BEYOND KNOWN-CLASS PROMPTS: SPARSE FACTOR OBJECTNESS FOR OPEN-WORLD OBJECT DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Open-world object detection (OWOD) requires recognizing known objects while discovering unknown ones. Open-vocabulary detectors (OVDs) provide strong visual representations but rely on class-name prompts for detection. Without unknown-class prompts, unknown objects may be missed even when their region representations retain informative object cues. We propose a **S**parse-**A**uto**E**ncoder-based **O**pen-**V**ocabulary **O**pen-**W**orld framework for OWOD (SAE-OVOW) to recover class-agnostic objectness directly from a frozen OVD without unknown-class annotations. SAE-OVOW uses a sparse autoencoder to decompose known-aligned region embeddings into sparse latent factors. It then selects factors that are reliably activated, exhibit similar distributions across known-aligned and unmatched regions, and have diverse decoder directions. The selected activations are aggregated into sparse factor objectness and combined with known-class uncertainty for unknown detection. Experiments on M-OWODB and S-OWODB show that SAE-OVOW consistently improves U-Recall while maintaining competitive known-class performance. On S-OWODB Task 2, it improves U-Recall by 12.7 points and overall mAP by 5.3 points over the best competing results. Latent removal-and-recovery experiments verify the functional relevance of the selected factors, while Task-1 post-hoc audits on M-OWODB and S-OWODB identify coherent visual patterns in 119 of 200 factors.

1 INTRODUCTION

Open-world object detection (OWOD) aims to recognize objects from known categories while discovering objects beyond the current label space (Joseph et al., 2021; Gupta et al., 2022; Ma et al., 2023). Recent progress in large-scale vision-language and generative models has enabled increasingly flexible visual understanding and generation, ranging from open-vocabulary recognition (Gu et al., 2021; Minderer et al., 2022; Cheng et al., 2024) to pose-guided person synthesis and customizable virtual dressing (Shen & Tang, 2024; Shen et al., 2025b). For OWOD, open-vocabulary detectors (OVDs) are particularly attractive because their region representations may preserve visual information beyond the class vocabulary supplied at inference.

However, OVD inference remains prompt-driven. Typically, only known class names are provided as prompts, forcing candidate regions to compete within the current vocabulary. If no suitable prompt exists, an unknown object may receive low matching scores and be discarded (Li et al., 2024a). Existing methods recover unknown objects through objectness modeling (Zohar et al., 2023b; Zhang et al., 2025), unknown pseudo-label mining (Fang et al., 2025; Wang et al., 2023), or heuristic uncertainty scores (Zhao et al., 2023; Yavuz & Güney, 2024). Although effective, these approaches do not explicitly exploit the fine-grained activation structure of frozen OVD region representations. A classification score reports only the strongest known-prompt match, whereas the internal representation may retain additional object-related information. Consequently, unknown objects and difficult known objects can exhibit similarly low prompt confidence while producing different internal activation patterns.

We therefore seek to recover class-agnostic objectness directly from the internal representations of a frozen OVD. The main challenge is that object-related cues are distributed across entangled feature dimensions, making stable objectness patterns difficult to access directly (Lim et al., 2025; Pach

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

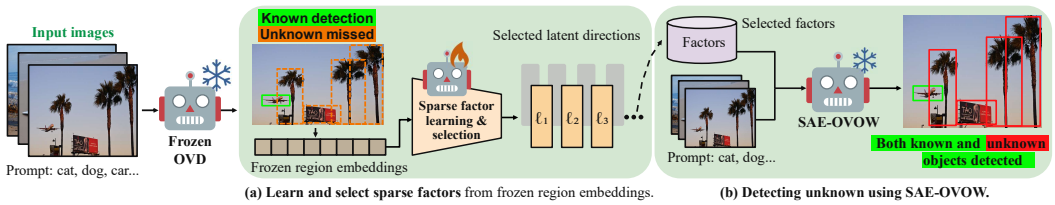


Figure 1: **Overall architecture of SAE-OVOW**, showing how frozen OVD region representations are used to (a) learn and select sparse latent factors that provide class-agnostic objectness and (b) detect unknown objects while preserving known-class prediction.

et al., 2025). As illustrated in Fig. 1, known-class prompts may miss objects outside the current vocabulary even when their region representations retain informative object structure.

To extract this structure, we propose a **S**parse-**A**uto**E**ncoder-based **O**pen-**V**ocabulary **O**pen-**W**orld object detection framework, termed SAE-OVOW. SAE-OVOW first learns an overcomplete sparse dictionary from known-aligned region embeddings. It then filters inactive or weak latent directions, favors directions with similar activation distributions across known-aligned and unmatched regions using Jensen–Shannon divergence, and promotes decoder diversity to reduce redundancy. The selected activations are aggregated into sparse factor objectness and combined with known-class uncertainty for unknown detection. SAE-OVOW requires neither unknown-class annotations nor predefined attribute descriptions, while keeping both the detector and class embeddings fixed.

Our contributions are summarized as follows:

- We introduce *sparse factor objectness* to recover class-agnostic object cues from frozen OVD representations beyond known-prompt scores.
- We propose SAE-OVOW, which learns sparse factors from known-aligned regions and selects reliable, transferable, and non-redundant directions using activation reliability, distribution similarity, and decoder diversity.
- SAE-OVOW achieves the highest U-Recall on M-OWODB and S-OWODB while maintaining competitive known-class performance, including gains of 12.7 U-Recall and 5.3 mAP points on S-OWODB Task 2. Intervention and audit studies further validate the selected factors.

2 Related Work

Open-World Object Detection. Open-world object detection extends closed-set detection by requiring models to recognize known classes while discovering objects outside the current label space (Joseph et al., 2021; Gupta et al., 2022; Ma et al., 2023). Existing methods identify unknown objects through objectness decoupling or feature separation (Sun et al., 2024; Majee et al., 2025; Heo et al., 2026), pseudo-label mining (Fang et al., 2025; Wang et al., 2023), output-level uncertainty (Zhao et al., 2023; Yavuz & Güney, 2024), and foundation-model or concept transfer (Ma et al., 2025; Lv et al., 2026). Most rely on task-specific heads, output scores, or external semantic spaces, leaving the internal activation structure of frozen OVD features largely unexplored. Foundation-model approaches often construct transferable cues from language-defined attributes (Zohar et al., 2023a; Xi et al., 2024; Yang et al., 2025). OW-OVD (Xi et al., 2025b) selects attributes shared across annotated and unannotated regions and combines attribute similarity with known-class uncertainty. OW-VAP (Xi et al., 2025a) instead learns a visual attribute parser to estimate region objectness. Nevertheless, their objectness signals remain organized around attribute vocabularies or predictions. SAE-OVOW directly extracts class-agnostic objectness from sparse latent directions in frozen OVD representations, without predefined attributes or semantic descriptions. Semantic descriptions are used only for post-hoc analysis.

Sparse Autoencoders. Useful visual structure can be introduced through explicit conditions or discovered from internal representations. Explicit conditioning has enabled pose-guided synthesis and consistent story visualization (Shen et al., 2024; 2025d), as well as layout-preserving editing, fine-grained garment generation, and long-term talking-face generation (Shen et al., 2025a; 2026;

2025c). In contrast, sparse autoencoders (SAEs) discover latent factors directly from model activations by decomposing dense representations into sparse coefficients over a learned dictionary. SAEs have been used to analyze superposed features and identify directions with measurable downstream effects (Zhang et al., 2026). In vision-language models, PatchSAE provides spatial attribution for CLIP representations (Lim et al., 2025), while MonoSemanticity Score and latent interventions evaluate semantic selectivity and functional influence (Pach et al., 2025). Top-activation examples further reveal recurring visual patterns (Cunningham et al., 2024; Lim et al., 2025). Unlike prior work that primarily uses SAEs for interpretation or steering, SAE-OVOW converts selected latent directions into an inference-time objectness signal for open-world detection.

3 METHODOLOGY

Fig. 2 provides an overview of SAE-OVOW. A task-specific SAE learns sparse coordinates from frozen OVD region embeddings (Sec. 3.1). Activation reliability, activation distributional similarity, and decoder diversity then select directions with stable within-task activation statistics (Sec. 3.2). Their activations form sparse factor objectness, which is combined with known-class uncertainty for unknown detection (Sec. 3.3). Appendix E provides the corresponding algorithms.

3.1 SPARSE COORDINATE LEARNING

We train an SAE to decompose frozen OVD region representations into a small active set of directions from an overcomplete dictionary.

Let $\mathcal{E}_{\text{vis}}$ be the region embeddings extracted from the frozen OVD. Using known-class annotations Y_I and text embeddings E_C, we define the known-aligned set (Xi et al., 2025b) as

$$E^{+}_{\text{vis}} = \{e_{\text{vis}} \in \mathcal{E}_{\text{vis}} \mid S(E_C, e_{\text{vis}}, Y_I) > \alpha\}, \tag{1}$$

where $S(\cdot)$ measures the region-to-class alignment using both classification and localization evidence (Feng et al., 2021), and α is a matching threshold. The remaining embeddings are collected as

$$E^{-}_{\text{vis}} = \{e_{\text{vis}} \in \mathcal{E}_{\text{vis}} \mid S(E_C, e_{\text{vis}}, Y_I) \leq \alpha\}. \tag{2}$$

E^{-}_{vis} is the unmatched set and may contain both background and objects outside the known label space. The SAE is trained on E^{+}_{vis}. For $\mathbf{z} \in \mathbb{R}^d$, its encoder produces

$$\mathbf{h}(\mathbf{z}) = \text{ReLU}(W_E^{\top}\mathbf{z} + \mathbf{b}_E), \quad \mathbf{h}(\mathbf{z}) \in \mathbb{R}^{d_{\text{SAE}}}, \tag{3}$$

and the decoder reconstructs the original representation as

$$\hat{\mathbf{z}} = W_D\mathbf{h}(\mathbf{z}), \quad W_D \in \mathbb{R}^{d \times d_{\text{SAE}}}. \tag{4}$$

Each unit-normalized column of W_D defines a dictionary direction, and its coordinate gives the corresponding activation strength. Sparsity is induced by ReLU activations and the ℓ_1-regularized objective

$$\mathcal{L}_{\text{SAE}} = \|\hat{\mathbf{z}} - \mathbf{z}\|_2^2 + \lambda_{\text{L1}} \|\mathbf{h}(\mathbf{z})\|_1. \tag{5}$$

3.2 JSD-GUIDED LATENT-DIRECTION SELECTION

Because unknown annotations are unavailable and the unmatched set mixes background with objects outside the known label space, we do not treat it as an unknown-object set. Instead, we select directions that are reliably active in both groups and have similar activation distributions, limiting specialization to the current known classes.

For the i-th latent direction, let a_i denote its encoder pre-activation. We map pre-activations to a common bounded support for distribution estimation:

$$q_i^{+}(e^{+}_{\text{vis}}) = \sigma\big([W_E^{\top}e^{+}_{\text{vis}} + \mathbf{b}_E]_i\big), \quad q_i^{-}(e^{-}_{\text{vis}}) = \sigma\big([W_E^{\top}e^{-}_{\text{vis}} + \mathbf{b}_E]_i\big). \tag{6}$$

Histograms of q_i^{+} and q_i^{-} yield the normalized distributions D_i^{+} and D_i^{-}; activation reliability and objectness estimation use the ReLU activations $d_i^{\pm} = \max(0, a_i^{\pm})$. We quantify the distribution discrepancy using the Jensen–Shannon divergence:

$$\text{JSD}(D_i^{+}, D_i^{-}) = \frac{1}{2}\text{KL}(D_i^{+}\|M_i) + \frac{1}{2}\text{KL}(D_i^{-}\|M_i), \quad M_i = \frac{1}{2}(D_i^{+} + D_i^{-}). \tag{7}$$

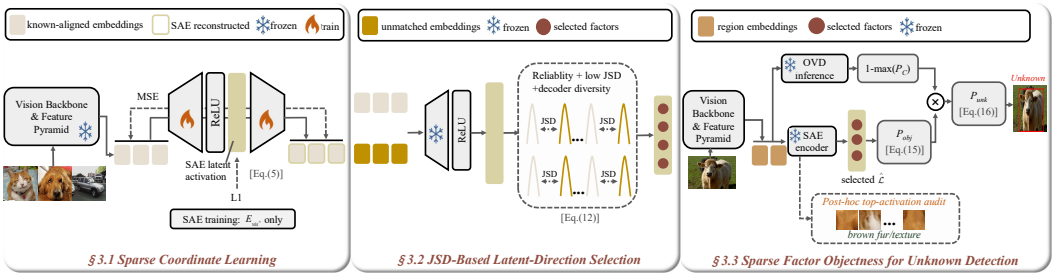


Figure 2: **Training and inference pipeline of SAE-OVOW**. A task-specific SAE decomposes frozen OVD region embeddings, and reliable and non-redundant directions are selected to estimate sparse factor objectness. The audit panel shows top-activating crops and is used only for post-hoc analysis.

A lower divergence indicates similar activation behavior across the two regions and favors directions with limited specialization to current known classes.

