Interface	Default audit pool	Response format	Prediction, proposals, executable gate, policy
Views	Default audit pool	Visual context	Candidates, verifier, executable gate, policy
Cross-base	Full pipeline	OVSS base	Same coefficients, thresholds, verifier, and rules
Budget	Ranked prefixes	Audit budget	Cached evidence, executable gate, policy

Table 10: **Deterministic execution contract.** Fixed constants and routing rules are shared by all benchmarks; invalid or abstained audits never modify pixels.

Component	Fixed rule/value	Role in execution
Priority score	$\pi_i = 1.20\log(1+100\alpha_i) + q_i + 0.75\max(\delta_i, -0.35) + 0.45\lambda_i + 0.20\eta_i - 0.90r_i$	Ranks proposals before sparse VLM audit; weights are fixed once and not benchmark-fitted.
Risk predicate	Object-like candidate/base, $r_i > 0.20$, or $\alpha_i < 0.015$	Blocks broad writes for risky, small, or boundary-sensitive proposals.
Score support	$(\tau_s, \tau_m, \tau_l) = (0.02, 0.20, 0.25)$ in Eq. 14	Chooses score-compatible pixels for conservative propagation.
Propagation route	Broad support for coherent low-risk evidence; conservative support for object, boundary, small, or risky evidence	Keeps boundary evidence executable but restricts writes to selected support.
Parser and gate	Invalid JSON, out-of-set label, `uncertain`, base label, low/medium confidence, `mixed`, or `unclear`	Abstains with no pixel update; deterministic parser, no retry.
VLM audit	Local Qwen3.5-9B; three-view prompt; constrained JSON; deterministic decoding; 96-token cap	One audited proposal gives one sparse VLM call; no closed oracle model is used.
View builder	Full image width 640; local crop fits 448^2; tight crop fits 336^2	Uses the same global, neighborhood, and tight views for every audit.
SAM2 proposals	Hiera-L fp16; points 8; IoU/stability 0.4; multimask off; crop layers 0; NMS 0.7	Supplies spatial domains only, never semantic predictions.
Audit budget	2,000 unique audits per benchmark	K counts VLM calls, not images or labels.

Table 11: **Proposal-source diagnostic.** The main protocol uses SAM2 to align with the CorrCLIP-style configuration; this diagnostic varies only the mask generator that supplies spatial domains.

Mask generator	VOC21	VOC20	PC59	PC60	City	ADE	Stuff	Object	Avg.
Mask2Former Cheng et al. (2022)	78.5	93.5	54.8	47.6	53.9	36.7	40.1	54.6	57.5
SAM2 Ravi et al. (2024)	79.6	94.2	55.2	48.2	56.9	38.3	38.6	52.9	58.0
EntitySeg Qi et al. (2023)	81.1	94.8	57.8	49.7	**58.1**	**40.6**	39.6	54.2	59.5
SAM3 Carion et al. (2025)	**82.3**	**95.6**	**58.6**	**50.5**	57.9	40.0	**40.4**	**56.5**	**60.2**

SEER uses prediction-side quantities at inference time. The SAM2 proposal bank is precomputed and fixed in the main protocol to match the CorrCLIP-style evaluation setup; Table 11 varies only

this spatial-domain source. Proposals define candidate supports for auditing and propagation, while semantic labels come from the OVSS score tensor and closed-set regional evidence. Fixing SAM2 keeps the protocol aligned and isolates the repair layer itself.

This diagnostic supports two points. First, Mask2Former, EntitySeg, and SAM3 all keep the pipeline functional when they provide reasonable spatial support. Second, stronger proposal coverage can lift the ceiling because proposals define the largest regions that can be audited and certified. The main experiments fix SAM2 for protocol alignment. SAM3 is strongest here, while its prompt-conditioned masks make it a diagnostic alternative outside the default protocol.

In the priority score, $\alpha_i = |M_i|/|\Omega|$, q_i is the dominant-label fraction, δ_i is the candidate-minus-base score gap, λ_i is the low-margin fraction, η_i is the mean entropy signal when available, and r_i is the fixed risk-label fraction. For the view builder, $d = \max(\text{box width}, \text{box height})$. The full view resizes the highlighted image to width 640, the local crop expands the proposal box by $\max(32\text{px}, 0.3d)$ before fitting it inside 448×448, and the tight crop expands by $\max(6\text{px}, 0.3d)$ before fitting it inside 336×336. The object-like routing set includes people/person, vehicles, animals, common furniture, and small handled objects when those names appear in the active dataset vocabulary. The structural risk-ratio set is fixed to common boundary-sensitive labels including background, wall, ceiling, floor, ground, road, building, tree, grass, sky, water, mountain, sidewalk, platform, rock, snow, wood, fence, door, window, curtain, light, sign, track, and shelves. These lists only affect routing and priority; they do not add candidate labels.

For a proposal M_i, the priority score combines four monotone signals: (i) the purity of the dominant baseline prediction within the proposal, (ii) proposal area normalized by image area, (iii) competition between the strongest non-baseline score and the baseline score, and (iv) the fraction of pixels predicted as risk-sensitive classes. This priority serves as a budget controller. Its coarse weights encode a fixed preference for impact, purity, score competition, uncertainty, and risk avoidance; they are not fitted per benchmark. It only orders proposals before VLM verification, and Appendix F shows that ranking alone is insufficient without closed-set evidence and executable propagation.

For score-grounded propagation, Eq. 14 uses three fixed constants. τ_s is a small tolerance that allows the target score to match the current baseline score, τ_m defines low-margin pixels, and τ_l is the wider tolerance allowed only inside low-margin support. We set $(\tau_s, \tau_m, \tau_l) = (0.02, 0.20, 0.25)$ once and keep the values unchanged across benchmarks. These rules form a deterministic execution contract: semantic evidence proposes a legal direction, while score support and risk predicates decide where the direction may write. Boundary-conservative propagation additionally requires low margin and is triggered when score support is weak after semantic certification. The default adaptive policy selects coherent propagation for certified low-risk regions and falls back to score-grounded or boundary-conservative propagation for object-like, boundary-like, large-risk, or high-risk proposals.

