VISUAL STRUCTURE-GUIDED TEXTUAL SEMANTIC SPACE ADAPTATION FOR DOMAIN-SHIFTED GENERALIZED CATEGORY DISCOVERY

Anonymous authors
Paper under double-blind review

ABSTRACT

Domain-Shifted Generalized Category Discovery (DS-GCD) requires categorizing an unlabeled target domain containing both source-seen and target-private categories under domain shift. Existing VLM-based approaches commonly cluster target visual features and represent discovered categories with anonymous visual prototypes, leaving the textual space unused as the target assignment coordinates. We investigate an alternative division of roles: target visual geometry provides unlabeled structural evidence, whereas text embeddings define the coordinates of the target assignment space. We propose Visual Structure Guided Textual Semantic Space Adaptation (ViSTA), which uses target visual structure to search a predefined noun vocabulary for textual anchors and then adapts CLIP within the resulting textual semantic space. In Stage 1, ViSTA filters candidate nouns and greedily selects anchors according to the compactness and separation of the target partitions they induce, under a given target category budget. In Stage 2, the selected words remain fixed while source supervision and target regularization adapt both CLIP encoders. Across four DS-GCD benchmarks and three seen/unseen splits, ViSTA attains a 75.76% average H-Score over 12 settings. These results support textual anchors as category coordinates while retaining target visual geometry as selection evidence.

1 INTRODUCTION

Domain-Shifted Generalized Category Discovery (DS-GCD) asks a model trained on labeled source samples from seen classes to categorize an unlabeled target domain containing both seen and unseen classes (Wen & Brbic, 2024; Rongali et al., 2024; Wang et al., 2025b; Feng & Ge, 2025). The task couples two challenges: domain shift changes the appearance distribution, while category discovery requires separating target-private samples into fine-grained categories. Conventional GCD usually assumes that labeled and unlabeled samples share a domain (Vaze et al., 2022; Wen et al., 2023). Universal Domain Adaptation (UniDA) handles distribution and label-set shifts but primarily detects target-private samples as unknown (Saito & Saenko, 2021; Saito et al., 2020; You et al., 2019). DS-GCD requires both seen-category recognition and fine-grained novel-category discovery within one target assignment space.

Existing DS-GCD methods primarily organize target samples in a visual representation space (Rongali et al., 2024; Wang et al., 2025b; Feng & Ge, 2025). CROW (Wen & Brbic, 2024), for example, clusters CLIP visual features into target prototypes, matches a subset to seen classes, and treats the remaining prototypes as novel categories. These methods establish that target visual geometry provides useful evidence for category discovery. Their category representation remains an anonymous visual prototype: the visual space supplies both the clustering evidence and the category coordinates, while CLIP's textual space does not participate in representing the discovered categories. This motivates using target geometry to evaluate candidate textual anchors through the partitions they induce.

In this formulation, category discovery partitions target samples into seen and unseen categories under a given category budget. Semantic retrieval selects textual anchors from a predefined candidate vocabulary, instantiated with WordNet nouns (Miller, 1995). We couple these operations by

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.