Distributional similarity alone may favor inactive directions whose activation distributions are trivially close. We impose reliability constraints based on activation frequency and activation strength. The activation frequencies are

$$f_i^+ = \frac{1}{|E_{\text{vis}}^+|} \sum_{e_{\text{vis}}^+ \in E_{\text{vis}}^+} \mathbb{I}\left(d_i^+(e_{\text{vis}}^+) > 0\right), \quad f_i^- = \frac{1}{|E_{\text{vis}}^-|} \sum_{e_{\text{vis}}^- \in E_{\text{vis}}^-} \mathbb{I}\left(d_i^-(e_{\text{vis}}^-) > 0\right), \tag{8}$$

and the average strengths over activated samples are

$$\mu_i^+ = \frac{\sum_{e_{\text{vis}}^+ \in E_{\text{vis}}^+} d_i^+(e_{\text{vis}}^+)\mathbb{I}(d_i^+(e_{\text{vis}}^+) > 0)}{\sum_{e_{\text{vis}}^+ \in E_{\text{vis}}^+} \mathbb{I}(d_i^+(e_{\text{vis}}^+) > 0) + \epsilon}, \quad \mu_i^- = \frac{\sum_{e_{\text{vis}}^- \in E_{\text{vis}}^-} d_i^-(e_{\text{vis}}^-)\mathbb{I}(d_i^-(e_{\text{vis}}^-) > 0)}{\sum_{e_{\text{vis}}^- \in E_{\text{vis}}^-} \mathbb{I}(d_i^-(e_{\text{vis}}^-) > 0) + \epsilon}. \tag{9}$$

We retain a direction only if

$$\sqrt{f_i^+ f_i^-} \geq \tau_f, \quad \sqrt{\mu_i^+ \mu_i^-} \geq \tau_\mu. \tag{10}$$

Here, ϵ stabilizes the denominators, while τ_f and τ_μ specify the minimum activation frequency and strength. Let $\mathcal{C}$ be the reliable candidate set. To balance distribution similarity and decoder diversity, we define the redundancy of candidate l_i with the selected set $\hat{\mathcal{L}}$ as

$$\mathcal{R}(l_i, \hat{\mathcal{L}}) = \begin{cases} 0, & |\hat{\mathcal{L}}| = 0, \\ \frac{1}{|\hat{\mathcal{L}}|} \sum_{l_j \in \hat{\mathcal{L}}} \sigma\left(\cos\left(l_i^{\text{dec}}, l_j^{\text{dec}}\right)\right), & |\hat{\mathcal{L}}| > 0, \end{cases} \tag{11}$$

where l_i^{dec} is the decoder vector associated with direction l_i, and σ is the Sigmoid function. At each step, we select

$$l^* = \underset{l_i \in \mathcal{C} \setminus \hat{\mathcal{L}}}{\text{argmin}} \left[\beta \cdot \text{JSD}(D_i^+, D_i^-) + (1-\beta) \cdot \mathcal{R}(l_i, \hat{\mathcal{L}})\right], \quad \hat{\mathcal{L}} = \hat{\mathcal{L}} \cup \{l^*\}. \tag{12}$$

Eq. (12) is repeated until $|\hat{\mathcal{L}}| = k$ or the candidate set is exhausted.

3.3 Sparse Factor Objectness for Unknown Detection

For each proposal, the selected latent activations define sparse factor objectness P_{obj}, complementing known-class scores under weak vocabulary alignment. At inference time, the original known-class prediction pathway is left unchanged.

For a candidate region embedding e_{vis}, the detector produces known-class scores

$$P_C = h_{\text{cls}}(e_{\text{vis}}, E_C), \qquad P_C = \{p_1, \ldots, p_K\}. \tag{13}$$

In parallel, we compute the SAE latent activation of the same region:

$$h(e_{\text{vis}}) = \text{ReLU}(W_E^\top e_{\text{vis}} + \mathbf{b}_E). \tag{14}$$

$h(e_{\text{vis}})$ maps the dense region embedding into the learned sparse coordinate system, where each positive entry is the response along one dictionary direction.

Restricting this activation to the selected latent-direction set $\hat{\mathcal{L}}$, where $|\hat{\mathcal{L}}| = k$, we define the active selected set $\mathcal{A}(e_{\text{vis}}) = \{j \in \hat{\mathcal{L}} : h_j(e_{\text{vis}}) > 0\}$. Sparse factor objectness aggregates positive factor activations through a softmax-weighted average after sigmoid:

$$P_{\text{obj}} = \begin{cases} \sum_{j \in \mathcal{A}(e_{\text{vis}})} \frac{\exp(\sigma(h_j(e_{\text{vis}})))}{\sum_{m \in \mathcal{A}(e_{\text{vis}})} \exp(\sigma(h_m(e_{\text{vis}})))} \sigma(h_j(e_{\text{vis}})), & |\mathcal{A}(e_{\text{vis}})| > 0, \\ 0, & |\mathcal{A}(e_{\text{vis}})| = 0, \end{cases} \tag{15}$$

P_{obj} aggregates the positive activations of the selected latent directions into a class-agnostic objectness score. Inactive directions contribute zero to this score.

We combine sparse factor objectness with known-class uncertainty:

$$P_{\text{unk}} = P_{\text{obj}} \cdot (1 - max(P_C)). \tag{16}$$

Eq.16 produces a high unknown-class score for proposals with high sparse factor objectness and low confidence for all known classes. Appending P_{unk} to the known-class scores enables unknown-object detection while preserving the original open-vocabulary classification pathway.

4 Experiments

We evaluate SAE-OVOW on M-OWODB, S-OWODB, and nu-OWODB, with additional incremental detection experiments on PASCAL VOC 2007. Our main experiments use YOLO-World-XL as the frozen detector. Each task uses an SAE with 2,048 latent units and 100 selected factors for objectness estimation. Dataset splits, evaluation metrics, and implementation details are provided in in Appendix A.2.

4.1 Main Results

Results on M-OWODB and S-OWODB. Table 1 shows that SAE-OVOW achieves the highest U-Recall in Tasks 1–3 on both benchmarks. On M-OWODB, it reaches 66.0, 66.7, and 71.0 U-Recall, outperforming the strongest baselines by 0.9, 0.5, and 1.4 points. It also obtains the best overall mAP in Tasks 1–4, with gains of 0.4, 3.7, 1.2, and 0.7 points. Against OWOBJ, its Task-1 known-class mAP is 5.2 points higher, indicating a favorable recall–precision balance. On the superclass-disjoint S-OWODB split, SAE-OVOW reaches 70.8, 75.6, and 61.4 U-Recall, improving over the strongest baselines by 2.1, 12.7, and 0.7 points. It achieves the best overall mAP in Tasks 2–4 (60.4, 56.1, and 56.3), while Task 1 mAP is 0.7 points below IPOW. These results demonstrate consistent gains under both standard and semantically disjoint task splits.

Known–Unknown Confusion Analysis. Table 2 evaluates whether higher unknown recall increases known–unknown confusion. SAE-OVOW combines U-Recall values of 66.0, 66.7, and 71.0 with WI values of 0.0047, 0.0037, and 0.0018 across Tasks 1–3. Relative to the lowest competing values, WI is reduced by 80.4%, 52.6%, and 66.7%. A-OSE also decreases to 1531, 1164, and 938, corresponding to reductions of 58.0%, 20.3%, and 16.7%. The simultaneous improvement in recall and confusion metrics indicates that the gains do not arise from indiscriminately assigning proposals to the unknown class. Instead, sparse factor objectness and known-class uncertainty improve unknown-object detection while reducing known–unknown confusion.

Results on nu-OWODB. On nu-OWODB Table 3 summarizes the results on nu-OWODB. The nu-OWODB benchmark evaluates SAE-OVOW in driving scenes characterized by distinctive object scales and contextual regularities. In Task 1, SAE-OVOW reaches 41.1 U-Recall, 4.4 points below OVOW, but achieves the lowest WI and A-OSE of 0.0008 and 215. In Task 2, it reverses the recall gap and obtains the best U-Recall of 43.8, exceeding OVOW by 3.0 points, together with the lowest WI of 0.0004, the lowest A-OSE of 151, and the highest overall mAP of 23.6. In Task 3, overall mAP reaches 25.8, surpassing OVOW by 1.6 points. These results show that sparse factor objectness remains effective in a driving-oriented OWOD setting, particularly after vocabulary expansion, without relying on benchmark-specific concept supervision.

Table 1: **Results on M-OWODB (top) and S-OWODB (bottom)**. We report known-class mAP and unknown-class U-Recall; Prev. and Curr. denote previously and currently introduced classes. Best results are bold.

Task IDs(→)	**Task 1**		**Task 2**				**Task 3**				**Task 4**		
Method	**U-Recall** (↑)	mAP (↑) Curr.	**U-Recall** (↑)	mAP(↑) Prev.	Curr.	Both	**U-Recall** (↑)	mAP(↑) Prev.	Curr.	Both	mAP(↑) Prev.	Curr.	Both
PROB (Zohar et al., 2023b)	19.4	59.5	17.4	55.7	32.2	44.0	19.6	43.0	22.2	36.0	35.7	18.9	31.5
OrthogonalDet (Sun et al., 2024)	24.6	61.3	26.3	55.5	38.5	47.0	29.1	46.7	30.6	41.3	42.4	24.3	37.9
SKDF (Ma et al., 2025)	39.0	56.8	36.7	52.3	28.3	40.3	36.1	36.9	16.4	30.1	30.1	14.7	26.9
OWOBJ (Zhang et al., 2025)	23.6	61.4	23.8	58.4	34.4	45.7	25.1	44.8	27.8	38.8	36.4	20.7	32.0
CROWD (Majee et al., 2025)	57.9	61.7	53.6	56.7	38.9	47.8	69.6	48.0	31.4	42.5	42.9	25.4	38.5
IPOW (Lv et al., 2026)	50.1	62.4	41.9	61.7	43.6	52.7	46.3	49.7	35.5	45.0	46.7	28.6	42.2
DEUS (Heo et al., 2026)	65.1	66.2	66.2	61.0	45.7	53.3	69.0	53.4	43.3	50.1	50.5	32.8	46.0
SAE-OVOW (ours)	**66.0**	**66.6**	**66.7**	**67.8**	**46.3**	**57.0**	**71.0**	**54.5**	**44.9**	**51.3**	**51.3**	**32.9**	**46.7**
PROB (Zohar et al., 2023b)	17.6	73.4	22.3	66.3	36.0	50.4	24.8	47.8	30.4	42.0	42.6	31.7	39.9
OrthogonalDet (Sun et al., 2024)	24.6	71.6	27.9	64.0	39.9	51.3	31.9	52.1	42.2	48.8	48.7	38.8	46.2
SKDF (Ma et al., 2025)	60.9	69.4	60.0	63.8	26.9	44.4	58.6	46.2	28.0	40.1	41.8	29.6	38.7
CROWD (Majee et al., 2025)	50.5	73.5	41.7	64.9	41.2	53.1	49.6	54.7	42.1	48.4	49.8	43.0	46.4
IPOW (Lv et al., 2026)	34.7	**73.6**	32.6	67.9	42.3	55.1	44.3	56.3	44.6	50.5	50.2	44.6	47.4
DEUS (Heo et al., 2026)	68.7	71.6	62.9	63.5	43.0	52.7	60.7	53.6	45.4	50.9	50.7	42.8	48.8
SAE-OVOW (ours)	**70.8**	72.9	**75.6**	**72.6**	**49.3**	**60.4**	**61.4**	**59.8**	**48.9**	**56.1**	**57.7**	**52.2**	**56.3**

Table 2: **Unknown-object confusion on M-OWODB**. We report U-Recall, wilderness impact (WI), and absolute open-set error (A-OSE) for Tasks 1–3; Task 4 contains no unknown classes.

Task IDs (→)	**Task 1**			**Task 2**			**Task 3**		
Method	**U-Recall** (↑)	WI (↓)	A-OSE (↓)	**U-Recall** (↑)	WI (↓)	A-OSE (↓)	**U-Recall** (↑)	WI (↓)	A-OSE (↓)
ORE (Joseph et al., 2021)	4.9	0.0621	10459	2.9	0.0282	10445	3.9	0.0211	7990
PROB (Zohar et al., 2023b)	19.4	0.0569	5195	17.4	0.0344	6452	19.6	0.0151	2641
RandBox (Wang et al., 2023)	10.6	0.0240	4498	6.3	0.0078	1880	7.8	0.0054	1452
OrthogonalDet (Sun et al., 2024)	24.6	0.0299	4148	26.3	0.0099	1791	29.1	0.0077	1345
CROWD (Majee et al., 2025)	57.6	0.0380	3823	53.6	0.0101	1508	69.6	0.0066	1266
IPOW (Lv et al., 2026)	50.1	0.0369	3648	41.9	0.0098	1460	46.3	0.0062	1126
SAE-OVOW (ours)	**66.0**	**0.0047**	**1531**	**66.7**	**0.0037**	**1164**	**71.0**	**0.0018**	**938**

Table 3: **Results on nu-OWODB**. We report U-Recall, WI, A-OSE, and known-class mAP; PK and CK denote previously and currently introduced known classes. Best results are bold.

Task IDs (→)	Task 1				Task 2						Task 3		
Methods	U-Recall (↑)	WI (↓)	A-OSE (↓)	mAP (↑) CK	U-Recall (↑)	WI (↓)	A-OSE (↓)	mAP (↑) PK	CK	Both	mAP (↑) PK	CK	Both
YOLO-UniOW-S (Liu et al., 2024)	37.5	0.0137	1658	21.5	30.0	0.0074	1265	21.5	9.8	16.7	16.7	15.6	16.4
YOLO-UniOW-M (Liu et al., 2024)	41.8	0.0147	1722	24.8	35.4	0.0067	1156	24.8	13.1	20.0	20.0	18.4	19.6
OVOW (Li et al., 2024b)	**45.5**	0.0185	1724	**28.1**	40.8	0.0106	1703	27.8	15.5	22.8	**23.8**	25.3	24.2
SAE-OVOW (ours)	41.1	**0.0008**	**215**	22.1	**43.8**	**0.0004**	**151**	**28.3**	**16.3**	**23.6**	23.6	**26.9**	**25.8**

Table 4: Overall mAP on PASCAL VOC 2007 under three **single-step incremental object detection** settings. Best results are bold.

Method	10+10	15+5	19+1
OWOBJ (Zhang et al., 2025)	69.9	73.3	75.8
CROWD (Majee et al., 2025)	73.5	74.4	74.2
IPOW (Lv et al., 2026)	73.8	75.0	74.5
SAE-OVOW (ours)	**87.9**	**87.4**	**87.0**

Results on IOD. SAE-OVOW is evaluated on PASCAL VOC 2007 under the 10+10, 15+5, and 19+1 settings. As summarized in Table 4 and detailed in Appendix Table 20, it achieves 87.9, 87.4, and 87.0 mAP. These results exceed CROWD by 14.4, 13.0, and 12.8 points and OWOBJ by 18.0, 14.1, and 11.2 points. The gains indicate that task-specific SAE relearning supports vocabulary expansion while retaining previously learned classes, demonstrating generalization beyond unknown-object detection.

