Appendix A: Proof for Theorem 1

Based on the definition of the target label space $\mathcal{Y}^t = \mathcal{Y}^s \cup \mathcal{Y}^u$, the total target risk can be decomposed as,

$$\mathcal{R}'(h) = (1-\pi^u)\cdot\mathcal{R}^{known}(h) + \pi^u \cdot \mathcal{R}^{OS}(h) \quad (1)$$

where $\mathcal{R}^{known}(h)$ and $\mathcal{R}^{OS}(h)$ (Fang et al., 2020) denote the target risks from known and unknown classes, respectively. π^u is the prior probability of unknown classes in the target domain. From the DG error bound in (Wang et al., 2022), the known-class risk satisfies,

$$\mathcal{R}^{known}(h) \le \sum_{i=1}^{M}\pi_i^*\mathcal{R}^i(h) + \frac{\gamma+\rho}{2} + \lambda_{\mathcal{H},(\mathcal{P}_X^t,\mathcal{P}_X^*)} \quad (2)$$

It is assumed that the source domain prior probabilities satisfy $\sum_{i=1}^{M}\pi_i^* = 1$, $\pi_i > 0$, for all $i = 1...M$. Substitute Equation (2) into Equation (1), we have

$$\mathcal{R}^t(h) \le (1-\pi^u)\left(\sum_{i=1}^{M}\pi_i^*\mathcal{R}^i(h) + \frac{\gamma+\rho}{2} + \lambda_{\mathcal{H},(\mathcal{P}_X^t,\mathcal{P}_X^*)}\right) + \pi^u\cdot\mathcal{R}^{OS}(h) \quad (3)$$

Since both π^u and $1-\pi^u$ belong to $[0,1]$, the inequality simplifies to

$$\mathcal{R}^t(h) \le \sum_{i=1}^{M}\pi_i^*\mathcal{R}^i(h) + \frac{\gamma+\rho}{2} + \lambda_{\mathcal{H},(\mathcal{P}_X^t,\mathcal{P}_X^*)} + \mathcal{R}^{OS}(h) \quad (4)$$

where $\mathcal{R}^i(h)$ is the risk of the i-th source domain. $\lambda_{\mathcal{H}(\mathcal{P}_{\mathcal{X}}^t,\mathcal{P}_{\mathcal{X}}^*)}$ is the ideal joint risk across the target domain and the domain with the best approximator distribution $\mathcal{P}_{\mathcal{X}}^*$. $\mathcal{R}^{\mathrm{OS}}(h)$ is the open space risk, which represents the risk of misclassifying unknown-class samples in target domain to known classes.

Appendix B: Proof for Lemma 1

The prompt of the c-th class is formulated as,

$$p_c = [\Phi(v_{\mathrm{dom}})], [classname] \quad (5)$$

where $\Phi(v_{\mathrm{dom}})$ is the domain token and *classname* is the specific class. Assuming that the mapping function F_t satisfies linear superposition, then the corresponding feature embedding is, p^c can be decomposed as,

$$F_t(p^c) = \Phi(v_{\mathrm{dom}}) + class\_emb(c) \quad (6)$$

where $\Phi(v_{\mathrm{dom}})$ represents the domain-level features, and *class_emb(c)* is the base embedding of the class name. Note that *class_emb(c)* is a coarse-grained embedding of the class name like "cat" or "dog", which contains extensive shared information between classes, usually resulting in small differences.

At the same time, the semantic-enhanced class prompt is,

$$p_{sem}^c = [\Phi(v_{\mathrm{dom}})], [\Psi_1(v_{sem}^{(1,c)}), \dots, \Psi_K(v_{sem}^{(K,c)})], [classname] \quad (7)$$

Thus the feature of the semantic-enhanced class prompt can be decomposed as,

$$F_t(p_{\mathrm{enh}}^c) = \Phi(v_{\mathrm{dom}}) + \sum_{k=1}^{K}\Psi_k(v_{\mathrm{sem}}^{(k,c)}) + class\_emb(c) \quad (8)$$

where $\Psi_k(v_{\mathrm{sem}}^{(k)})$ denotes the k-th fine-grained semantic token, corresponding to c-th local features, e.g., "vertical pupils of a cat" or "sharp beak of a bird". For different classes, it is commonly that $v_{\mathrm{sem}}^{(k,c)} \perp v_{\mathrm{sem}}^{(k,d)}$ that is, the fine-grained features of distinct classes are orthogonal or weakly correlated.