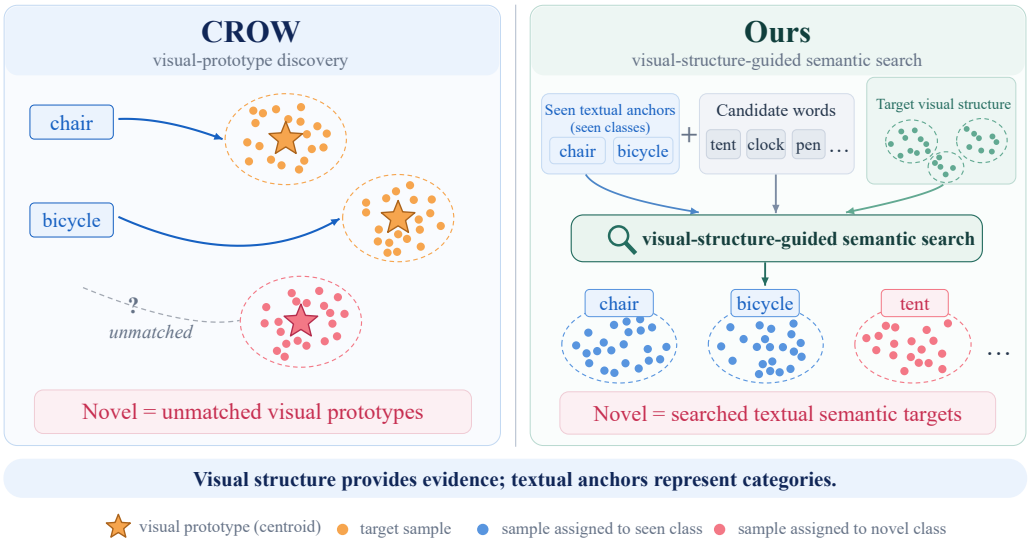


Figure 1: Conceptual distinction between visual-prototype discovery and ViSTA's textual-anchor search. Target visual structure provides selection evidence, while textual anchors represent the category coordinates in ViSTA. Orange dots denote target samples; blue and red dots indicate assignments to seen and novel classes, respectively.

preferring anchors whose image–text assignments agree with category-level target visual structure. Partition quality thus guides the construction of the textual semantic space.

We propose Visual Structure Guided Textual Semantic Space Adaptation (ViSTA), a two-stage framework for DS-GCD. Stage 1 filters the noun vocabulary and greedily selects textual anchors according to the compactness and separation of their induced target partitions. Stage 2 keeps the selected words fixed and adapts CLIP with source supervision and target regularization. ViSTA uses target visual geometry as structural evidence and text embeddings as category coordinates (Figure 1).

Our contributions are summarized as follows:

- We formulate a vocabulary-constrained textual-anchor view of DS-GCD that separates visual structural evidence from the textual coordinates used to represent target categories.
- We develop ViSTA, which uses target visual organization to construct a textual semantic space and adapts CLIP within the selected anchor vocabulary.
- We evaluate ViSTA on four DS-GCD benchmarks and three seen/unseen splits. ViSTA attains a 75.76% average H-Score across the 12 settings, while component analyses examine semantic filtering, search sensitivity, and the Stage 2 objective.

2 Related Work

Generalized Category Discovery. Novel Class Discovery transfers knowledge from labeled seen classes to discover categories in unlabeled data (Han et al., 2019; 2020; Fini et al., 2021; Troisemaine et al., 2023). GCD extends this setting to an unlabeled mixture of seen and unseen classes (Vaze et al., 2022). Methods such as GCD, ORCA, and SimGCD use contrastive learning, clustering, or parametric classification (Vaze et al., 2022; Cao et al., 2022; Wen et al., 2023). They usually assume a shared domain for labeled and unlabeled data. DS-GCD additionally requires transferring seen-class knowledge across domains while discovering target-private categories (Rongali et al., 2024; Wang et al., 2025b; Feng & Ge, 2025).

Universal Domain Adaptation. Domain adaptation addresses distribution shift between labeled source and unlabeled target data (Ganin & Lempitsky, 2015; Long et al., 2017). UniDA also allows unknown label-set overlap (Chang et al., 2022; Saito & Saenko, 2021; You et al., 2019; Saito et al., 2020). DANCE, OVANet, UniOT, and GLC align transferable classes and identify target-private samples through neighborhood structure, one-vs-all prediction, optimal transport, or clustering (Saito et al., 2020; Saito & Saenko, 2021; Chang et al., 2022; Qu et al., 2023). DCC, NCDDA, and SAN further study class separation across domains (Li et al., 2021; Yu et al., 2022; Zang et al., 2023). DS-GCD requires fine-grained discovery within the target-private set, in addition to recognizing shared classes.