4.2 Ablation Study

Component Contributions to Sparse Factor Objectness. Table 5(a) measures the contribution of each selection stage on M-OWODB Task 1. The OOD-only control uses $\gamma(1 - \max(P_C))$ with

Table 5: **Ablation studies on M-OWODB Task 1.** (a) OOD: calibrated uncertainty ($\gamma = 0.5$); Obj.: generic object prompt; All: all SAE directions; JSD: Jensen–Shannon divergence ranking; SAE-OVOW: JSD ranking with decoder diversity. (b) PCA: principal components; Rand. SAE: random SAE directions; Dense SAE: SAE without sparsity; Rand. Orth.: random orthogonal directions. (c) Binary: foreground/background classification; Centerness: FCOS centerness; IoU: IoU prediction; Gaussian: PROB-style Gaussian likelihood.

Method	U-Recall (↑)	U-mAP (↑)	Known mAP (↑)
OOD only	29.2	0.2	57.7
Obj. prompt	20.8	**5.5**	**73.5**
All directions	31.7	2.3	71.0
JSD rank	65.8	5.1	66.6
SAE-OVOW	**66.0**	5.3	66.6

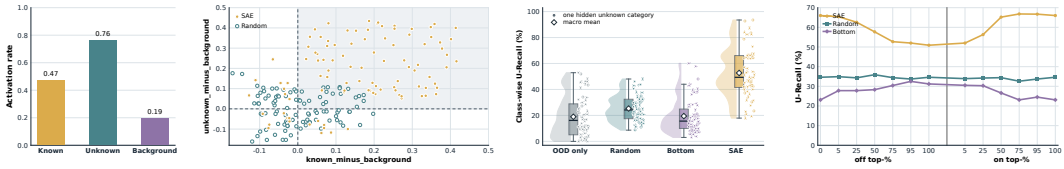
(a) Component ablation

Method	U-Recall (↑)	U-mAP (↑)	Known mAP (↑)
PCA	46.0	1.5	61.1
Rand. Orth.	44.0	1.1	61.6
Dense SAE	29.2	0.2	65.5
Rand. SAE	29.2	1.2	66.5
SAE-OVOW	**66.0**	**5.3**	**66.6**

(b) Dictionary and direction

Method	U-Recall (↑)	U-mAP (↑)	Known mAP (↑)
Binary Obj.	**70.6**	1.8	0.7
Centerness	49.7	3.9	17.2
IoU Obj.	36.9	1.3	5.8
Gaussian Obj.	24.0	0.1	50.0
SAE-OVOW	66.0	**5.3**	**66.6**

(c) Objectness branches



(a) Overall activation profile (b) Per-direction activation gaps (c) Class-wise U-Recall gains (d) Feature-space removal and recovery

Figure 3: **Activation structure and functional relevance of sparse factor objectness on M-OWODB Task 1**. (a) Activation rates across known, unknown, and background regions. (b) Per-direction activation-rate gaps between object regions and background for selected and random directions. (c) Class-wise U-Recall under OOD-only scoring and random, bottom-ranked, or selected SAE directions; points show one-hidden-unknown-category evaluations and diamonds denote macro means. (d) U-Recall after progressively removing latent components from the feature representation and restoring them in ranked order.

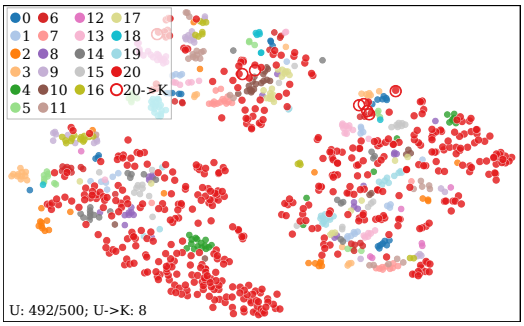
Table 6: **Effect of frozen YOLO-World backbone scale on M-OWODB**. We compare the calibrated OOD-only baseline with SAE-OVOW using YOLO-World-M and YOLO-World-L across Tasks 1–3. Each task reports U-Recall, U-mAP, and known-class mAP.

Task IDs (→)	Task 1			Task 2			Task 3		
Models	U-Recall (↑)	U-mAP (↑)	mAP (↑)	U-Recall (↑)	U-mAP (↑)	mAP (↑)	U-Recall (↑)	U-mAP (↑)	mAP (↑)
YWorld-M + OOD only	20.7	0.1	53.4	22.3	0.1	52.8	27.1	0.0	44.4
YWorld-L + OOD only	28.5	0.1	55.1	29.3	0.1	55.0	36.6	0.1	46.0
YWorld-M + SAE-OVOW	48.6	4.1	52.3	55.9	1.8	52.0	60.1	0.4	43.7
YWorld-L + SAE-OVOW	49.8	4.5	54.2	58.0	1.5	54.1	65.4	0.7	45.6

$\gamma = 0.5$ and obtains 29.2 U-Recall, 0.2 U-mAP, and 57.7 known mAP. The generic object prompt raises U-mAP and known mAP to 5.5 and 73.5, with U-Recall decreasing to 20.8. Using all SAE directions yields 31.7 U-Recall, 2.3 U-mAP, and 71.0 known mAP. Reliability filtering and JSD ranking raise U-Recall to 65.8 and U-mAP to 5.1; decoder diversity gives the final 66.0 U-Recall and 5.3 U-mAP while retaining 66.6 known mAP. Relative to OOD only, SAE-OVOW gains 36.8 U-Recall, 5.1 U-mAP, and 8.9 known-mAP points, showing the cumulative contribution of sparse direction selection.

Sparse Dictionary and Direction Selection. Table 5(b) compares the learned sparse dictionary with alternative coordinate systems and selection schemes. PCA and random orthogonal directions obtain 46.0 and 44.0 U-Recall, respectively, while reaching only 1.5 and 1.1 U-mAP, with known mAP values of 61.1 and 61.6. Removing sparsity reduces Dense SAE to 29.2 U-Recall and 0.2 U-mAP. Randomly sampling directions from the trained SAE preserves 66.5 known mAP but gives only 29.2 U-Recall and 1.2 U-mAP. SAE-OVOW combines the sparse data-adaptive dictionary with structured direction selection and achieves 66.0 U-Recall, 5.3 U-mAP, and 66.6 known mAP. The gains over generic coordinates and random SAE directions support the joint contribution of sparse representation learning and reliability-, distribution-, and diversity-aware selection.

Conventional Objectness Branches. Table 5(c) compares conventional objectness branches on frozen OVD features. The RPN-style binary classifier (Ren et al., 2015) attains the highest U-



Unknown Predictions in Frozen Feature Space. Fig. 4 shows that unknown regions occupy multiple neighborhoods and partly overlap with known-class groups in the frozen OVD feature space. Nevertheless, SAE-OVOW assigns 492 of the 500 visualized unknown regions (98.4%) to the unknown class, with only eight assigned to known classes. This suggests that sparse factor evidence remains effective even when unknown and known regions are close in the projected feature space.

Figure 4: **t-SNE of GT-matched frozen OVD region embeddings on M-OWODB Task 1**. Labels 0– 19 denote known classes. Filled and open red circles denote unknown regions predicted as unknown and known.

Recall of 70.6, together with 1.8 U-mAP and 0.7 known mAP. The accompanying loss of known-class accuracy is consistent with the category bias of binary proposal classifiers (Kim et al., 2022). FCOS centerness (Tian et al., 2019) reaches 49.7 U-Recall, 3.9 U-mAP, and 17.2 known mAP, while IoU prediction (Jiang et al., 2018; Kim et al., 2022) gives 36.9, 1.3, and 5.8, respectively. The PROB-style Gaussian likelihood (Zohar et al., 2023b) preserves 50.0 known mAP but reaches only 24.0 U-Recall and 0.1 U-mAP. SAE-OVOW obtains 66.0 U-Recall, the highest U-mAP of 5.3, and the highest known mAP of 66.6, providing the strongest balance between unknown discovery and known-class recognition.

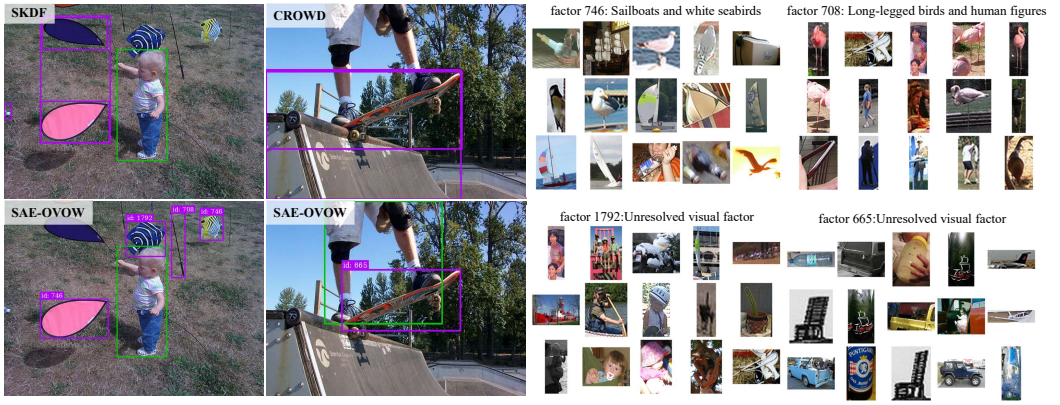
Activation Structure and Category-level Gains. Fig. 3(a)–(c) characterizes sparse factor objectness at the population, direction, and category levels. At the population level, the selected factors activate on 0.76 of unknown-object regions and 0.47 of known-object regions, compared with 0.19 of background regions. This foreground preference extends across the current vocabulary boundary and is particularly strong for unknown objects. At the direction level, most selected factors occupy the positive-positive quadrant, with higher activation rates on both known and unknown objects than on background. Random directions remain concentrated near the origin and frequently exhibit weak or negative activation gaps. The selected set therefore captures multiple object-responsive directions with consistent foreground-background separation. At the category level, these activation properties shift the class-wise U-Recall distribution upward across the one-hidden-unknown-category evaluations. Selected SAE directions achieve the highest macro mean and substantially stronger per-category recall than OOD-only scoring, random directions, and bottom-ranked directions. The progression across the three panels connects population-level activation patterns and individual factor behavior to detection gains across diverse unknown categories.

Functional Relevance of Selected Directions. Fig. 3(d) evaluates the selected directions by removing their latent components directly from the feature representation and restoring them in ranked order. U-Recall declines progressively as larger fractions of the selected SAE directions are removed, falling from the full-set level of about 66% to about 51% after the complete selected set is removed. Restoring half of the selected directions recovers U-Recall to approximately 65%, and restoring most of the set returns performance to the full-set range. The random and bottom-ranked sets remain substantially lower and vary less across the intervention ratios. This removal-and-recovery response shows that the selected directions form a functionally concentrated subspace carrying feature evidence for unknown-object recall; complete interventions for both benchmarks and Tasks 1–3 appear in Appendix B.2.

Sparse Factor Objectness across Detector Scales. Table 6 evaluates whether SAE-OVOW generalizes across detector capacities. With YOLO-World-M, SAE-OVOW improves U-Recall by 27.9–33.6 points and U-mAP by 0.4–4.0 points across Tasks 1–3, while reducing known-class mAP by at most 1.1 points. With YOLO-World-L, the corresponding gains are 21.3–28.8 points in U-Recall and 0.6–4.4 points in U-mAP, with a maximum known-class mAP reduction of 0.9 points. Scaling from YOLO-World-M to YOLO-World-L improves SAE-OVOW's U-Recall by 1.2–5.3 points and known-class mAP by 1.9–2.1 points. These results show that sparse factor objectness remains effective across backbone scales rather than depending on a particular frozen detector.

Table 7: **Post-hoc semantic audit of selected factors on OWODB Task 1**.

Benchmark	Coherent / Audited	Unresolved	Coherence	Coherent among top 10	Representative recurring visual patterns
M-OWODB	54/100	46	54.0%	7/10	vehicle rear panels; circular openings; surface textures
S-OWODB	65/100	35	65.0%	7/10	sports fields and crowds; standing people; collars;
Overall	119/200	81	59.5%	14/20	parts and shapes; local textures; scene-level context



(a) Baseline and SAE-OVOW predictions (b) Top-activating regions of selected factors

Figure 5: **Qualitative comparison on M-OWODB Task 1**. (a) Baselines appear above SAE-OVOW; purple and green boxes denote unknown and known predictions. (b) Top-activation regions for the displayed factors.

4.3 Qualitative Analysis

Qualitative Analysis. SAE-OVOW aggregates the complete selected-factor set when scoring each region. The factor ID displayed in Fig. 5 denotes only the strongest selected activation and is provided as a representative trace rather than a single-factor explanation. In the left example, SAE-OVOW separates disc-like objects missed or merged by SKDF; in the right example, it removes CROWD's oversized ramp response and localizes the skateboard. The corresponding top-activation crops show coherent recurring patterns for factors 746 and 708, whereas latents 1792 and 665 remain unresolved.

Post-hoc Semantic Audit. SAE-OVOW selects objectness-relevant latent directions using activation reliability, activation-distribution similarity, and decoder diversity. Semantic descriptions enter only the post-hoc audit, after the selected sets have been fixed. We manually inspect the top-15 activating region crops of each factor and record recurring visual patterns. Table 7 summarizes the 100 selected factors from Task 1 of each OWODB benchmark. Among the 200 audited factors, 119 (59.5%) receive human-readable descriptions: 54/100 on M-OWODB and 65/100 on S-OWODB. Among the ten highest-ranked factors within each benchmark, 14/20 exhibit coherent patterns. The complete rank-ordered audit of all 200 selected factors, including unresolved factors and their top-activation regions, is provided in the supplementary material.

5 Conclusion

SAE-OVOW shows that frozen OVD region embeddings encode class-agnostic objectness information beyond known-class scores. It learns sparse coordinates, selects reliable and non-redundant latent directions, and combines their activations with known-class uncertainty for unknown detection. Experiments on M-OWODB, S-OWODB, and nu-OWODB demonstrate strong unknown recall, low known–unknown confusion, and compatibility with vocabulary expansion. Ablations and feature-space removal interventions verify that the selected directions drive the detection gains, while post-hoc audits reveal coherent visual patterns in a subset of selected factors. We hope this work encourages broader use of internal objectness information for reliable and interpretable open-world detection.

ETHICS STATEMENT

This work studies open world object detection using public research benchmarks and does not introduce new personal or sensitive data. Potential applications include safety monitoring and autonomous perception; the same capability may also support surveillance or profiling. Performance can vary across object classes and environments because of dataset bias. Real-world deployment should therefore include domain-specific evaluation, privacy safeguards, human oversight, and procedures for contesting incorrect detections. The proposed method does not infer personal identity or protected attributes.