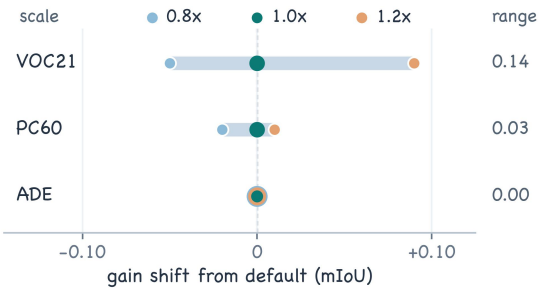

Figure 6: **Support-threshold stability.** Joint scaling around the default constants $(0.02, 0.20, 0.25)$ changes gains by at most 0.14 mIoU.

This support-threshold replay varies fixed constants around their default values. Using the main closed top-3 reference, gains are $+6.72$, $+4.19$, and $+9.67$ on VOC21, PC60, and ADE. Joint scaling shifts the replayed gains by at most 0.14 mIoU; PC60 and ADE move by 0.03 and 0.00. One-factor sweeps over τ_s, τ_m, and τ_l change the four-dataset mean after-mIoU by 0.00007, 0.02093, and 0.01554, respectively; the largest range on any individual dataset is 0.08291. The audit interface, cached VLM evidence, parser, and propagation policy stay fixed in every replay.

VLM decoding. All VLM verifier outputs used by the main mechanism studies are produced with deterministic decoding (`do_sample=False`; greedy decoding with zero or implicit temperature). The replay diagnostics therefore reuse cached verifier outputs exactly as generated, which removes sampling noise from the protocol and keeps the audit sequence reproducible. The threshold stability

diagnostic in Fig. 6 checks the fixed support thresholds around their default setting with joint and one-factor perturbations.

D CLOSED-SET EVIDENCE DIAGNOSTICS

The main paper reports evidence-interface gains using Qwen3.5-9B as the default verifier. The supplementary view adds absolute mIoU in Table 12, execution counts in Table 13, and VLM filtering behavior in Table 15. These diagnostics should be read together because a stronger interface is useful only when it remains executable under the OVSS score field. Boundary-sensitive COCO-Object is analyzed separately in the routing and propagation diagnostics because its stuff-to-object label collapse makes broad regional evidence less directly executable, so the same label can be semantically plausible yet spatially risky.

Table 12: **Evidence interfaces.** Absolute mIoU for the main gain table.

Interface	VOC21	PC60	City	ADE	Avg.
Initial prediction	72.88	44.01	47.41	28.63	48.23
Free-form label	75.70	44.01	47.41	28.63	48.94
Full label list	78.85	44.56	53.43	35.19	53.01
Closed top-1	78.56	45.39	55.36	37.06	54.09
Closed top-2	**79.82**	46.23	54.20	37.56	54.45
Closed top-3	79.60	**48.20**	**56.90**	**38.30**	**55.75**

Table 13: **Execution statistics.** Closed evidence executes useful regions.

Interface	Gain	Exec.	Ben. px	Harm. px
Free-form label	+0.70	455	3.49M	0.04M
Full label list	+4.78	4,193	160.53M	9.06M
Closed top-1	+5.86	3,700	174.80M	3.56M
Closed top-2	+6.22	3,951	183.11M	4.07M
Closed top-3	**+7.52**	**5,217**	**220.68M**	5.11M

The candidate-size diagnostic in Table 14 compares the default Base + 3 interface with smaller and larger candidate lists. Target-in-set coverage is an analysis-only quantity computed with validation labels; it is unavailable to the verifier and propagation policy at inference time. All outcome rates use audited proposals as the denominator. Adding candidates changes both the verifier routing and abstention pattern, so correct outputs need not increase monotonically with k. The best gain comes from the balance between target coverage, repair outputs, execution filtering, and conservative exits. Top-3 increases repair and correct outputs while allowing more abstain/filter exits than Top-2, which is intended under richer candidate context.

Table 14: **Candidate-set size and verifier outcomes.** Rates are averaged over the four mechanism benchmarks with 2,000 audited proposals each. Base + k alt. denotes the current/base label plus up to k non-base score-derived alternatives; `uncertain` is an abstention output. Target-in-set coverage is diagnosis-only. Repair and correct-output rates are normalized by all audited proposals; executed precision is reported separately in Table 15.

Interface	Target-in-set cov.	Sem. cand.	Repair output	Correct output	Base/no edit	Abstain/filter	Avg. gain
Base + 1 alt.	56.3%	2.00	41.3%	40.7%	50.1%	8.6%	+5.86
Base + 2 alt.	85.2%	2.97	47.6%	39.4%	46.3%	6.1%	+6.22
Base + 3 alt.	**99.9%**	**3.35**	**58.8%**	**46.4%**	**28.7%**	12.5%	**+7.52**

Table 15: **Execution filtering of VLM evidence.** Candidate, filtered, and invalid/OOS rates use audited proposals as the denominator; raw accuracy and executed precision are measured on their corresponding pre- and post-filter subsets.

Benchmark	Cand. present	Raw high-conf. acc.	Executed precision	Filtered/no exec.	Invalid/OOS
VOC21	100.0%	62.4%	81.9%	38.9%	0.2%
PC60	99.5%	67.7%	92.7%	35.2%	0.4%
City	100.0%	48.5%	74.3%	36.2%	0.1%
ADE	100.0%	53.1%	80.2%	34.5%	0.1%
Average	99.9%	57.9%	82.3%	36.2%	0.2%

These diagnostics support treating the VLM as a regional evidence source that still requires execution filtering. High-confidence answers alone have limited calibration. The parser, candidate-set validity checks, abstention path, and execution gate substantially improve the precision of decisions that are allowed to affect pixels. Table 15 measures audited proposals before and after filtering, while Table 32 summarizes routed proposals; the denominators are therefore different.

E Propagation Policy Diagnostics

The propagation replay is extended to all eight benchmarks in Table 16 and to pixel-level exposure in Table 17. These endpoint policies show why propagation choice matters, not that one support type is universally superior. Coherent region propagation is strong when verified regions correspond to semantic units; boundary-conservative propagation is useful when object boundaries or proposal noise make broad propagation risky. The adaptive SEER policy is evaluated separately in Table 28, where routing among strict, lenient, and default support gives the strongest overall trade-off.

Table 16: **Full propagation-policy replay over eight benchmarks.** Gains are measured over the fixed prediction. Beneficial/harmful pixels are diagnostic statistics computed after evaluation.