Vision-Language Model Discovery. Vision-language models such as CLIP provide aligned image and text representations and have become a strong foundation for open-vocabulary recognition (Radford et al., 2021). Recent category discovery methods exploit this multimodal structure by using CLIP features, retrieved descriptions, generated textual embeddings, or cross-modality co-training to improve discovery performance (Ouldnoughi et al., 2023; Wang et al., 2025a; Zheng et al., 2024). In the domain-shifted setting, CROW is the closest prior work to ours. CROW shows that foundation-model visual features provide a strong basis for DS-GCD, but its discovery process remains centered on visual prototypes. SpectralGCD (Caselli et al., 2026) instead learns from image–concept similarities with teacher-guided concept selection. ViSTA selects a category-sized set of textual anchors through target partitions and retains the selected words during adaptation. We include a DS-GCD adaptation of SpectralGCD in the main comparison.

3 METHOD

3.1 PROBLEM FORMULATION

In DS-GCD settings (Wen & Brbic, 2024), we have a labeled source dataset $D_s = \{(x_i^s, y_i^s)\}_{i=1}^{N_s}$, for each image x_i^s , y_i^s is its corresponding label. Let $D_t = \{x_i^t\}_{i=1}^{N_t}$ denote the unlabeled target dataset. We denote the set of classes in the labeled source dataset as Y_s and the set of classes in the unlabeled target dataset as Y_t. We consider Y_s as a set of seen classes, so that $Y_u = Y_t \setminus Y_s$ is the set of unseen classes which contains novel categories. Besides, under the domain shift, we assume that $P(x^s) \neq P(x^t)$, where $P(x^s)$ and $P(x^t)$ denote the source and target image distributions. The goal is to assign each target sample to a fine-grained category in Y_t, including both seen and unseen classes.

3.2 MODELS

We build upon vision-language model CLIP, which has an image encoder $f_\theta(\cdot)$ with a learnable parameter θ and a text encoder $g_\varphi(\cdot)$ with a learnable parameter φ (Radford et al., 2021). For an image x and a set of predicted text labels $Y = \{y_1, y_2, ..., y_n\}$, CLIP can compute the cosine similarity of their embeddings for matching purposes. Let $I = f_\theta(x)$ and $T_i = g_\varphi(y_i)$ denote the image and text embeddings. The probability that image x is assigned to label y_i is:

$$P(y_i|x) = \frac{\exp(\text{sim}(I, T_i)/\tau)}{\sum_{j=1}^{n} \exp(\text{sim}(I, T_j)/\tau)}, \tag{1}$$

where $sim(\cdot, \cdot)$ denotes cosine similarity and τ denotes the softmax temperature. We use CLIP to construct a textual semantic space $\mathcal{A}_u$, and we use the source class names as textual anchors for seen classes, so we have $\mathcal{A}_s = \{a_y \mid y \in \mathcal{Y}_s\}$. The resulting target textual semantic space is $\mathcal{A}_t = \mathcal{A}_s \cup \mathcal{A}_u$. Finally, we use LoRA to fine-tune CLIP in order to efficiently perform category discovery based on the text semantic space (Hu et al., 2022).

3.3 OVERVIEW OF VISTA

ViSTA separates the evidence used to evaluate a target partition from the coordinates used to represent its assignments. Target visual geometry supplies structural evidence, while textual anchors define the target assignment coordinates. Figure 2 shows the two stages.