REPRODUCIBILITY STATEMENT

The manuscript specifies the architecture, objectives, factor-selection criteria, inference equations, algorithms, dataset splits, metrics, training settings, and hyperparameter studies. The implementation builds on publicly available YOLO-World and OWOD benchmarks. Appendix A gives the experimental protocol and implementation details; the subsequent appendices report hyperparameter analyses, controlled interventions, semantic audits, and additional incremental results. The supplementary material provides complete selected-factor audit records, activation visualizations, and additional artifacts beyond the representative examples included in the appendices. Task-specific SAEs are trained with the reported thresholds, expansion ratio, optimizer settings, and hardware configuration. These details are intended to support independent reproduction without access to proprietary data.

AI USE STATEMENT

Generative AI tools were used to assist language polishing, drafting, LaTeX organization, and migration of the manuscript to the official ICLR 2027 template. All technical claims, equations, experiments, figures, references, and final wording were checked by the authors. AI-generated data, labels, or model outputs were not used to train or evaluate SAE-OVOW, and no generative AI model is part of the proposed inference pipeline. The authors take responsibility for the final content of this work, including text and artifacts produced with the aid of generative AI.

REFERENCES

Bart Bussmann, Noa Nabeshima, Adam Karvonen, and Neel Nanda. Learning multi-level features with matryoshka sparse autoencoders. *arXiv preprint arXiv:2503.17547*, 2025.

David Chanin, Tomáš Dulka, and Adrià Garriga-Alonso. Feature hedging: Correlated features break narrow sparse autoencoders. *arXiv preprint arXiv:2505.11756*, 2025.

David Chanin, James Wilken-Smith, Tomáš Dulka, Hardik Bhatnagar, Satvik Golechha, and Joseph Bloom. A is for absorption: Studying feature splitting and absorption in sparse autoencoders. *Advances in Neural Information Processing Systems*, 38:82318–82355, 2026.

Tianheng Cheng, Lin Song, Yixiao Ge, Wenyu Liu, Xinggang Wang, and Ying Shan. Yolo-world: Real-time open-vocabulary object detection. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16901–16911. IEEE, 2024.

Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=F76bwRSLeK.

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

Ruohuan Fang, Guansong Pang, Wenjun Miao, Xiao Bai, Jin Zheng, and Xin Ning. Unsupervised recognition of unknown objects for open-world object detection. *IEEE Transactions on Neural Networks and Learning Systems*, 36(6):11340–11354, 2025.

Chengjian Feng, Yujie Zhong, Yu Gao, Matthew R. Scott, and Weilin Huang. TOOD: Task-aligned one-stage object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3510–3519, 2021.

Leo Gao, Tom Dupre la Tour, Henk Tillman, Gabriel Goh, Rajan Troll, Alec Radford, Ilya Sutskever, Jan Leike, and Jeffrey Wu. Scaling and evaluating sparse autoencoders. In *International Conference on Learning Representations*, volume 2025, pp. 26721–26754, 2025.

Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. *arXiv preprint arXiv:2104.13921*, 2021.

Akshita Gupta, Sanath Narayan, KJ Joseph, Salman Khan, Fahad Shahbaz Khan, and Mubarak Shah. Ow-detr: Open-world detection transformer. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9225–9234. IEEE, 2022.

Jun-Woo Heo, Keonhee Park, and Gyeong-Moon Park. Detecting unknown objects via energy-based separation for open world object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

Borui Jiang, Ruixuan Luo, Jiayuan Mao, Tete Xiao, and Yuning Jiang. Acquisition of localization confidence for accurate object detection. In *Proceedings of the European Conference on Computer Vision*, pp. 784–799, 2018.

KJ Joseph, Salman Khan, Fahad Shahbaz Khan, and Vineeth N Balasubramanian. Towards open world object detection. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5826–5836. IEEE, 2021.

Dahun Kim, Tsung-Yi Lin, Anelia Angelova, In So Kweon, and Weicheng Kuo. Learning open-world object proposals without learning to classify. *IEEE Robotics and Automation Letters*, 7(2): 5453–5460, 2022.

Yiming Li, Yi Wang, Wenqian Wang, Dan Lin, Bingbing Li, and Kim-Hui Yap. Open world object detection: A survey. *IEEE Transactions on Circuits and Systems for Video Technology*, 35(2): 988–1008, 2024a.

Zizhao Li, Zhengkang Xiang, Joseph West, and Kourosh Khoshelham. From open vocabulary to open world: Teaching vision language models to detect novel objects. *arXiv preprint arXiv:2411.18207*, 2024b.

Hyesu Lim, Jinho Choi, Jaegul Choo, and Steffen Schneider. Sparse autoencoders reveal selective remapping of visual concepts during adaptation. In *The Thirteenth International Conference on Learning Representations*, 2025.

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

Lihao Liu, Juexiao Feng, Hui Chen, Ao Wang, Lin Song, Jungong Han, and Guiguang Ding. YOLO-UniOW: Efficient universal open-world object detection. *arXiv preprint arXiv:2412.20645*, 2024.

Xueqiang Lv, Shizhou Zhang, Yinghui Xing, Di Xu, Peng Wang, and Yanning Zhang. Knowing the unknown: Interpretable open-world object detection via concept decomposition model. *arXiv preprint arXiv:2602.20616*, 2026.

Shuailei Ma, Yuefeng Wang, Ying Wei, Jiaqi Fan, Thomas H Li, Hongli Liu, and Fanbing Lv. Cat: Localization and identification cascade detection transformer for open-world object detection. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19681–19690. IEEE, 2023.

Shuailei Ma, Yuefeng Wang, Ying Wei, Enming Zhang, Jiaqi Fan, Xinyu Sun, and Peihao Chen. Skdf: a simple knowledge distillation framework for distilling open-vocabulary knowledge to open-world object detector. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

Anay Majee, Amitesh Gangrade, and Rishabh Iyer. Looking beyond the known: Towards a data discovery guided open-world object detection. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Matthias Minderer, Alexey Gritsenko, Austin Stone, Maxim Neumann, Dirk Weissenborn, Alexey Dosovitskiy, Aravindh Mahendran, Anurag Arnab, Mostafa Dehghani, Zhuoran Shen, et al. Simple open-vocabulary object detection. In *European conference on computer vision*, pp. 728–755. Springer, 2022.

Mateusz Pach, Shyamgopal Karthik, Quentin Bouniot, Serge Belongie, and Zeynep Akata. Sparse autoencoders learn monosemantic features in vision-language models. In *Advances in Neural Information Processing Systems*, volume 38, pp. 95706–95742, 2025.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems*, volume 28, 2015.

Fei Shen and Jinhui Tang. Imagpose: A unified conditional framework for pose-guided person generation. *Advances in neural information processing systems*, 37:6246–6266, 2024.

Fei Shen, Hu Ye, Jun Zhang, Cong Wang, Xiao Han, and Yang Wei. Advancing pose-guided image synthesis with progressive conditional diffusion models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=rHzapPnCgT.

Fei Shen, Xiaoyu Du, Yutong Gao, Jian Yu, Yushe Cao, Xing Lei, and Jinhui Tang. Imagharmony: Controllable image editing with consistent object quantity and layout. *arXiv preprint arXiv:2506.01949*, 2025a.

Fei Shen, Xin Jiang, Xin He, Hu Ye, Cong Wang, Xiaoyu Du, Zechao Li, and Jinhui Tang. Imagdressing-v1: Customizable virtual dressing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 6795–6804, 2025b.

Fei Shen, Cong Wang, Junyao Gao, Qin Guo, Jisheng Dang, Jinhui Tang, and Tat-Seng Chua. Long-term talkingface generation via motion-prior conditional diffusion model. *arXiv preprint arXiv:2502.09533*, 2025c.

Fei Shen, Hu Ye, Sibo Liu, Jun Zhang, Cong Wang, Xiao Han, and Yang Wei. Boosting consistency in story visualization with rich-contextual conditional diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 6785–6794, 2025d.

Fei Shen, Jian Yu, Cong Wang, Xin Jiang, Xiaoyu Du, and Jinhui Tang. Imaggarment: Fine-grained garment generation for controllable fashion design. *IEEE Transactions on Visualization and Computer Graphics*, 2026.

Zhicheng Sun, Jinghan Li, and Yadong Mu. Exploring orthogonality in open world object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17302–17312, 2024.

Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. Fcos: Fully convolutional one-stage object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9627–9636, 2019.

Yanghao Wang, Zhongqi Yue, Xian-Sheng Hua, and Hanwang Zhang. Random boxes are open-world object detectors. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6210–6220. IEEE, 2023.

Xing Xi, Yangyang Huang, Zhijie Zhong, and Ronghua Luo. Umb: Understanding model behavior for open-world object detection. *Advances in Neural Information Processing Systems*, 37:74233–74261, 2024.

Xing Xi, Xing Fu, Weiqiang Wang, and Ronghua Luo. OW-VAP: Visual attribute parsing for open world object detection. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 68190–68207. PMLR, 2025a. URL `https://proceedings.mlr.press/v267/xi25b.html`.

Xing Xi, Yangyang Huang, Ronghua Luo, and Yu Qiu. OW-OVD: Unified open world and open vocabulary object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 25454–25464, June 2025b.

Muli Yang, Gabriel James Goenawan, Huaiyuan Qin, Kai Han, Xi Peng, Yanhua Yang, and Hongyuan Zhu. Detecting open world objects via partial attribute assignment. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20318–20328. IEEE, 2025.

Mısra Yavuz and Fatma Güney. O1o: Grouping of known classes to identify unknown objects as odd-one-out. In *Asian Conference on Computer Vision*, pp. 394–410. Springer Nature Singapore Singapore, 2024.

Qi Zhang, Yifei Wang, Xiaohan Wang, Jiajun Chai, Guojun Yin, Wei Lin, and Yisen Wang. SAE as a crystal ball: Interpretable features predict cross-domain transferability of LLMs without training. In *The Fourteenth International Conference on Learning Representations*, 2026.

Shan Zhang, Yao Ni, Jinhao Du, Yuan Xue, Philip Torr, Piotr Koniusz, and Anton Van Den Hengel. Open-world objectness modeling unifies novel object detection. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 30332–30342. IEEE, 2025.

Xiaowei Zhao, Yuqing Ma, Duorui Wang, Yifan Shen, Yixuan Qiao, and Xianglong Liu. Revisiting open world object detection. *IEEE transactions on circuits and systems for video technology*, 34 (5):3496–3509, 2023.

Orr Zohar, Alejandro Lozano, Shelly Goel, Serena Yeung, and Kuan-Chieh Wang. Open world object detection in the era of foundation models. *arXiv preprint arXiv:2312.05745*, 2023a.

Orr Zohar, Kuan-Chieh Wang, and Serena Yeung. Prob: Probabilistic objectness for open world object detection. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11444–11453. IEEE, 2023b.

A Experimental Setup and SAE Training

This section specifies the OWOD protocol and the training and evaluation details used throughout the paper.

- §A.1 defines the problem and summarizes the benchmark splits.
- §A.2 describes training, inference, post-processing, and evaluation metrics.
- §A.3 reports task-wise SAE configurations and representation statistics.

A.1 Problem Definition and Benchmarks

Problem Definition. Open world object detection requires a detector to recognize known object classes while discovering objects from unknown classes. At task t, let $\mathcal{K}^t = \{O_1^t, O_2^t, \ldots, O_{K^t}^t\}$ denote the set of known classes. Object instances that do not belong to $\mathcal{K}^t$ are regarded as unknown and are associated with an unknown class set $\mathcal{U}^t = \{O_{K^t+1}^t, O_{K^t+2}^t, \ldots, O_{K^t+U^t}^t\}$. Given an image, the detector is expected to localize and classify instances from $\mathcal{K}^t$, while assigning instances outside $\mathcal{K}^t$ to the unknown class. After task t, a subset of discovered unknown classes is annotated by an Oracle and incorporated into the known class set: $\mathcal{K}^{t+1} = \mathcal{K}^t \cup \Delta\mathcal{U}^t$, where $\Delta\mathcal{U}^t \subseteq \mathcal{U}^t$ denotes the newly labeled classes. The updated detector should recognize $\mathcal{K}^{t+1}$ while preserving performance on previously learned classes.

Datasets. Table 8 provides the split details of M-OWODB and S-OWODB. M-OWODB is constructed by combining VOC and COCO. In Task 1, all VOC classes are used as known classes, while the remaining COCO classes are treated as unknown. The following tasks are organized according to COCO superclasses. For testing, the VOC (Everingham et al., 2010) and COCO (Lin et al., 2014) validation sets are merged as the M-OWODB test set. During evaluation, all classes introduced up to the current task are considered known; for instance, aeroplane and bicycle are known in Task 1, whereas bed and sofa are treated as unknown. Since VOC and COCO contain overlapping subclasses, their semantic similarity may introduce dataset bias. S-OWODB addresses this issue by using only COCO and strictly partitioning subclasses by superclass. This protocol prevents later tasks from containing objects that belong to the same superclass as earlier tasks, making S-OWODB a more general benchmark for open-world evaluation.

A.2 Implementation, Inference, and Metrics

Training. We use YOLO-World-XL (Cheng et al., 2024) as the frozen open-vocabulary detector. For each OWOD subtask, one forward-only pass over the training split collects the known-aligned region embeddings in Eq. (1) to train a task-specific SAE. The matching threshold α is 0.55 on M-OWODB and 0.75 on S-OWODB.

SAE Configuration. Given region-embedding dimension d, the SAE uses an expansion ratio of $r = 4$, yielding $d_{\mathrm{SAE}} = 4d = 2048$. For N training embeddings, sparsity is quantified by the active-latent ratio $s = \frac{100}{N d_{\mathrm{SAE}}} \sum_{n=1}^{N} \sum_{j=1}^{d_{\mathrm{SAE}}} \mathbb{I}[h_j(\mathbf{z}_n) > 0]\%$, where smaller values indicate sparser representations. Table 15 evaluates controlled sparsity levels while holding all other training settings fixed. Each reported level denotes the center of its corresponding activation-ratio interval, and the representative checkpoint is selected closest to that center. The reference setting uses $s = 0.5\%$ with a dead-latent ratio below 0.8%. Table 9 summarizes the task-specific configurations and held-out representation statistics.