	VOC20	VOC21	PC59	PC60	City	ADE	Stuff	Object	Avg.
Initial prediction	87.62	72.88	48.68	44.01	47.41	28.63	33.38	46.19	51.10
Coherent mIoU	91.92	77.38	53.18	46.91	53.61	35.03	36.38	49.19	55.45
Coherent gain	+4.30	+4.50	+4.50	+2.90	+6.20	+6.40	+3.00	+3.00	+4.35
Boundary mIoU	90.72	76.28	51.98	46.31	51.91	33.83	35.68	48.59	54.41
Boundary gain	+3.10	+3.40	+3.30	+2.30	+4.50	+5.20	+2.30	+2.40	+3.31

Table 17: **Pixel exposure of propagation policies.** Harmful pixels are shown in millions; full mIoU gains, including COCO-Object, are reported in Table 16.

	VOC20	VOC21	PC59	PC60	City	ADE	Stuff
Coherent harmful px	0.02	1.32	1.14	13.85	1.41	0.76	1.18
Boundary harmful px	0.00	1.28	0.85	13.29	0.67	0.47	0.58
Coherent beneficial px	2.32	8.22	50.97	99.13	90.54	48.00	125.65
Boundary beneficial px	1.24	6.48	35.88	71.09	38.63	34.13	84.63

F Region Prioritization Diagnostics

The priority module controls the audit budget. Its coefficients are coarse, fixed ordering preferences for area, purity, score competition, uncertainty, and risk exposure; they are not tuned per benchmark and are not used after a region has been audited. The replay in Table 18 reorders the same audited pool and reuses cached VLM decisions, isolating proposal ordering without new VLM calls. Area is useful, while the full priority is stronger on average. This supports a modest claim: prediction-side ranking exposes a compact auditable set, and executable evidence propagation produces the main gain.

Table 18: **Audited-pool priority replay.** Values are mIoU gains over the fixed prediction when the same audited decisions are replayed under different proposal orderings.

Benchmark	Random	Area-only	Uncertainty-only	Competition-only	Full priority
VOC21	+0.53	+4.55	+1.79	+3.70	**+6.72**
PC60	+0.27	+2.12	+2.00	+3.39	**+4.19**
City	+1.32	+4.73	+1.64	+3.55	**+9.49**
ADE	+1.17	+4.90	+3.42	+4.75	**+9.67**
Average	+0.82	+4.08	+2.21	+3.85	**+7.52**

Candidate-order stability. We replayed 8,000 cached decisions with the original candidate order and three independent shuffles. The corresponding after-mIoU values are 49.4254, 49.4820, 49.4558, and 49.4701. The largest decline on any individual dataset is 0.0778 mIoU. Candidate position therefore has little influence under deterministic decoding and the closed executable gate.

G Executable Gate Diagnostics

The relaxed-gate diagnostic in Table 19 removes the target/legal certification constraint and allows broader executable actions. Relaxing the gate can increase recall, while also expanding the number of writeback actions substantially and making conservative containment harder to preserve. The accompanying increase in routed support supports the gate's role as an executable certification constraint.

Table 19: **Relaxing executable certification.** Removing the target/legal gate increases recall but permits many more writeback actions.

Benchmark	Initial prediction	After relaxed gate	Gain	Executed actions
VOC21	72.88	78.58	+5.70	5,024
PC60	44.01	49.59	+5.58	7,572
City	47.41	56.56	+9.15	1,955
ADE	28.63	38.21	+9.58	1,975
COCO-Object	46.19	50.55	+4.36	1,608

H Control Interfaces and Non-VLM Diagnostics

Protocol-matched controls are compared in Table 20, with pixel statistics in Table 21. The score tensor contains useful repair signal, but score-only rules do not produce regional semantic evidence. A CLIP crop classifier tests ordinary region recognition, score-grounded support removes SEER's adaptive executable propagation, and Dense-CRF-style smoothing tests generic local regularization. All controls remain below SEER under the same fixed-prediction protocol.

Table 20: **Protocol-matched controls.** Same prediction state; only the repair signal or execution rule changes.

Method	VOC21	PC60	City	ADE	Avg.
OVSS score rule	+2.10	+3.80	+2.53	+2.13	+2.64
CLIP crop classifier	+3.88	+2.38	+2.70	+5.80	+3.69
VLM score support	+2.09	+0.87	+1.74	+2.11	+1.70
Dense-CRF smoothing	+0.01	-0.01	+0.02	+0.01	+0.01
SEER	**+6.72**	**+4.19**	**+9.49**	**+9.67**	**+7.52**

Table 21: **Pixel statistics.** Four-benchmark averages over executed repairs.

Control	Exec.	Ben.	Harm	Gain
Cons. score	107.0	1.66M	0.05M	+0.74
Margin score	768.3	18.51M	3.44M	+2.64
Closed evidence	1,304.3	55.17M	1.28M	+7.52

Table 22: **Training-free cost.** Selective proposal-level auditing with effective sharded verifier latency.

Benchmark	Images	Calls/img	VLM s/img	Prop. s/img	Total s/img
VOC21	1,449	1.38	0.58	0.043	0.62
PC60	5,104	0.39	0.16	0.043	0.21
City	500	4.00	1.68	0.043	1.72
ADE	2,000	1.00	0.42	0.043	0.46
COCO-Object	5,000	0.40	0.17	0.043	0.21
Weighted avg.	14,053	0.71	0.30	0.043	0.34

Table 23: **Qwen3.5 resources.** Unsharded local single-call profile.

Scale	s/call	Peak	Weights
0.8B	0.53	2.4 GB	1.63 GiB
4B	1.79	9.4 GB	8.68 GiB
9B	2.09	18.7 GB	17.98 GiB

Table 24: **End-to-end timing.** Validation-weighted cost with the default verifier.

Component	Time / image
Proposal generation	0.508s
Frozen OVSS base forward	5.042s
VLM verifier audit	0.424s
Propagation / writeback	0.043s
Cold end-to-end total	6.018s
SEER repair overhead	0.468s

Table 25: **Cross-base uncertainty.** Bootstrap intervals for the plug-in gains in Table 26.