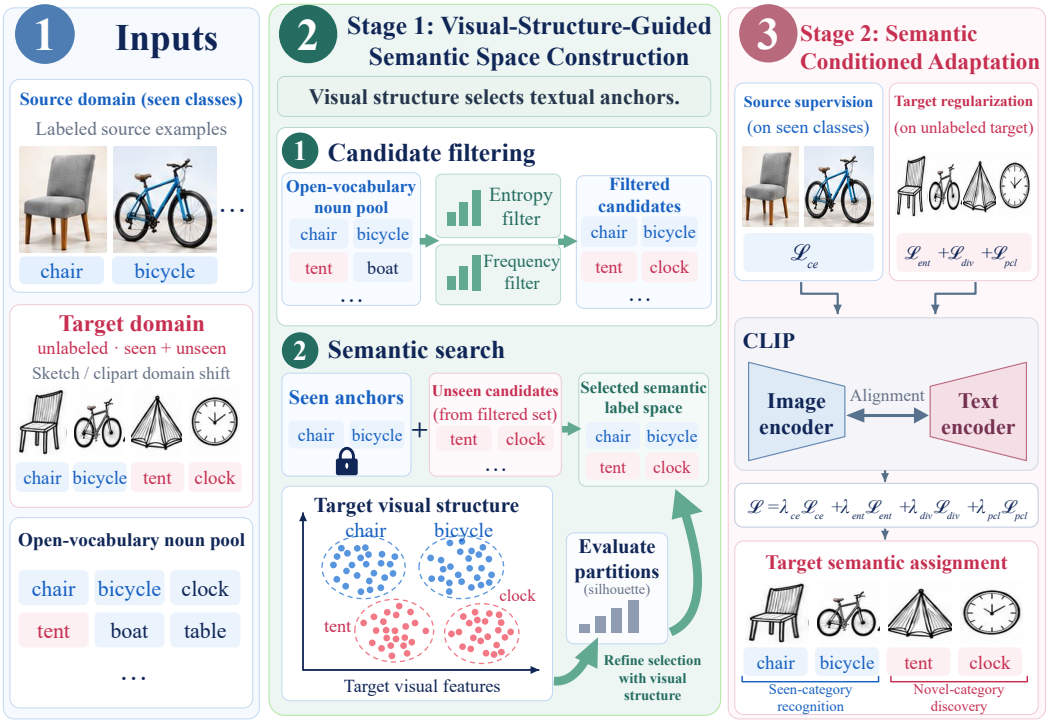


Figure 2: Overview of ViSTA. Stage 1 uses frozen target visual features to evaluate the partitions induced by candidate textual anchors and constructs a fixed vocabulary-constrained assignment space. Stage 2 keeps the selected anchors fixed and adapts the image–text assignment function using source supervision and target regularization.

In Stage 1, each candidate anchor set partitions target images through image–text similarity. We evaluate the compactness and separation of these groups in the frozen visual feature space. Stage 2 uses source supervision and target regularization to improve assignment confidence, diversity, and consistency across augmentations. The selected words remain fixed, while their embeddings change with encoder adaptation.

3.4 Visual-Structure-Guided Textual Semantic Space Construction

Stage 1 constructs the target textual semantic space by searching unseen textual semantic targets from an open-vocabulary noun pool. Let $\mathcal{A}_s = \{a_y \mid y \in \mathcal{Y}_s\}$ denote the textual anchors of source seen classes, and let $\mathcal{V} = \{v_m\}_{m=1}^{M}$ denote the candidate noun pool, instantiated by WordNet nouns in our implementation (Miller, 1995). Given the target category budget $K_0 = |\mathcal{Y}_t|$, ViSTA searches unseen textual semantic targets $\mathcal{A}_u \subset \mathcal{V}$ and constructs

$$\mathcal{A}_t = \mathcal{A}_s \cup \mathcal{A}_u, \qquad |\mathcal{A}_t| = K_0. \tag{2}$$

The category budget is given; the search determines which nouns occupy the unseen slots.