Optimization. We train each SAE with AdamW, a batch size of 2048, zero weight decay, and a maximum trajectory of $10{,}000$ optimizer steps. The learning rate is 1×10^{-3} with cosine decay and 200 warmup steps. The ℓ_1 coefficient $\lambda_{\mathrm{L1}} = 0.5$ is independently warmed up for 200 steps. Both linear maps use learnable biases, with the decoder bias initialized by the geometric median of at most $50{,}000$ training embeddings. Ghost-gradient training uses coefficient 0.5 and begins after a latent remains inactive for 100 steps; Top-K selection is disabled. During OWOD testing, we set $\beta = 0.6$, $\tau_f = 1 \times 10^{-7}$, and $\tau_\mu = 1 \times 10^{-3}$.

JSD Distribution Construction. For each target task, we construct latent-wise activation distributions from the training regions available through that task. Regions whose maximum assignment score reaches α form E_{vis}^{+}, while the remaining regions form E_{vis}^{-}; following the incremental proto-

Table 8: **Task composition in M-OWODB and S-OWODB**.

M-OWODB	**Task 1**	**Task 2**	**Task 3**	**Task 4**
Semantic Split	VOC Classes	Outdoor, Accessories Appliances, Truck	Sports, Food	Electronic, Indoor Kitchen, Furniture
# train images	16551	45520	39402	40260
# test images	4952	1914	1642	1738
# train instances	47223	113741	114452	138996
# test instances	14976	4966	4826	6039
S-OWODB	**Task 1**	**Task 2**	**Task 3**	**Task 4**
Semantic Split	Animals, Person Vehicles	Appliances, Accessories Outdoor, Furniture	Sports, Food	Electronic, Indoor Kitchen
# train images	89490	55870	39402	38903
# test images	3793	2351	1642	1691
# train instances	421243	163512	114452	160794
# test instances	17786	7159	4826	7010

col, scores for classes introduced in preceding tasks are omitted before assignment. For every latent, sigmoid-transformed encoder pre-activations are discretized into 10,000 equal-width bins on $[0, 1]$, including samples with zero post-ReLU activation. Counts are aggregated across the available training sources according to their empirical frequencies, and the positive and unmatched histograms are normalized separately. We use 10^{-12} to stabilize normalization and logarithmic terms. These distributions define the JSD in Eq. (7), after which the frequency and activation-strength criteria in Eq. (10) determine eligible factors.

Incremental Learning. SAE-OVOW accommodates incremental learning by expanding the detector vocabulary after each task. Newly annotated classes are added to K^{t+1}, and their text embeddings are generated by the frozen text encoder. The OVD recognizes the enlarged known vocabulary through its original zero-shot classification pathway, with the visual backbone, box regression head, and contrastive classification head fixed. Under the updated vocabulary, known-aligned training regions define the task-specific embedding set used to learn a new sparse dictionary and reselect latent directions from the updated known-aligned and unmatched sets. At each task, the SAE branch is rebuilt from cumulative stored region embeddings, with detector parameters fixed.

Inference and Post-processing. We append P_{unk} to the K known-class scores, yielding $\widetilde{P} = [p_1, \ldots, p_K, P_{\text{unk}}]$. Multi-label inference evaluates every proposal–class pair independently. All score channels use a threshold of 0.001, with at most 30,000 candidates entering class-aware NMS at IoU 0.7. A class-agnostic NMS step at IoU 0.5 resolves overlaps across known and unknown labels. Each image retains at most 300 predictions. Proposal generation and post-processing settings are shared by all compared variants.

Metrics. Following existing OWOD evaluation protocols, we report mean average precision (mAP) for known classes, unknown recall (U-Recall) for unknown classes, wilderness impact (WI), and absolute open set error (A-OSE). Since U-Recall can become less sensitive when using stronger foundation models, we report unknown mean average precision (U-mAP) in our ablation and sensitivity studies. This metric supports analysis of how individual hyperparameters and modules affect unknown-prediction ranking; most prior OWOD methods do not report U-mAP, which precludes direct comparison on this metric.

A.3 SAE Training Configuration and Statistics

Table 9 reports the amount of task-wise training data together with held-out reconstruction, sparsity, and dead-latent statistics for the learned sparse dictionaries.

Table 9: **Task-wise SAE training configuration and representation statistics**. Each task uses cumulative known-aligned region embeddings through the current task, split into 95% training and 5% validation subsets. Shared architecture and optimization settings are described in Appendix A.2. MSE, EV, L1, and L0 denote validation reconstruction error, explained variance, mean absolute activation, and mean active-latent count. Dead is the percentage of latents inactive on the training split.

Benchmark	Task	N_{train}	N_{val}	Step	MSE ↓	EV ↑	L1 ↓	L0 ↓	Dead (%) ↓
M-OWODB	1	232,869	12,258	300	1.302×10^{-4}	0.9334	3.990×10^{-4}	10.31	0.342
M-OWODB	2	560,822	29,518	400	1.459×10^{-4}	0.9253	4.181×10^{-4}	10.86	0.586
M-OWODB	3	935,770	49,252	400	2.093×10^{-4}	0.8928	4.323×10^{-4}	10.33	0.537
M-OWODB	4	1,272,300	66,964	400	2.572×10^{-4}	0.8683	4.356×10^{-4}	9.80	0.439
S-OWODB	1	352,259	18,540	400	8.863×10^{-5}	0.9546	3.581×10^{-4}	9.86	0.781
S-OWODB	2	689,264	36,276	400	1.427×10^{-4}	0.9269	4.115×10^{-4}	9.57	0.586
S-OWODB	3	1,033,722	54,405	400	1.927×10^{-4}	0.9013	4.337×10^{-4}	9.30	0.293
S-OWODB	4	1,348,480	70,971	500	2.421×10^{-4}	0.8760	4.316×10^{-4}	9.44	0.781

B Hyperparameter and Design Analysis

This section groups the sensitivity and design studies by the component they probe.

- §B.1 analyzes region matching in the initial and incremental settings.
- §B.2 studies the diversity balance, reliability criteria, and functional relevance of selected directions.
- §B.3 examines dictionary capacity and activation sparsity.
- §B.4 compares aggregation rules and selected-set sizes.
- §B.5 evaluates robustness to stochastic SAE training and latent reselection.

Analysis Setup. Tables 10–15 study the sensitivity of the main hyperparameters in SAE-OVOW. Unless otherwise specified, only one parameter is varied at a time while the remaining settings are kept fixed. Gray shading indicates the default setting used in the main experiments. The results show a consistent trend: for reasonable non-degenerate settings, these hyperparameters do not change the overall ability of the detector to recall unknown objects. Their influence is mainly reflected in the precision and ranking quality of unknown predictions, which is why we report U-mAP in the ablation and sensitivity studies.

Reference Configuration and Sensitivity Analysis. For each task, SAE fitting, latent-statistic estimation, and factor selection use region embeddings collected from the corresponding training data. We reserve 5% of the known-aligned training embeddings for monitoring SAE reconstruction error, explained variance, activation sparsity, and checkpoint quality (Table 10). The reference configuration uses benchmark-specific matching thresholds of $\alpha = 0.55$ for M-OWODB and $\alpha = 0.75$ for S-OWODB, together with shared settings $\beta = 0.6$, $\tau_f = 10^{-7}$, $\tau_\mu = 10^{-3}$, $d_{\text{SAE}}/d = 4$, $s = 0.5\%$, and $k = 100$. These settings remain fixed across incremental tasks. Tables 11–16 vary one parameter at a time around this reference configuration while keeping the others fixed.

B.1 Region Matching Threshold

Initial-task Matching. The threshold α in Eqs. (1) and (2) partitions region embeddings into the known-aligned set E^{+}_{vis} and unmatched set E^{-}_{vis}, determining the positive distribution used to learn sparse coordinates. Table 10 shows limited sensitivity: M-OWODB U-Recall remains 65.4–66.0, S-OWODB U-Recall remains 70.5–70.8, and known-class mAP is stable. The larger U-mAP variation indicates that α affects unknown-confidence ranking more than proposal coverage. Extreme values mainly alter known-aligned set composition, so recall remains stable while score calibration changes. A permissive α admits weak positives that blur the learned dictionary; a strict α reduces sample diversity. The near-flat central range suggests that the distribution-level SAE objective averages local assignment errors. Thus, α acts primarily as a data-curation threshold, and the plateau indicates that performance has no narrow optimum.

Table 10: **Effect of the matching threshold α on Task 1 performance**.

Setting	M-OWODB			S-OWODB		
α	U-mAP	U-Recall	mAP	U-mAP	U-Recall	mAP
0.5	5.2	65.8	66.6	12.5	70.5	72.9
0.55	5.3	66.0	66.6	12.6	70.7	72.9
0.6	5.3	66.0	66.6	12.6	70.7	72.9
0.65	5.3	65.9	66.6	12.7	70.6	72.9
0.7	5.3	66.0	66.6	12.6	70.7	72.9
0.75	5.3	66.0	66.6	12.8	70.8	72.9
0.8	5.3	66.0	66.6	12.6	70.8	72.9
0.85	5.2	66.0	66.6	12.7	70.8	72.9
0.9	5.2	65.9	66.6	12.8	70.6	72.9
0.95	5.1	65.4	66.6	12.8	70.7	72.9

Table 11: **Effect of the matching threshold α under incremental learning**.

Setting	M-OWODB			S-OWODB		
α	U-mAP	U-Recall	mAP	U-mAP	U-Recall	mAP
0.5	3.5	64.8	67.7	8.1	74.8	72.6
0.55	3.7	66.7	67.7	8.5	74.8	72.6
0.6	3.7	66.7	67.7	8.5	74.8	72.6
0.65	3.8	66.6	67.7	8.4	74.8	72.6
0.7	3.8	66.7	67.8	8.3	75.6	72.6
0.75	3.8	66.7	67.8	8.3	75.6	72.6
0.8	3.8	66.7	67.7	7.8	75.7	72.6
0.85	3.8	66.7	67.7	7.8	75.7	72.6
0.9	3.8	66.7	67.7	7.8	75.7	72.6
0.95	3.9	66.7	67.7	7.8	76.5	72.6

Table 12: **Effect of the JSD–diversity balance β**.

Setting	M-OWODB			S-OWODB		
β	U-mAP	U-Recall	mAP	U-mAP	U-Recall	mAP
0.1	4.8	64.8	66.6	9.6	68.1	72.6
0.2	5.5	66.7	66.6	11.0	68.0	72.6
0.3	5.4	66.1	66.6	11.9	70.0	72.9
0.4	5.5	65.8	66.6	12.8	70.8	72.8
0.5	5.4	66.1	66.6	12.9	70.9	72.9
0.6	5.3	66.0	66.6	12.8	70.8	72.9
0.7	5.2	66.0	66.6	12.5	70.8	72.8
0.8	5.3	65.9	66.6	12.0	71.1	72.9
0.9	5.3	65.9	66.6	12.0	71.0	72.8
1.0	5.3	65.8	66.6	12.0	71.0	72.8

Table 13: **Effect of reliability thresholds τ_f and τ_μ**.

Setting		M-OWODB			S-OWODB		
τ_f	τ_μ	U-mAP	U-Recall	mAP	U-mAP	U-Recall	mAP
-1	-1	0.0	17.4	56.5	0.0	15.3	61.6
-4	-1	0.0	18.1	59.1	0.1	16.3	63.8
-7	-1	0.0	19.1	59.1	0.1	16.3	63.8
-1	-2	0.1	29.2	59.6	0.1	20.4	64.6
-4	-2	5.2	60.8	66.4	8.2	66.6	72.7
-7	-2	5.3	60.9	66.4	8.3	66.6	72.7
-1	-3	0.1	28.2	59.6	0.0	2.4	64.6
-4	-3	5.3	66.0	66.6	12.8	70.8	72.9
-7	-3	5.3	66.0	66.6	12.8	70.8	72.9
$-\infty$	$-\infty$	5.2	65.8	66.6	12.6	70.8	72.8

Table 14: **Effect of the SAE expansion ratio d_{SAE}**.

Setting	M-OWODB			S-OWODB		
d_{SAE}	U-mAP	U-Recall	mAP	U-mAP	U-Recall	mAP
2x	5.2	63.5	66.7	9.5	69.2	72.9
4x	5.3	66.0	66.6	12.8	70.8	72.9
8x	6.9	71.2	66.6	9.6	68.2	72.7
16x	3.8	70.5	66.7	7.8	71.9	72.7

Table 15: **Effect of the SAE activation sparsity s**.

Setting	M-OWODB			S-OWODB		
s (%)	U-mAP	U-Recall	mAP	U-mAP	U-Recall	mAP
25	5.1	63.3	66.6	8.9	62.9	71.2
5	5.1	66.1	66.5	9.9	69.6	71.4
0.5	5.3	66.0	66.6	12.8	70.8	72.9
0.1	4.6	56.7	66.5	9.7	68.3	71.5

Table 16: **Effect of aggregation rules for selected latent directions**.

Setting	M-OWODB			S-OWODB		
Aggr.	U-mAP	U-Recall	mAP	U-mAP	U-Recall	mAP
Bin Gate	0.4	50.0	57.7	0.4	43.0	62.4
Max	3.2	67.0	67.6	9.7	67.8	72.4
Mean	3.8	67.6	67.4	10.3	69.7	72.3
ours	5.3	66.0	66.6	12.8	70.8	72.9

Table 17: **Effect of the number k of selected latent direct**.

Setting	M-OWODB			S-OWODB		
k	U-mAP	U-Recall	mAP	U-mAP	U-Recall	mAP
25	5.1	67.5	61.2	12.6	71.2	67.0
50	5.3	66.4	64.1	12.8	71.1	70.8
75	5.3	66.5	64.1	12.8	71.0	72.2
100	5.3	66.0	66.6	12.8	70.8	72.9

Incremental Matching. Table 11 evaluates α as the vocabulary grows and visually similar classes increase boundary ambiguity. U-Recall changes moderately, whereas U-mAP is more sensitive because loose or strict matching includes unrepresentative known embeddings or removes useful selected directions. On M-OWODB, U-mAP increases from 3.5 at $\alpha = 0.5$ to 3.9 at $\alpha = 0.95$, while U-Recall remains between 64.8 and 66.7; S-OWODB maintains stable mAP throughout the sweep. Its increase to 76.5 U-Recall at $\alpha = 0.95$ occurs with unchanged mAP of 72.6, reflecting a different precision–recall trade-off. Overall, the broad plateau indicates that the method is not sensitive to the precise choice of α. We retain the benchmark-specific defaults throughout incremental learning.