OVSS base	Δ_{avg}	Bootstrap 95% CI
ProxyCLIP	+4.53	[+4.08, +4.92]
ResCLIP	+5.08	[+4.55, +5.50]
Trident	+5.93	[+5.34, +6.41]
CorrCLIP	+5.48	[+4.86, +5.97]

The cost and resource measurements in Tables 22, 23, and 24 are local reproducibility diagnostics and do not imply real-time inference. The default verifier is a locally deployable open VLM. The 0.42s/call value is the effective latency after sharded execution, while the 2.09s/call resource profile is an unsharded local single-call measurement. Each VLM call audits one proposal using the three-view prompt, and the weighted average is computed over 14,053 validation images. SEER is training-free and does not update OVSS weights; its extra work scales with the number of audited proposals.

I Cross-Base Plug-in Diagnostics

The main paper applies the same complete SEER pipeline to multiple OVSS bases. The full-validation per-benchmark results appear in Table 26, with bootstrap intervals summarized in Table 25. Coefficients, thresholds, candidates, verifier, parser, gates, and propagation rules are unchanged; only the base prediction and score cache come from the chosen OVSS method. This controlled plug-in test complements the full eight-benchmark comparison in Tab. 1.

Table 26: **Full-validation cross-base plug-in test.** The same SEER rules are used for each OVSS base; Δ_{avg} averages four benchmark gains.

OVSS base	Benchmark	Base	+SEER	Δ	Δ_{avg}
ProxyCLIP	VOC21	61.3	64.9	+3.6	+4.53
	PC60	35.3	39.5	+4.2	
	City	38.1	42.6	+4.5	
	ADE	20.2	26.0	+5.8	
ResCLIP	VOC21	61.1	65.2	+4.1	+5.08
	PC60	33.5	38.3	+4.8	
	City	35.9	41.1	+5.2	
	ADE	18.0	24.2	+6.2	
Trident	VOC21	67.1	73.0	+5.9	+5.93
	PC60	38.6	44.0	+5.4	
	City	42.9	48.8	+5.9	
	ADE	21.9	28.4	+6.5	
CorrCLIP	VOC21	74.8	79.3	+4.5	+5.48
	PC60	44.2	47.7	+3.5	
	City	49.4	55.0	+5.6	
	ADE	26.9	35.2	+8.3	

Cross-base reading. This test is not intended to replace the full eight-benchmark comparison. It asks whether the same deterministic execution contract remains effective when the OVSS prediction and score cache change. All four bases improve on all four benchmarks, with average gains from +4.53 to +5.93 mIoU. This pattern indicates that the repaired residues are not idiosyncratic to CorrCLIP; the evidence-to-execution interface supports plug-in behavior when the new base provides a compatible score field.

J Prompt, Policy, and Repair-Space Diagnostics

This section audits the repair mechanism from three angles. Prompt and policy sensitivity are reported in Tables 27 and 28; proposal-space headroom is measured with validation-label oracles in Table 29; residual harm and routing behavior are summarized in the remaining diagnostics.

Table 27: **Prompt schemas.** The default prompt has the best average.

Prompt	VOC21	PC60	ADE	Avg.
Concise	+3.75	+3.26	+2.58	+3.20
Neutral	+5.98	+2.73	+3.49	+4.07
Default	**+6.72**	**+4.19**	**+9.67**	**+6.86**

Table 28: **Policy sensitivity.** SEER gives the largest gains with broader certified support.

Policy	VOC	PC	ADE	Support
Strict	+3.73	+2.47	+7.06	0.66M
Lenient	+4.56	+3.91	+8.23	0.82M
SEER	**+6.72**	**+4.19**	**+9.67**	1.14M

Prompt templates. All prompts use the same legal candidate set $\mathcal{A}_i$ and the same parser; only the amount of visual context and instruction detail changes.

Concise. "A region is highlighted in the image. Candidate labels are $\mathcal{A}_i$. Choose exactly one candidate label or `uncertain`. Return JSON only: {`label`, `confidence`}."

Neutral. "Inspect the highlighted region. The legal labels are $\mathcal{A}_i$; do not use outside labels. Decide whether the region visually matches one candidate or should be `uncertain`. Return JSON only: {`label`, `confidence`, `region_form`}, where `region_form` is one of `coherent`, `boundary`, `mixed`, or `unclear`."

Default. "You are verifying evidence for open-vocabulary semantic segmentation repair. You are shown the global image, a local neighborhood crop, and a tight crop of the same highlighted region. Use global context for scene semantics, local context for surrounding boundaries, and the tight crop for appearance. Candidate labels are $\mathcal{A}_i$ and include `uncertain`. Do not invent labels. Return JSON only: {`label`, `confidence`, `region_form`}. Use `uncertain` if the evidence is weak, mixed, outside the candidate set, or not executable as a coherent region."

Table 29: **Proposal-space repair headroom.** Oracle rows are analysis-only diagnostics, not inference-time baselines. The full-pool oracle is the headroom row; the top-500 oracle is budget-limited and not a strict upper bound.

Benchmark	SEER closed	Oracle top-500	Oracle full pool	SEER/full
VOC21	+6.72	+13.01	+16.38	0.41
PC60	+4.19	+2.33	+6.56	0.64
City	+9.49	+13.92	+21.74	0.44
ADE	+9.67	+6.37	+19.54	0.49
Average	+7.52	+8.91	+16.05	0.50

The oracle rows quantify remaining repair headroom inside the same proposal and action space. The full-pool oracle is the strict headroom diagnostic, while the top-500 oracle preserves the audit budget and can be lower than SEER when the budgeted oracle misses actions found by the default replay. Across the four mechanism benchmarks, SEER recovers about half of the full-pool oracle gain on average. COCO-Object is excluded because broad full-pool writes are not a safe upper bound in boundary-sensitive object settings.

Table 30: **Image/class harm.** Harmed-image rates and losses are measured after SEER repair; negative classes count class-IoU drops.

Benchmark	Images	Harmed img.	Med. loss	P75 loss	Mean gain	Neg. cls.
VOC21	1,449	5.8%	3.40	8.70	+4.70	0
PC60	5,104	0.9%	3.26	8.95	+2.83	0
City	500	5.0%	0.43	1.92	+6.18	0
ADE	2,000	3.8%	0.82	2.47	+2.91	5

Table 31: **Residual harm sources.** Harmful-pixel rates are measured among modified pixels; source shares are normalized only over residual harmful pixels.