For classes c and d ($c \neq d$) in the same domain, the class discrepancy based on traditional prompts can be written as,

$$\begin{aligned}\mathrm{dis}(c,d) &= \mathrm{dis}\Big(F_t(p^c), F_t(p^d)\Big) \\ &= \mathrm{dis}\Big(class\_emb(c), class\_emb(d)\Big)\end{aligned} \quad (9)$$

While the discrepancy in terms of the enhanced prompts is,

$$\begin{aligned}\mathrm{dis}_{sim}(c,d) &= \mathrm{dis}\Big(F_t(p_{sim}^c), F_t(p_{sim}^d)\Big) \\ &\approx \mathrm{dis}\Big(\sum_{k=1}^{K}\Psi_k(v_{\mathrm{sem}}^{(k,c)}), \sum_{k=1}^{K}\Psi_k(v_{\mathrm{sem}}^{(k,d)})\Big) \\ &\quad + \mathrm{dis}\Big(class\_emb(c), class\_emb(d)\Big)\end{aligned} \quad (10)$$

The orthogonal or weakly correlated fine-grained features of distinct classes satisfy,

$$\mathrm{dis}\Big(\sum_{k=1}^{K}\Psi_k(v_{\mathrm{sem}}^{(k,c)}), \sum_{k=1}^{K}\Psi_k(v_{\mathrm{sem}}^{(k,d)})\Big) > 0$$

Consequently, the discrepancy semantic-enhanced prompts are diluted by class-specific fine-grained features, resulting in a larger value,

$$dis_{sem}(c,d) > dis(c,d)$$

In scenarios where classes c and d belong to different domains, the above inequality can be derived through a similar reasoning process.

Appendix C: Hyperparameter Sensitivity Analysis

Hyperparameter Sensitivity Analysis. Figure 1 shows the sensitivity analysis of two critical hyper-parameters. For attention heads number K, the optimal performance occurs at K=4. Fewer heads reduce accuracy due to insufficient semantic modeling, while more heads also degrade performance via overfitting. For σ in diffusion generation, the best value is σ=0.2. Lower values limit the pseudo-unknown diversity, whereas higher values introduce noise that erodes semantic coherence.

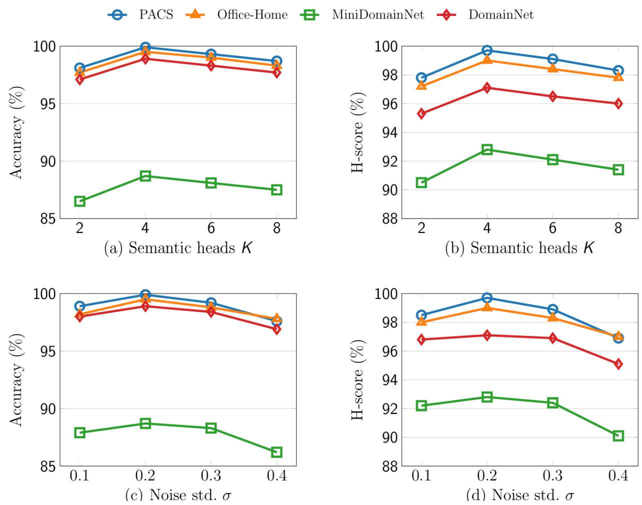

Figure 1: Hyperparameter Sensitivity Analysis for SeeCLIP.

References

Fang Z, Lu J, Liu F, et al. Open set domain adaptation: Theoretical bound and algorithm[J]. IEEE transactions on neural networks and learning systems, 2020, 32(10): 4309-4322.

Wang J, Lan C, Liu C, et al. Generalizing to unseen domains: A survey on domain generalization[J]. IEEE transactions on knowledge and data engineering, 2022, 35(8): 8052-8072.