Table 18: **Robustness to SAE training seeds on M-OWODB and S-OWODB Task 1**. Each run independently trains the SAE and reselects the latent directions, while keeping the frozen region embeddings, data split, hyperparameters, and inference settings fixed.

Seed	M-OWODB Task 1			S-OWODB Task 1		
	U-mAP (↑)	U-Recall (↑)	mAP (↑)	U-mAP (↑)	U-Recall (↑)	mAP (↑)
42	5.3	66.0	66.6	12.8	70.8	72.9
5407	5.4	66.9	66.7	13.1	70.3	71.6
6407	5.7	65.0	66.6	12.6	70.5	71.6
Mean ± std	**5.5 ± 0.2**	**66.0 ± 1.0**	**66.6 ± 0.1**	**12.8 ± 0.3**	**70.5 ± 0.3**	**72.0 ± 0.8**

B.2 Factor Selection and Functional Validation

JSD–diversity Balance β. The parameter β in Eq. (12) balances the JSD term against redundancy among selected decoder directions. Table 12 shows that $\beta = 0.1$ places greater emphasis on decoder diversity and selects directions with less similar activation distributions between the known-aligned and unmatched sets. At larger values, U-Recall remains stable while U-mAP improves, especially on S-OWODB. M-OWODB remains within 64.8–66.0 U-Recall and S-OWODB within 68.0–71.1. S-OWODB U-mAP rises from 9.6 at $\beta = 0.1$ to 12.8 at the default $\beta = 0.6$. The highest values leave known-class mAP unchanged and yield marginal U-mAP gains. Decoder diversity limits repeated selection of geometrically similar directions and preserves coverage of diverse latent directions. The plateau beyond 0.6 indicates diminishing returns from greater JSD weight.

Reliability Thresholds τ_f and τ_μ. The thresholds τ_f and τ_μ in Eq. (10) filter directions by activation frequency and strength; in Table 13, -1, -4, and -7 denote 10^{-1}, 10^{-4}, and 10^{-7}. Strict thresholds remove useful directions and collapse U-mAP and U-Recall. At the most restrictive pairs, M-OWODB U-Recall falls below 30 and S-OWODB below 21, confirming a loss of eligible latent directions. In the feasible region, several settings recover 66.0/70.8 U-Recall on M-OWODB/S-OWODB. The competitive unfiltered row shows the contribution of the JSD–diversity objective; reliability filtering removes rare or negligible activations before ranking. The chosen thresholds lie on a broad plateau and tolerate small estimation errors in activation statistics. Relative to no filtering, $\tau_f = 10^{-7}$ and $\tau_\mu = 10^{-3}$ preserve M-OWODB U-Recall while improving S-OWODB mAP and maintaining strong U-mAP. We use these values.

Feature-space Removal and Recovery. We next test the functional contribution of selected directions to unknown assignment. Fig. 6 removes the latent components associated with top-ranked directions directly from each feature representation and restores them in ranked order. Equal-sized random and bottom-ranked latent sets provide comparison interventions. Removing selected components produces the largest and most consistent degradation in U-mAP and U-Recall across the evaluated benchmark-task settings, while restoring a fraction of the selected set recovers performance rapidly. Random and bottom-ranked components produce weaker and less systematic responses. These interventions show that the ranking criterion identifies a feature subspace with a functional role in sparse factor scoring and unknown detection.

B.3 Dictionary Capacity and Activation Sparsity

SAE Expansion Ratio d_{SAE}. Table 14 analyzes sparse latent-space size. The expansion ratio controls the capacity and granularity of the SAE dictionary. A small ratio can leave multiple feature directions entangled within one coordinate (Gao et al., 2025; Chanin et al., 2025); a larger ratio provides finer decomposition and may improve unknown recall. Excessive dimensionality can induce feature splitting across specialized latents (Chanin et al., 2026; Bussmann et al., 2025). In our experiments, the $16\times$ setting retains high U-Recall but reduces U-mAP, particularly on S-OWODB, showing that additional dimensions do not ensure higher unknown precision. The $2\times$ setting lowers recall on both datasets, while $8\times$ and $16\times$ raise recall without improving known-class mAP or unknown precision. A larger dictionary may capture more unmatched structure while dispersing objectness responses across correlated coordinates. The $4\times$ setting provides the most balanced cross-dataset behavior.

Activation Sparsity s. Tables 14 and 15 isolate complementary representation properties: the former changes dictionary capacity, while the latter fixes the default $4\times$ dictionary ($d_{\mathrm{SAE}} = 2048$)

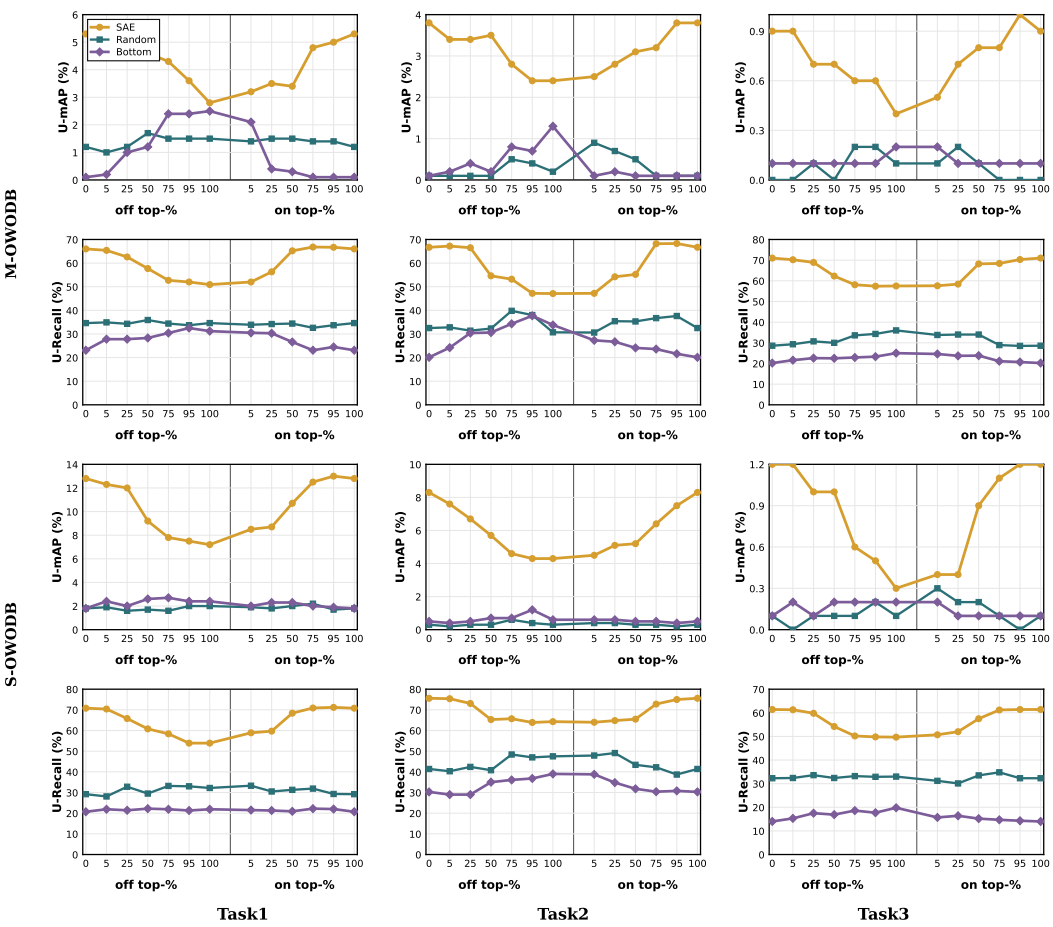


Figure 6: **Feature-space removal and recovery of SAE directions on M-OWODB and S-OWODB Tasks 1–3**. Curves compare the top-100 selected (SAE:100), independently sampled random-100, and bottom-100 ranked (SAE:-100) directions using U-mAP and U-Recall. Random controls use the independent seeds of their corresponding intervention runs; "off" and "on" denote removal from and restoration to the feature representation.

and varies activation sparsity s. The intermediate $s = 0.5\%$ setting gives the best cross-dataset balance. Relaxing sparsity to 25% produces less selective activations, reducing S-OWODB U-mAP/U-Recall/mAP from 12.8/70.8/72.9 to 8.9/62.9/71.2. Excessive sparsity at 0.1% removes useful distributed activations, lowering M-OWODB U-Recall from 66.0 to 56.7 and S-OWODB U-Recall from 70.8 to 68.3. Dense activations admit broad background and weak responses; extreme sparsity discards cues that contribute jointly to the objectness score. The 0.5% setting preserves distributed support for recall with a compact active set. Additional capacity and sparsity require balance within a fixed dictionary. This optimum agrees with SAE findings on the trade-off between activation sparsity and reconstruction fidelity (Cunningham et al., 2024; Gao et al., 2025).

B.4 Scoring Aggregation and Selected-Set Size

Aggregation of Selected Directions. Table 16 compares alternative mappings from selected activations to sparse factor objectness. Bin Gate sets $P_{\text{obj}} = 0.5$ whenever at least one selected factor is active and 0 otherwise. The binary activation gate discards response magnitude and yields only 0.4/0.4 U-mAP on M-OWODB/S-OWODB, while collapsing known-class mAP to 57.7/62.4. Thus, the population-level activation-rate differences observed in Fig. 3(a) are informative but insufficient as a binary proposal score. Max and mean aggregation recover reasonable recall and known-class performance, but their best U-mAP values remain 3.8 and 10.3. The proposed softmax-weighted

average raises U-mAP to 5.3 and 12.8, improvements of 1.5 and 2.5 points over the strongest alternative, while retaining a competitive recall–mAP balance.

Number of Selected Latent Directions k. Table 17 studies the size of the selected direction set. Across $k = 25$–100, U-mAP varies only from 5.1 to 5.3 on M-OWODB and from 12.6 to 12.8 on S-OWODB, while U-Recall remains within 66.0–67.5 and 70.8–71.2 respectively. The main sensitivity appears in known-class mAP, which increases from 61.2 to 66.6 on M-OWODB and from 67.0 to 72.9 on S-OWODB. Thus, the slightly higher U-Recall obtained with a small selected set comes with worse known-class performance, consistent with excessive assignment to the unknown class rather than better known–unknown separation. Increasing k preserves unknown-prediction ranking while reducing interference with the known-class pathway.

B.5 Robustness to SAE Training Seeds

We evaluate whether downstream detection performance is sensitive to stochastic SAE training and latent reselection. For each seed, we independently train the SAE and reselect the latent directions, while keeping the frozen region embeddings, data split, hyperparameters, and inference settings fixed.

Seed Robustness. Table 18 shows limited run-to-run variation across three trained SAEs. On M-OWODB, the standard deviations are 0.2 U-mAP, 1.0 U-Recall, and 0.1 mAP; on S-OWODB, they are 0.3, 0.3, and 0.8 respectively. The mean results also remain close to the seed-42 reference run. Although S-OWODB exhibits slightly greater variation in known-class mAP, unknown-prediction ranking and unknown recall remain stable. These results indicate that, under fixed detector and evaluation settings, the observed performance is not driven by a single favorable SAE initialization: retraining the SAE and reselecting the latent directions preserve the overall known–unknown performance balance.

Overall Sensitivity. The hyperparameter analysis supports the motivation for reporting U-mAP. With a strong frozen open-vocabulary detector, unknown recall is often close to the achievable coverage provided by the region proposals and is therefore relatively insensitive to moderate parameter changes. The more important question is whether the sparse factor objectness branch ranks ground-truth unknown detections above background proposals and known-object confusions. U-mAP captures this precision-oriented effect, while U-Recall alone can obscure it. Across Tables 10–18, moderate changes preserve most proposal-level recall, and settings that alter the quality or selectivity of latent evidence produce clearer movement in U-mAP. The shared configuration exposes a broad cross-dataset operating region, and the surrounding sweeps characterize its stability under protocol changes.

C Qualitative Results and Semantic Audit

This section complements the quantitative evaluation with detection examples and a post-hoc audit of selected factors. §C.1 presents additional detection results on both OWODB protocols. §C.2 audits factor semantics across datasets and incremental tasks.

C.1 Additional Detection Examples

Figs. 7 and 8 extend the qualitative comparison in Fig. 5 to both OWODB protocols and three representative baselines. Each column pairs the baseline prediction in the upper row with the corresponding SAE-OVOW result in the lower row. On M-OWODB, SAE-OVOW produces more localized hypotheses for bananas, desktop objects, and mobile phones while retaining surrounding known detections such as bottles, monitors, and human regions. On S-OWODB, the bathroom, desktop, and beach scenes show that the task-specific sparse factor objectness remains effective under the superclass-separated split, including for small, overlapping, and contextually diverse regions. Across both benchmarks, the visualizations illustrate how SAE-OVOW reduces coarse or merged unknown responses while exposing the factor most strongly associated with each prediction.

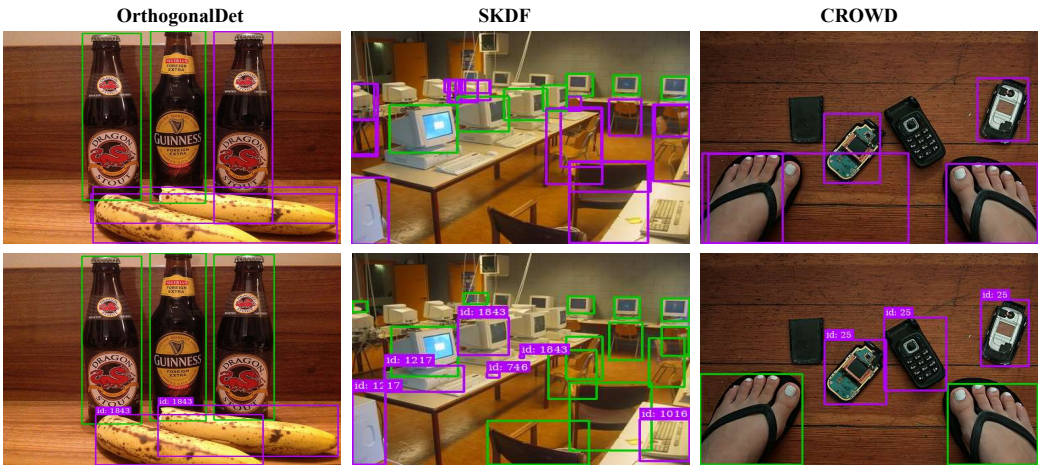


Figure 7: **Additional qualitative comparisons on M-OWODB Task 1**. OrthogonalDet, SKDF, and CROWD predictions appear above the corresponding SAE-OVOW results. Purple and green boxes denote unknown and known predictions.