Benchmark	Harm/mod.	Support spill	Sem. error
ADE	2.1%	69.8%	30.2%
City	2.1%	72.4%	27.6%
PC60	2.6%	57.2%	42.8%
VOC21	1.4%	58.7%	41.3%

The harm diagnostics address whether average mIoU gains hide systematic damage. Harmed-image rates stay low, no negative classes appear on VOC21, PC60, or City, and ADE has five small negative class deltas, led by stairway -0.57, ball -0.53, and base -0.32. Residual harmful pixels are sparse among modified pixels. Their sources split between support spill, where plausible evidence crosses ambiguous boundaries, and semantic error, where the regional label is wrong. This pattern supports the need for both closed evidence filtering and conservative score-space execution.

Table 32: **Regional-evidence routing taxonomy.** Rates are computed over routed audited proposals. Conservative routing records boundary-aware execution; residual pixel harm is reported in the preceding table.

Benchmark	Uncertain/low	Candidate miss	Correct high-conf.	Conservative route
VOC21	4.8%	0.2%	89.6%	5.4%
PC60	3.7%	0.5%	87.2%	8.6%
City	3.4%	0.3%	89.5%	6.8%
ADE	4.6%	0.8%	84.9%	9.7%
Object	6.9%	0.7%	81.6%	10.8%

Limitations. SEER is a plug-in repair layer over a fixed OVSS prediction, so the complete-system comparison reports the accuracy-cost trade-off of adding frozen proposal-level regional auditing to an OVSS predictor, while matched-input diagnostics isolate repair. Its priority, gate, and routing rules form a deterministic execution contract, fixed across datasets, bases, and auditors to enforce legal labels, high-confidence evidence, and score-compatible support. SEER remains shaped by proposal coverage, candidate recall, regional verifier errors, and latent score support; improving thin structures, mixed proposals, crowded boundaries, and candidate-set misses in boundary-sensitive images is an important direction for future work.

K Visual Examples and Figure Data

Failure and boundary-sensitive cases. Fig. 7 shows residual cases where regional recognition is easier than executable support. Crowded street, food/table, and small-object scenes expose four residual risks: over-confident verifier outputs, candidate-set misses, mixed proposals, and weak score-field support. In these cases, a proposal may cover several semantic parts, or the correct label may be visible only in fragmented score support. SEER reduces their impact through abstention and

conservative routing, which can leave part of a region unchanged when support is unreliable. Such cases also show the risk of direct VLM pixel authority. Larger galleries in Figs. 8, 9, 10, 11, 12, 13, 14, and 15 provide held-out examples.

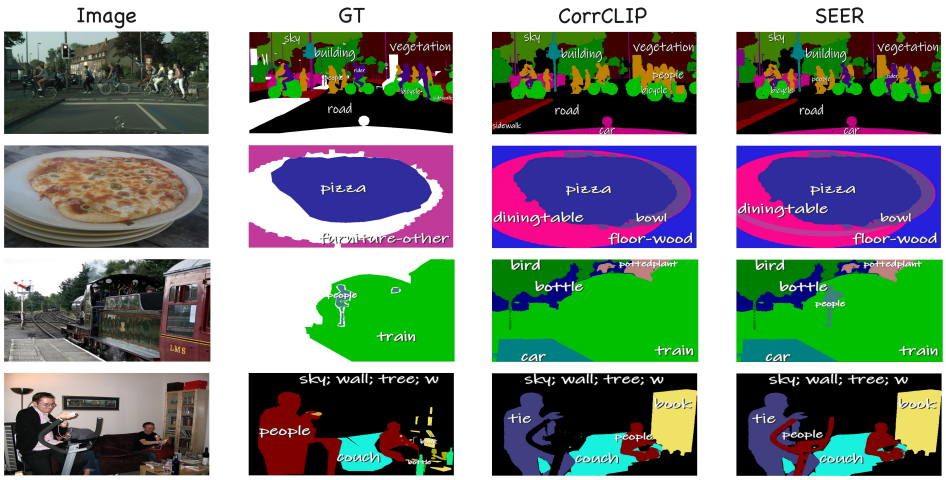

Figure 7: **Boundary-sensitive and residual failure cases.** SEER can recover proposal-supported regions absent from the initial top-1 prediction, but errors remain when proposals cross semantic boundaries, candidate labels are ambiguous, or the OVSS score field is weak. These cases motivate the execution gate and conservative propagation: plausible regional VLM evidence still needs score-field-constrained pixel support.



Figure 8: **ADE20K visual examples.** SEER repairs local semantic residues in indoor and scene-centric images while preserving the surrounding dense prediction structure.



Figure 9: **Cityscapes visual examples.** SEER targets structured road-scene residues such as missing stuff regions, fragmented objects, and label-space confusions in a fixed OVSS prediction.



Figure 10: **PC59 visual examples.** The repairs are localized to semantically supported regions and keep unrelated objects and stuff categories unchanged.