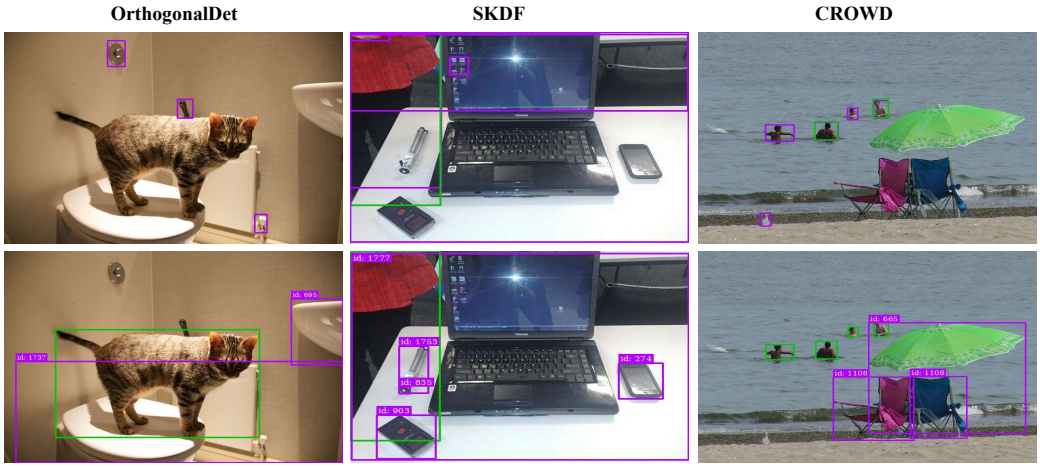


Figure 8: **Additional qualitative comparisons on S-OWODB Task 1**. OrthogonalDet, SKDF, and CROWD predictions appear above the corresponding SAE-OVOW results. Purple and green boxes denote unknown and known predictions.

C.2 Task-1 Semantic Audits on Two Benchmarks.

Audit Protocol. Following established top-activation analysis protocols (Cunningham et al., 2024; Lim et al., 2025; Pach et al., 2025), we inspect the 15 highest-activation region crops for each selected factor. The audit begins after factor selection has been completed, and semantic descriptions remain confined to this analysis. A factor receives a human-readable description when its crops exhibit a coherent recurring visual pattern; the remaining cases are recorded as unresolved. Table 19 reports the ten highest-ranked factors from M-OWODB and S-OWODB Task 1. Figs. 9 and 10 visualize the top five factors from the two benchmarks, including their activation distributions over known-aligned and unmatched regions and the corresponding top-activating crops.

Cross-dataset Semantic Audit. Within each benchmark, four of the five displayed task-local factors exhibit coherent crop-level patterns. M-OWODB factors emphasize vehicle rear panels, furniture structures, framed parts, and local textures. S-OWODB factors emphasize sports or crowd scenes, human regions, collars, and vehicle appearances. Across the ten highest-ranked factors from each benchmark, 14 of 20 receive descriptions. The complete Task 1 summary in Table 7 records 119 coherent factors among 200 audited factors. These observations indicate that selected directions

Table 19: **Top-ten selected factors on M-OWODB and S-OWODB Task 1**. A factor receives a human-readable description when its top-15 activation regions exhibit a coherent recurring pattern; remaining factors are marked unresolved.

Dataset	Factor_id	JSD	$\mathcal{R}$	Score	Rank	Top-Activation Interpretation / Status
M-OWODB	85	0.002	0.000	0.001	1	Unresolved factor
	773	0.007	0.475	0.194	2	Coherent pattern: Vehicle rear panels and rounded structural parts
	1602	0.009	0.481	0.198	3	Coherent pattern: Chair backs and narrow furniture frames
	1646	0.007	0.490	0.200	4	Coherent pattern: Circular openings and framed display surfaces
	743	0.007	0.492	0.201	5	Coherent pattern: Fur, feather, and patterned surface textures
	352	0.007	0.495	0.202	6	Coherent pattern: Vehicle bodies and glossy rectangular panels
	75	0.005	0.496	0.202	7	Unresolved factor
	35	0.005	0.496	0.202	8	Coherent pattern: Bird and poultry body shapes
	34	0.005	0.498	0.202	9	Coherent pattern: Bird and small-animal body shapes
	23	0.003	0.502	0.202	10	Unresolved factor
S-OWODB	2000	0.004	0.000	0.003	1	Coherent pattern: Outdoor sports fields and crowd scenes
	1108	0.015	0.470	0.197	2	Unresolved factor
	1282	0.007	0.489	0.200	3	Coherent pattern: Standing people and torso regions
	1411	0.013	0.481	0.200	4	Coherent pattern: Neckties and shirt collars
	838	0.009	0.490	0.202	5	Coherent pattern: Cars and vehicle body regions
	444	0.008	0.494	0.202	6	Unresolved factor
	702	0.007	0.499	0.204	7	Coherent pattern: Baseball players and stadium fields
	939	0.007	0.498	0.203	8	Coherent pattern: Large transportation vehicles
	323	0.007	0.496	0.203	9	Unresolved factor
	1807	0.007	0.496	0.203	10	Coherent pattern: Bicycle symbols and signage

capture heterogeneous visual structure spanning object parts, textures, appearances, and contextual regularities.

Audit scope. The descriptions in Table 19 provide local annotations for task-specific factors learned independently on each benchmark. Figs. 9–10 expose the activation distributions and top-activating crops supporting these annotations,the complete rank-ordered audit of all 200 selected factors, including unresolved factors and their top-activation regions, is provided in the supplementary material. SAE-OVOW computes sparse factor objectness from the complete numerical activation set, and the tables display ranked subsets for inspection. Unresolved status records limited crop-level coherence for a stable description. Functional relevance is evaluated through the masking and recovery experiments in Fig. 6.

D Additional Incremental Evaluation

This section evaluates SAE-OVOW beyond the primary OWODB tables and visualizes how predictions evolve as the known vocabulary expands. §D.1 reports single-step incremental object detection results. §D.2 visualizes task-wise behavior on nu-OWODB.

D.1 IOD Benchmark Results

IOD Benchmark Comparison. Single-step incremental object detection can be viewed as a simplified OWOD setting, where previously unknown classes are introduced once as new known classes. Table 20 reports the results on PASCAL VOC 2007 under the 10+10, 15+5, and 19+1 settings. SAE-OVOW achieves the best mAP in all three settings, reaching 87.9, 87.4, and 87.0 respectively. Compared with OWOBJ, our method improves mAP by 18.0, 14.1, and 11.2 points, and outperforms CROWD and IPOW by more than 12 points across all settings.

Incremental Adaptation. These results are consistent with the design of SAE-OVOW. The method expands the OVD vocabulary and retrains the SAE-based objectness branch for each task. Newly introduced classes are recognized through the expanded vocabulary, while the reselected latent directions provide class-agnostic objectness for detecting objects outside that vocabulary. Stable performance across incremental splits shows that the formulation preserves OVD zero-shot capability and supports open-world class expansion.

Table 20: Comparison with state-of-the-art methods in **single-step incremental object detection** settings on PASCAL VOC 2007. The 10, 5 and 1 classes in gray background are introduced to a detector trained on the remaining 10, 15 and 19 classes, respectively.

10 + 10 setting	aero	cycle	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	bike	person	plant	sheep	sofa	train	tv	mAP
ORE (Joseph et al., 2021)	63.5	70.9	58.9	42.9	34.1	76.2	80.7	76.3	34.1	66.1	56.1	70.4	80.2	72.3	81.8	42.7	71.6	68.1	77.0	67.7	64.5
OW-DETR (Gupta et al., 2022)	61.8	69.1	67.8	45.8	47.3	78.3	78.4	78.6	36.2	71.5	57.5	75.3	76.2	77.4	79.5	40.1	66.8	66.3	75.6	64.1	65.7
PROB (Zohar et al., 2023b)	70.4	75.4	67.3	48.1	55.9	73.5	78.5	75.4	42.8	72.2	64.2	73.8	76.0	74.8	75.3	40.2	66.2	73.3	64.4	64.0	66.5
CAT (Ma et al., 2023)	76.5	75.7	67.0	51.0	62.4	73.2	82.3	83.7	42.7	64.4	56.8	74.1	75.8	79.2	78.1	39.9	65.1	59.6	78.4	67.4	67.7
OrthogonalDet (Sun et al., 2024)	82.9	80.1	75.8	64.3	60.6	81.5	87.9	54.9	48	82.1	57.7	63.5	80.5	77.6	78.2	38.9	69.8	62.8	76.9	64.2	69.4
OWOBJ (Zhang et al., 2025)	75.9	80.7	73.3	52.1	58.8	77.7	81.9	79.1	47.9	77.8	70.4	78.9	80.3	79.3	80.0	45.1	70.2	78.4	68.5	68.4	69.9
CROWD (Majee et al., 2025)	84.1	84.5	73.9	60.0	65.1	80.1	89.3	82.7	53.3	77.4	63.4	78.5	80.9	83.4	83.9	46.5	72.6	60.9	77.9	71.5	73.5
IPOW (Lv et al., 2026)	79.8	83.7	78.0	63.1	65.1	81.3	87.0	84.8	62.0	71.3	64.9	82.5	79.0	79.1	83.8	48.6	66.0	67.8	74.9	74.2	73.8
SAE-OVOW (ours)	97.5	91.5	92.3	75.6	76.8	95.4	95.3	92.4	81.6	87.0	83.8	89.6	94.9	96.6	95.7	54.9	87.7	85.1	94.2	89.3	**87.9**
15 + 5 setting	aero	cycle	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	bike	person	plant	sheep	sofa	train	tv	mAP
ORE (Joseph et al., 2021)	75.4	81.0	67.1	51.9	55.7	77.2	85.6	81.7	46.1	76.2	55.4	76.7	86.2	78.5	82.1	32.8	63.6	54.7	77.7	64.6	68.5
OW-DETR (Gupta et al., 2022)	77.1	76.5	69.2	51.3	61.3	79.8	84.2	81.0	49.7	79.6	58.1	79.0	83.1	67.8	85.4	33.2	65.1	62.0	73.9	65.0	69.4
PROB (Zohar et al., 2023b)	77.9	77.0	77.5	56.7	63.9	75.0	85.5	82.3	50.0	78.5	63.1	75.8	80.0	78.3	77.2	38.4	69.8	57.1	73.7	64.9	70.1
CAT (Ma et al., 2023)	75.3	81.0	84.4	64.5	56.6	74.4	84.1	86.6	53.0	70.1	72.4	83.4	85.5	81.6	81.0	32.0	58.6	60.7	81.6	63.5	72.2
OrthogonalDet (Sun et al., 2024)	81.8	79.3	71.0	71.0	58.8	62.1	82.6	89.7	79.8	47.0	80.5	61.1	79.9	80.2	81.6	44.2	65.5	71.5	75.6	74.2	72.6
OWOBJ (Zhang et al., 2025)	82.8	80.0	82.4	60.1	68.0	79.9	90.0	86.4	54.4	83.1	64.2	77.3	85.1	80.3	80.1	42.1	73.2	61.8	77.9	68.7	73.3
CROWD (Majee et al., 2025)	82.8	80.6	72.5	59.6	61.3	83.1	89.3	83	49.2	86.1	62.2	83.7	86	80.3	82.8	46.1	80	63.7	79.5	75.6	74.4
IPOW (Lv et al., 2026)	77.9	79.5	77.0	65.7	63.0	82.2	86.2	87.7	56.6	82.8	68.9	86.7	85.9	79.6	84.8	44.6	71.7	68.6	74.6	75.2	75.0
SAE-OVOW (ours)	96.9	91.0	92.0	75.3	76.4	95.4	95.1	92.1	81.2	86.3	83.2	88.7	94.6	95.8	95.2	54.5	87.5	84.7	93.8	88.8	**87.4**
19 + 1 setting	aero	cycle	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	bike	person	plant	sheep	sofa	train	tv	mAP
ORE (Joseph et al., 2021)	67.3	76.8	60	48.4	58.8	81.1	86.5	75.8	41.5	79.6	54.6	72.8	85.9	81.7	82.4	44.8	75.8	68.2	75.7	60.1	68.8
OW-DETR (Gupta et al., 2022)	70.5	77.2	73.8	54.0	55.6	79.0	80.8	80.6	43.2	80.4	53.5	77.5	89.5	82.0	74.7	43.3	71.9	66.6	79.4	62.0	70.2
PROB (Zohar et al., 2023b)	80.3	78.9	77.6	59.7	63.7	75.2	86.0	83.9	53.7	82.8	66.5	82.7	80.6	83.8	77.9	48.9	74.5	69.9	77.6	48.5	72.6
CAT (Ma et al., 2023)	86.0	85.8	78.8	65.3	61.3	71.4	84.8	84.8	52.9	78.4	71.6	82.7	83.8	81.2	80.7	43.7	75.9	58.5	85.2	61.1	73.8
OrthogonalDet (Sun et al., 2024)	81.8	82.6	77.0	56.3	66.0	74.4	88.5	78.7	51.2	84.3	63.1	84.4	81.3	78.8	80.9	46.8	77.9	68.6	74.1	74.5	73.6
OWOBJ (Zhang et al., 2025)	86.1	83.9	83.4	62.9	65.9	79.9	90.6	87.3	56.9	86.5	70.3	85.9	84.7	86.9	81.6	51.9	78.8	73.5	80.7	53.8	75.8
CROWD (Majee et al., 2025)	81.7	80.3	77.4	57.2	66.8	80.7	87.1	67.9	49.4	87.3	65.6	84.2	85.4	79.9	81.6	48.6	77.0	69.0	82.2	75.3	74.2
IPOW (Lv et al., 2026)	81.1	78.7	78.2	56.3	63.5	77.2	86.2	86.8	59.5	79.8	64.7	86.3	83.0	78.8	83.2	49.3	74.4	71.6	81.0	71.0	74.5
SAE-OVOW (ours)	96.5	91.0	91.1	74.9	76.0	95.2	95.0	91.6	80.5	86.2	82.2	88.5	93.8	95.1	95.0	53.7	86.8	83.8	93.6	88.6	**87.0**

D.2 Incremental Learning Visualization on nu-OWODB

Task-wise Visualization. Fig. 11 tracks the same driving scene across the three nu-OWODB tasks. In Task 1, OVOW assigns vehicle fragments to unknown and misses several pedestrians and traffic cones; SAE-OVOW reduces this confusion and recalls more unknown objects. When pedestrian classes enter the vocabulary in Task 2, SAE-OVOW assigns them through the expanded known-class pathway and continues to discover traffic cones. In Task 3, the previously unknown pedestrians and traffic cones appear as known predictions together with the vehicle and bicycles. Vocabulary expansion, task-specific SAE retraining, and direction reselection adapt sparse factor objectness to each known–unknown boundary. Equation (16) then incorporates the scores of the expanded vocabulary.