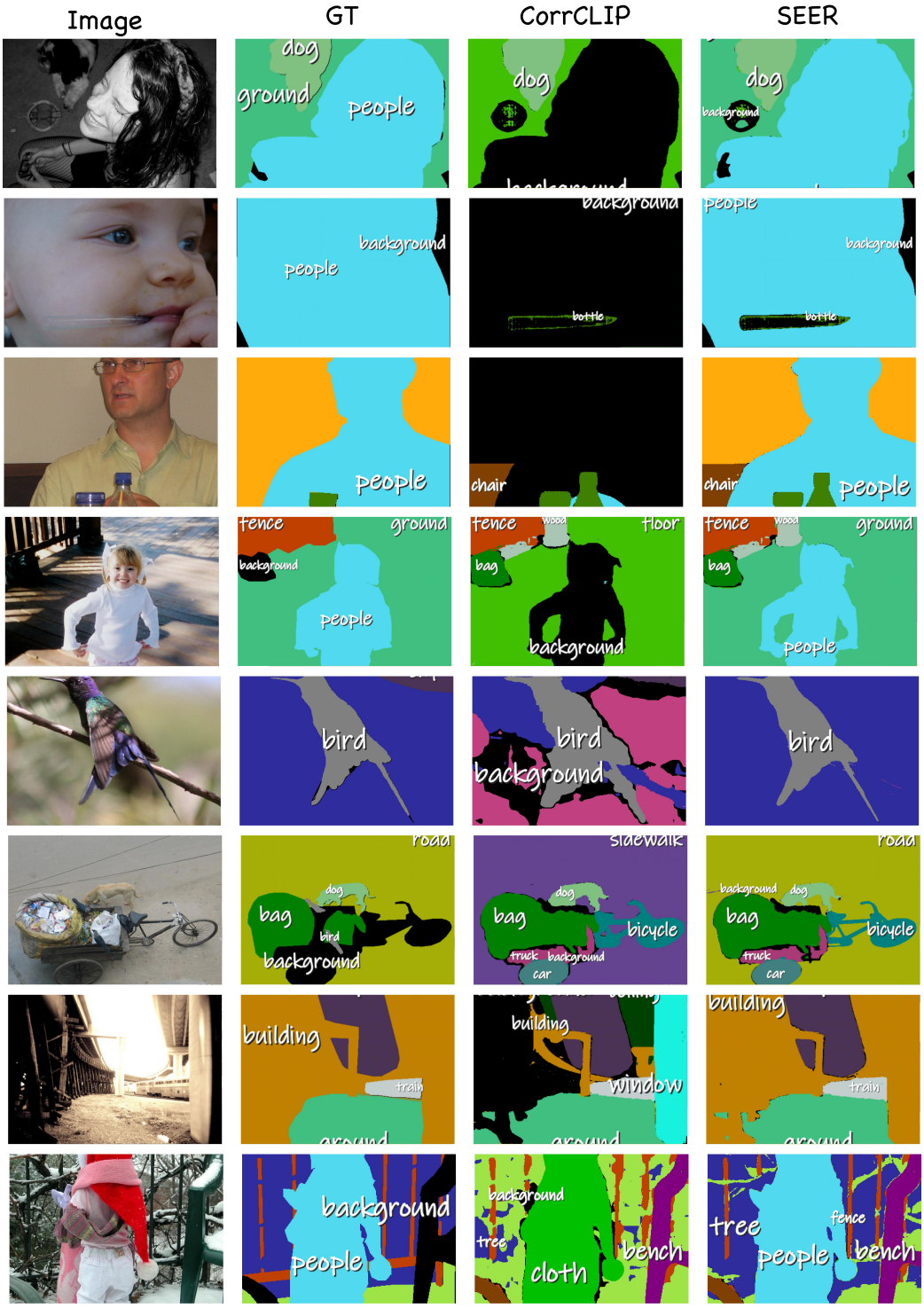

Figure 11: **PC60 visual examples.** The fixed prediction often contains plausible but incomplete regions; SEER uses closed-set evidence to propagate executable labels where the OVSS score field supports them.

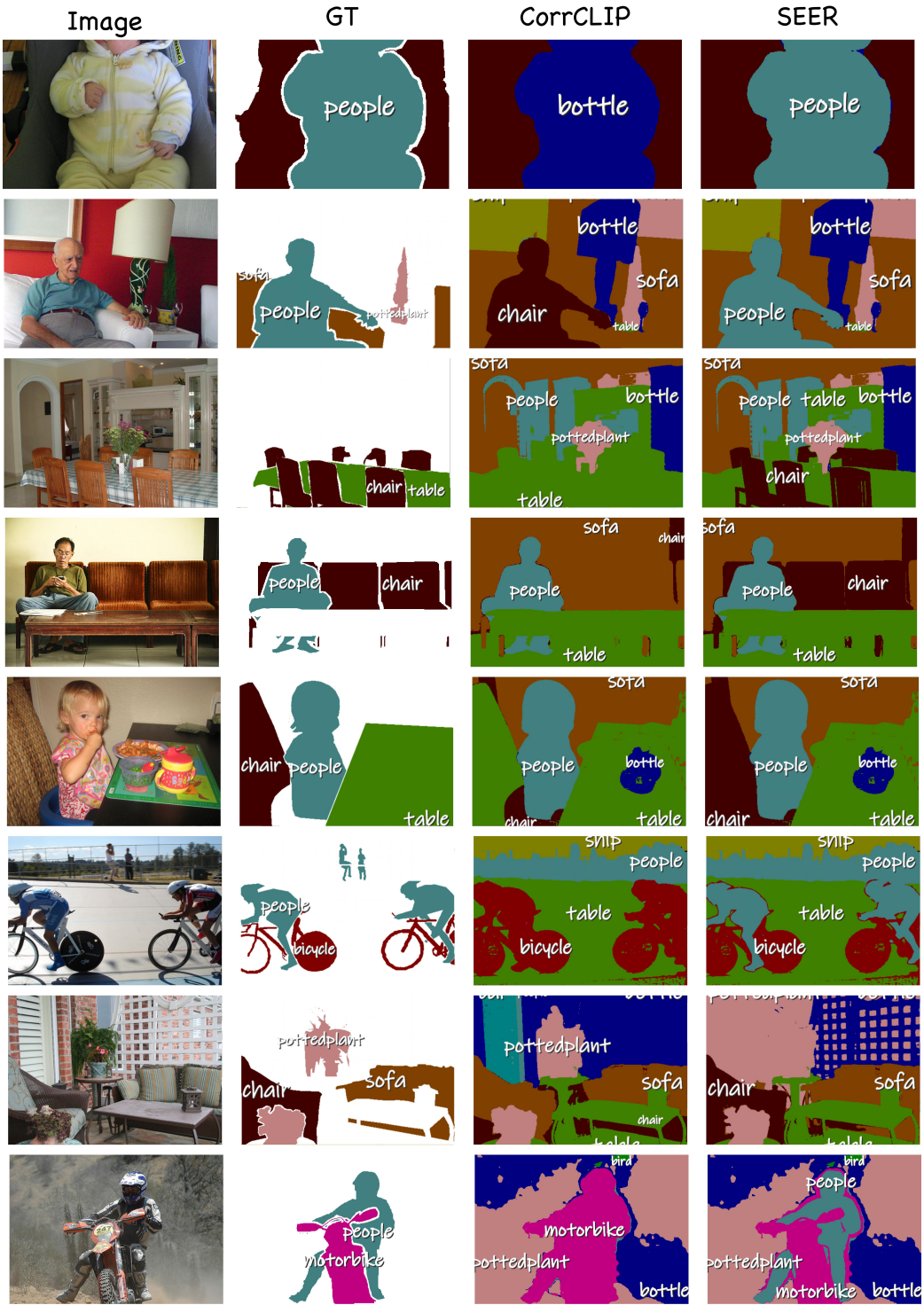

Figure 12: **VOC20 visual examples.** Object-centric cases show that SEER repairs missing or confused object regions through closed-set evidence and score-space execution.