E Algorithms

This section gives executable summaries of the two stages used after task-specific SAE training. §E.1 specifies JSD-based latent-direction selection. §E.2 specifies sparse factor objectness and unknown scoring.

E.1 JSD-Based Latent-Direction Selection

JSD-based Latent-direction Selection. Algorithm 1 compares known-aligned and unmatched activation distributions after task-specific SAE training. Activation-frequency and strength constraints exclude rarely active directions; JSD measures distribution discrepancy, while decoder-vector redundancy expands geometric coverage. The algorithm adds the lowest-scoring candidates until $|\hat{\mathcal{L}}| = k$ and fixes the resulting set before test-time scoring.

E.2 Sparse Factor Objectness and Unknown Scoring

Sparse Factor Objectness for Unknown Detection. Algorithm 2 combines frozen-OVD known-class scores P_C with sparse SAE coordinates. It aggregates positive selected activations into P_{obj} and weights them by known-class uncertainty, favoring strong sparse factor objectness with weak known-prompt alignment.

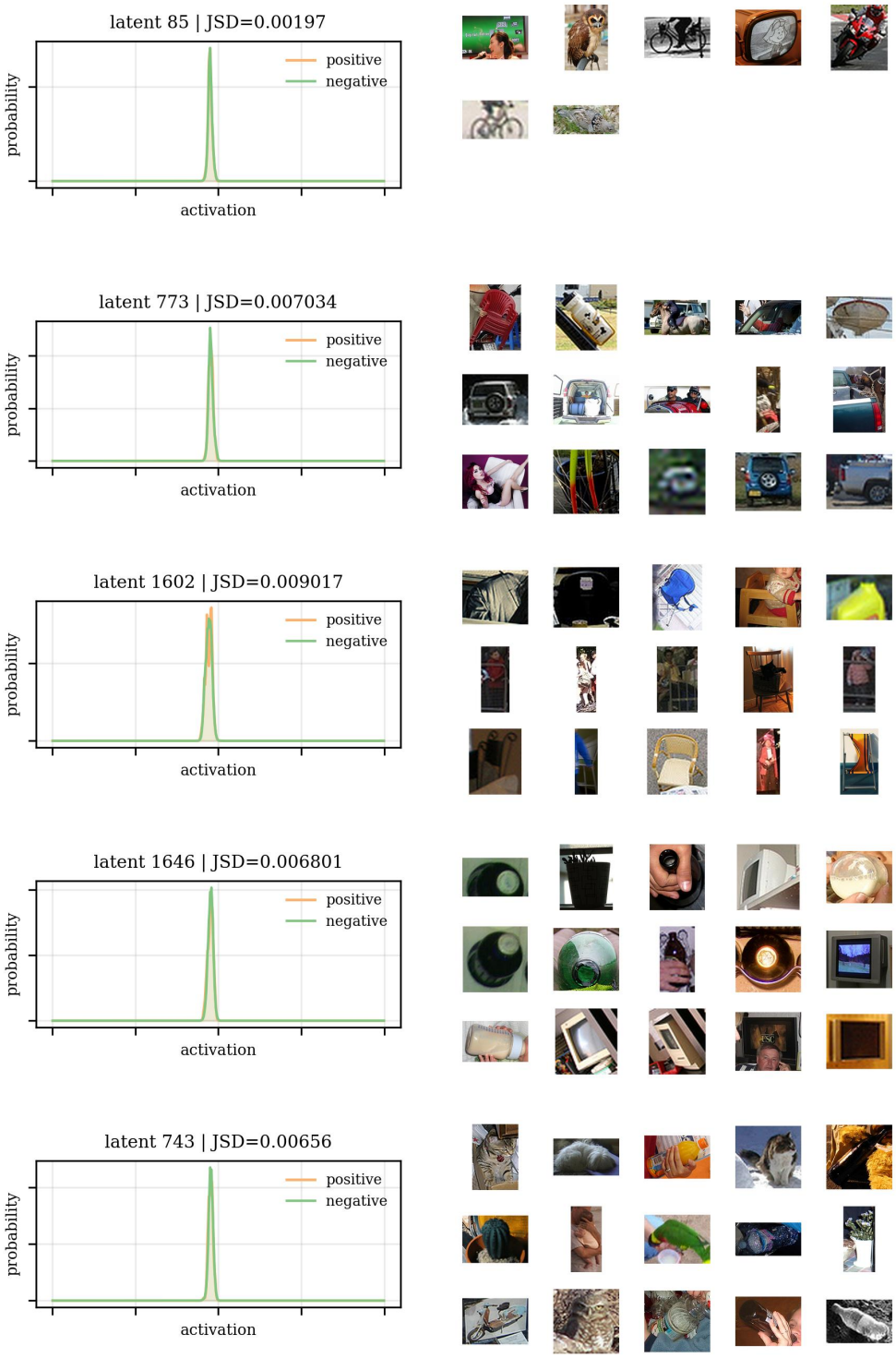


Figure 9: **Top-five selected factors on M-OWODB Task 1**. Each row compares activation distributions over known-aligned and unmatched regions and presents the 15 highest-activation crops used for the semantic audit in Table 19.

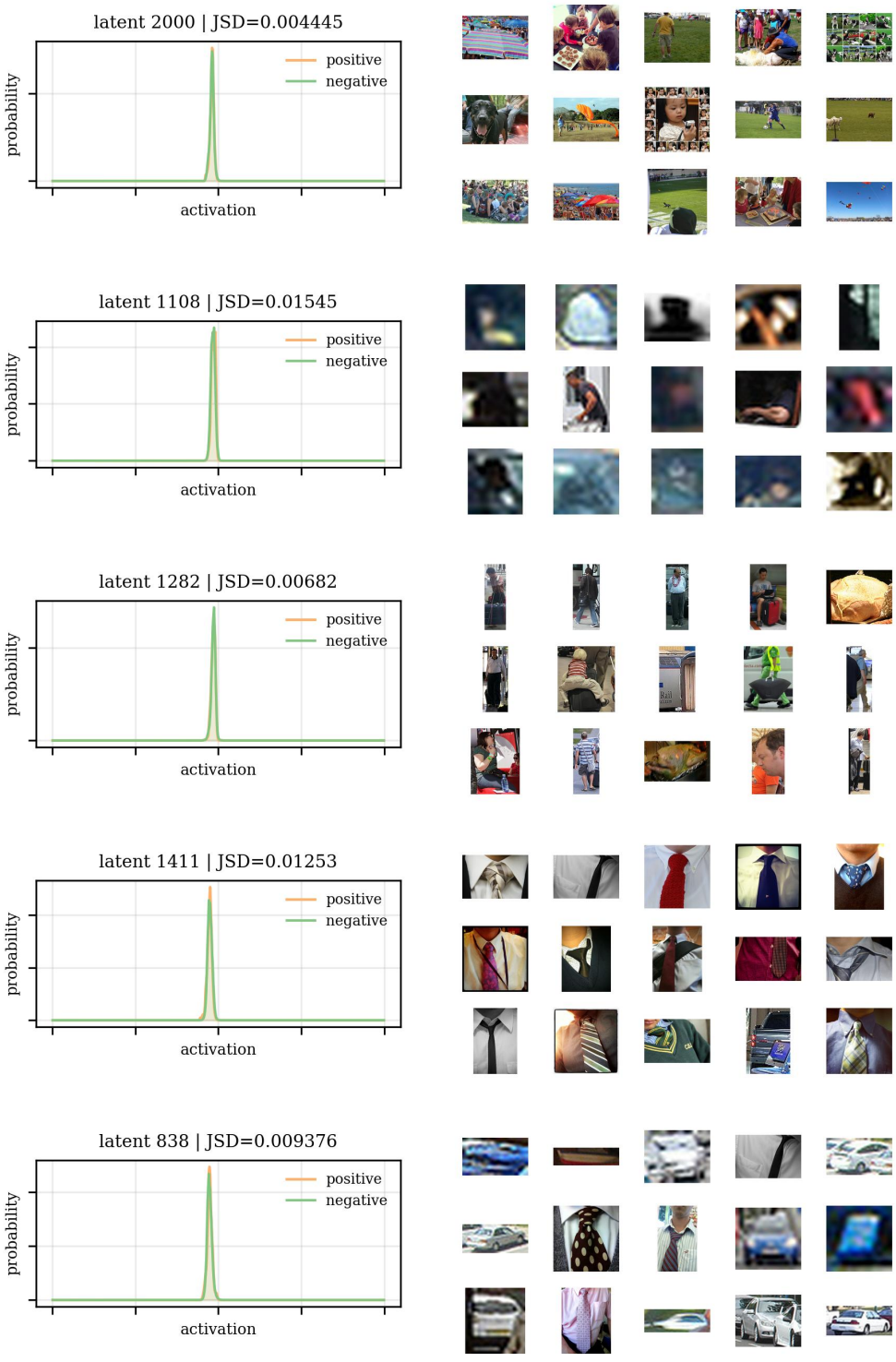


Figure 10: **Top-five selected factors on S-OWODB Task 1**. Each row compares activation distributions over known-aligned and unmatched regions and presents the 15 highest-activation crops used for the semantic audit in Table 19.



Figure 11: **Task-wise qualitative comparison on nu-OWODB**. Rows compare OVOW (top) and SAE-OVOW (bottom), and columns correspond to Tasks 1–3 for the same scene. Tasks 1 and 2 display only predictions assigned to the unknown class; Task 3 displays all predictions after all classes have been introduced. Purple and green boxes denote unknown and known predictions respectively. The visualization traces the discovery of pedestrians and traffic cones in early tasks and their recognition through the expanded known-class vocabulary in later tasks.

Algorithm 1 JSD-based latent-direction selection

Require: Known-aligned embeddings E^{+}_{vis}, unmatched embeddings E^{-}_{vis}; SAE encoder $(W_E, \mathbf{b}_E)$ and decoder dictionary W_D; thresholds τ_f, τ_μ; trade-off β; selected direction-set size k
Ensure: Selected latent-direction set $\hat{\mathcal{L}}$
$\mathcal{C} \leftarrow \emptyset$, $\hat{\mathcal{L}} \leftarrow \emptyset$
for each SAE direction l_i **do**
  Compute bounded responses q_i^+ on E^{+}_{vis} and q_i^- on E^{-}_{vis}
  Estimate normalized histogram distributions D_i^+ and D_i^-
  Compute $\text{JSD}(D_i^+, D_i^-)$
  Compute activation frequencies f_i^+, f_i^- and activated strengths μ_i^+, μ_i^-
  if $\sqrt{f_i^+ f_i^-} \geq \tau_f$ and $\sqrt{\mu_i^+ \mu_i^-} \geq \tau_\mu$ **then**
    $\mathcal{C} \leftarrow \mathcal{C} \cup \{l_i\}$
  end if
end for
while $|\hat{\mathcal{L}}| < k$ and $\mathcal{C} \setminus \hat{\mathcal{L}} \neq \emptyset$ **do**
  for each $l_i \in \mathcal{C} \setminus \hat{\mathcal{L}}$ **do**
    Compute redundancy $\mathcal{R}(l_i, \hat{\mathcal{L}})$ using decoder-vector similarity
    $s_i \leftarrow \beta\, \text{JSD}(D_i^+, D_i^-) + (1-\beta)\mathcal{R}(l_i, \hat{\mathcal{L}})$
  end for
  $l^\star \leftarrow \arg\min_{l_i \in \mathcal{C} \setminus \hat{\mathcal{L}}} s_i$
  $\hat{\mathcal{L}} \leftarrow \hat{\mathcal{L}} \cup \{l^\star\}$
end while
return $\hat{\mathcal{L}}$

Algorithm 2 Sparse factor objectness for unknown detection

Require: Candidate region embedding e_{vis}; known text embeddings E_C; SAE encoder $(W_E, \mathbf{b}_E)$; selected direction set $\hat{\mathcal{L}}$ with $|\hat{\mathcal{L}}| = k$
Ensure: Known-class scores P_C and unknown score P_{unk}
Compute known-class scores $P_C = h_{\text{cls}}(e_{\text{vis}}, E_C)$
Compute sparse latent activation $h = \text{ReLU}(W_E^\top e_{\text{vis}} + \mathbf{b}_E)$
$\mathcal{A} \leftarrow \{j \in \hat{\mathcal{L}} : h_j > 0\}$
Compute sparse factor objectness from the active selected directions:

$$P_{\text{obj}} = \begin{cases} \displaystyle\sum_{j \in \mathcal{A}} \frac{\exp(\sigma(h_j))}{\sum_{m \in \mathcal{A}} \exp(\sigma(h_m))} \sigma(h_j), & |\mathcal{A}| > 0, \\ 0, & |\mathcal{A}| = 0. \end{cases}$$

Compute known-class uncertainty $P_{\text{unc}} = 1 - max(P_C)$
$P_{\text{unk}} \leftarrow P_{\text{obj}} \cdot P_{\text{unc}}$
return P_C, P_{unk}