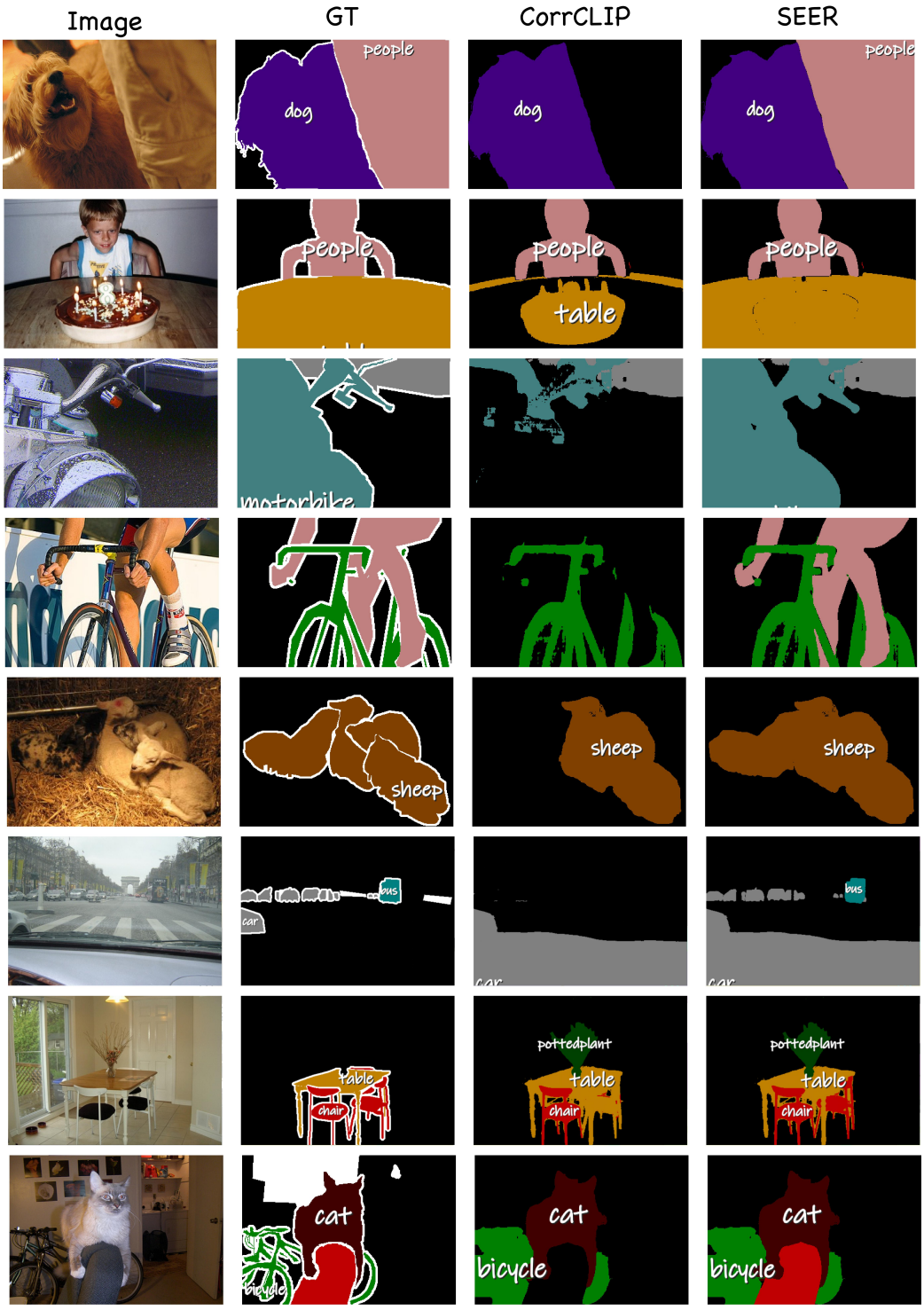

Figure 13: **VOC21 visual examples.** SEER improves residual object and background errors under the active label space, complementing the quantitative gains reported in the main paper.

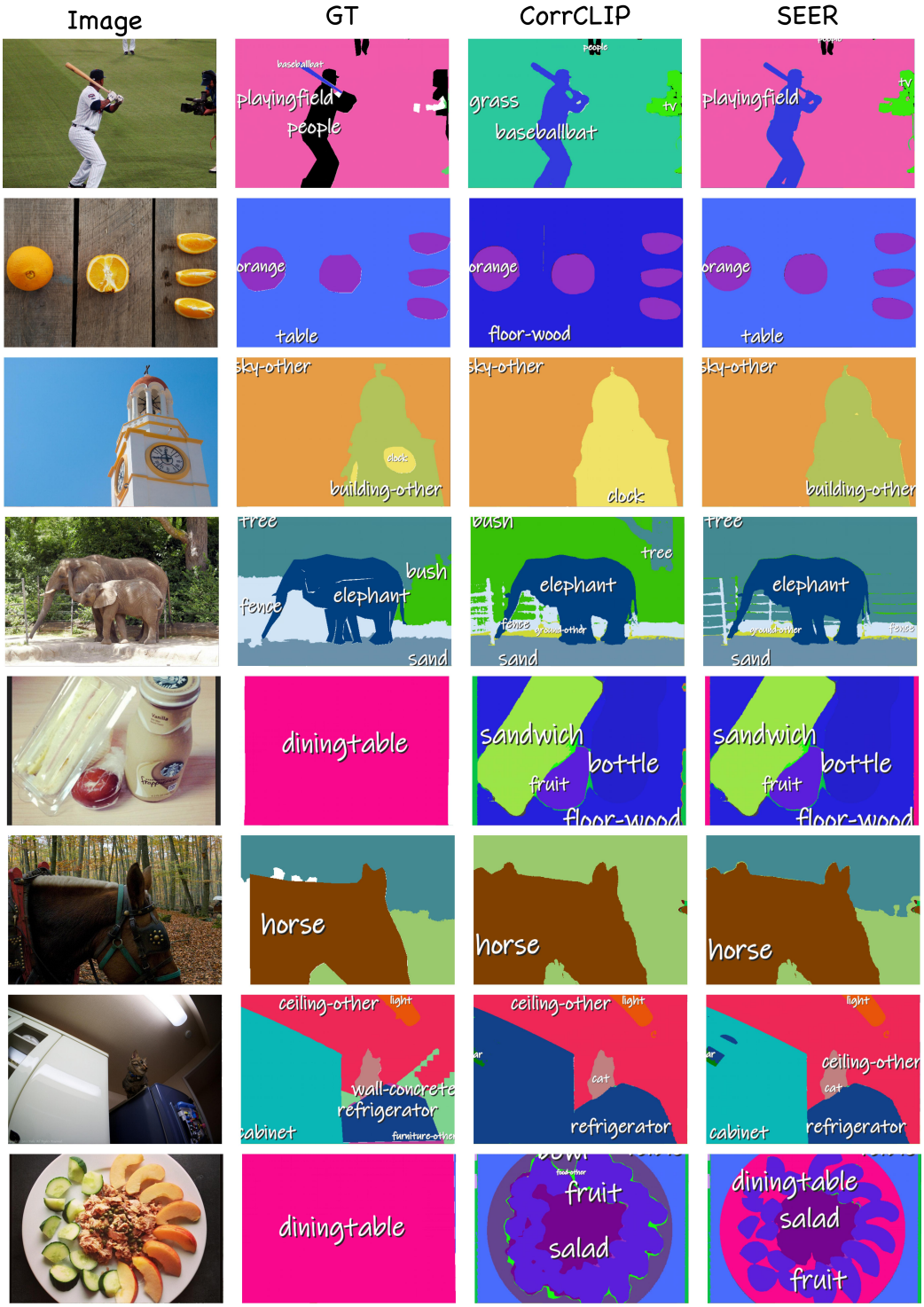

Figure 14: **COCO-Stuff visual examples.** Stuff-heavy scenes highlight the value of label-space-aware evidence: SEER corrects broad semantic regions while retaining pixel-level execution in the OVSS model.

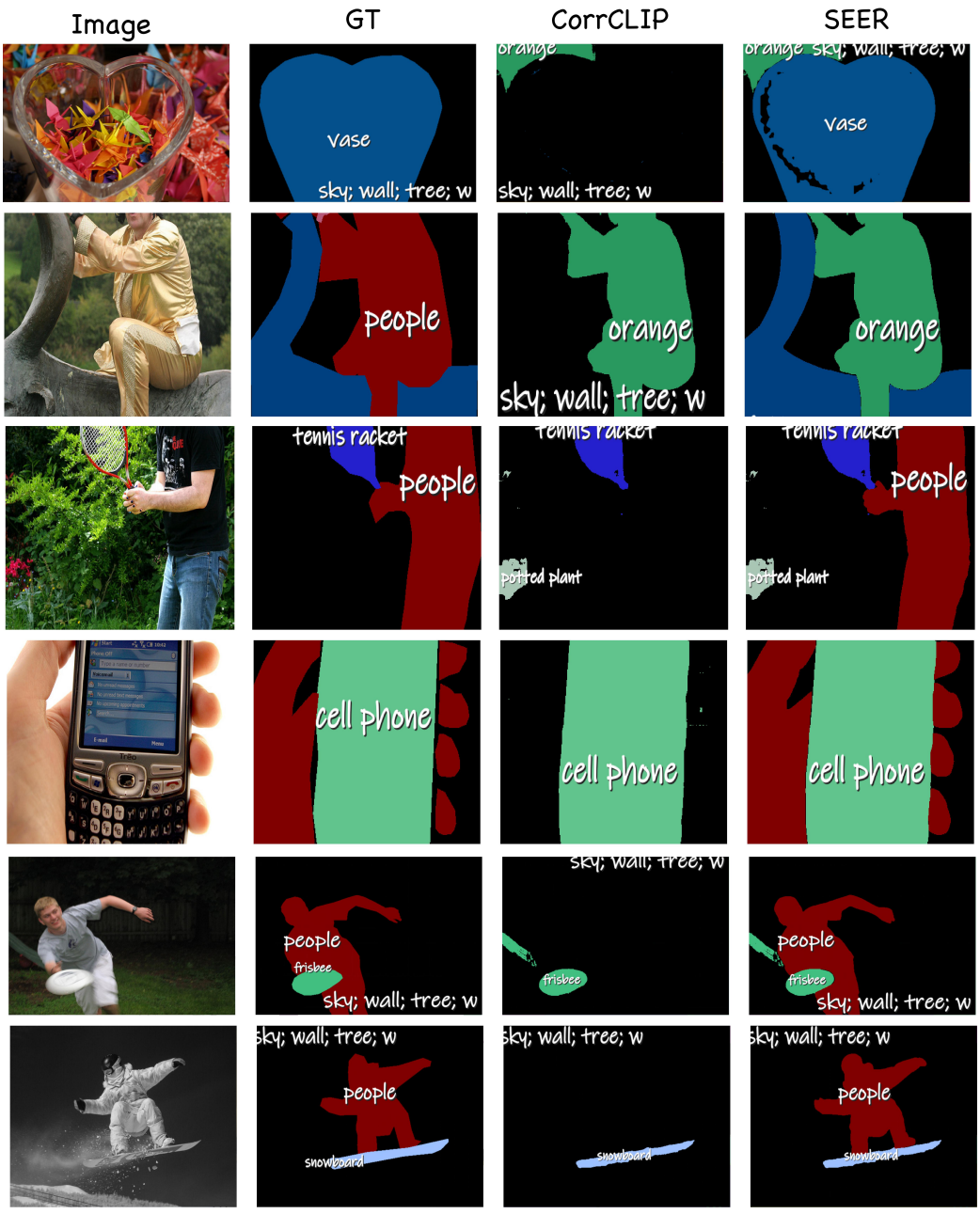

Figure 15: **COCO-Object visual examples.** Object-boundary cases are more conservative, but SEER still repairs reliable regions when the closed candidate and execution gate agree.