Candidate Filtering. Directly searching over $\mathcal{V}$ is noisy and inefficient. We therefore apply visual-aware dual filtering to obtain a target-relevant candidate pool $\mathcal{V}'$. Let Z_i denote normalized target-domain image features used for filtering, and let T_{v_m} denote the normalized CLIP text feature of noun v_m. We first obtain K_0 visual centroids $\{\mu_k\}_{k=1}^{K_0}$ with K-means (Lloyd, 1982) and compute the response distribution of each noun over these centroids:

$$q_{k|m} = \frac{\exp(\mu_k^\top T_{v_m}/\tau_e)}{\sum_{k'=1}^{K_0} \exp(\mu_{k'}^\top T_{v_m}/\tau_e)}. \tag{3}$$

The entropy filter keeps nouns with concentrated centroid responses, measured by the normalized entropy $H_e(v_m)$. The frequency filter further keeps nouns that appear frequently among the top textual matches of target samples, measured by $F(v_m)$. The filtered pool is

$$\mathcal{V}' = \{v_m \in \mathcal{V} \mid H_e(v_m) \leq \gamma_{ent},\ F(v_m) \geq \delta_{fre}\}. \tag{4}$$

The detailed definitions of $H_e(v_m)$, $F(v_m)$, and δ_{fre} are provided in Appendix A.1. This filtering step removes visually ambiguous or weakly supported nouns before semantic target search.

Structure-Guided Scoring. Given a candidate textual semantic space $\mathcal{S}$, ViSTA assigns each target sample to the most compatible textual anchor:

$$\pi_{\mathcal{S}}(x_i^t) = \arg\max_{a \in \mathcal{S}} (Z_i^t)^\top T_a, \tag{5}$$

where $Z_i^t = f_\theta(x_i^t)/\|f_\theta(x_i^t)\|_2$ is the normalized target image feature and $T_a = g_\varphi(a)/\|g_\varphi(a)\|_2$ is the normalized text-anchor feature. Thus, each anchor set defines a partition of the target images.

ViSTA evaluates the induced partition using the target visual structure. Specifically, we define

$$\mathrm{Score}(\mathcal{S}) = \mathrm{Sil}\left(\{\pi_{\mathcal{S}}(x_i^t)\}_{i=1}^{N_t}, \{Z_i^t\}_{i=1}^{N_t}\right), \tag{6}$$

where $\mathrm{Sil}(\cdot)$ is the silhouette score computed in the target visual feature space (Rousseeuw, 1987). A higher score indicates that the textual anchors induce more compact intra-class structure and larger inter-class separation among target samples. The full definition is provided in Appendix A.2. In our implementation, this score is computed on unlabeled target samples, with optional subsampling for memory efficiency.

Greedy Semantic Search. Stage 1 searches for textual semantic targets by maximizing the visual-structure compatibility of the induced target partition:

$$\mathcal{A}_u^\star = \arg\max_{\mathcal{A}_u \subset \mathcal{V}'} \mathrm{Score}(\mathcal{A}_s \cup \mathcal{A}_u), \qquad |\mathcal{A}_s \cup \mathcal{A}_u| = K_0. \tag{7}$$

Since exact optimization is combinatorial, ViSTA uses greedy semantic target search. It maintains a semantic slot set

$$\mathcal{S} = \{s_1, \ldots, s_{K_0}\}, \tag{8}$$

where source slots are anchored by $\mathcal{A}_s$ and the remaining slots are initialized from $\mathcal{V}'$. For each unseen slot s_i, ViSTA builds a local candidate set $\mathcal{C}_i \subseteq \mathcal{V}'$ from semantically nearby nouns while suppressing redundant anchors. The slot is updated by

$$s_i \leftarrow \arg\max_{v \in \mathcal{C}_i} \mathrm{Score}\left(\mathcal{S}_{i \leftarrow v}\right), \tag{9}$$

where $\mathcal{S}_{i \leftarrow v}$ denotes the temporary semantic space obtained by replacing s_i with v. The current slot is included in $\mathcal{C}_i$, so the update can keep the slot unchanged if no candidate improves the score. The construction of $\mathcal{C}_i$ is provided in Appendix A.3. After multiple rounds, the optimized slots define $\mathcal{A}_t$.

The resulting anchors define the category coordinates used in Stage 2.

3.5 Textual Semantic Space-Conditioned Adaptation

Stage 2 adapts both CLIP encoders with LoRA while keeping the selected anchor vocabulary $\mathcal{A}_t$ fixed (Hu et al., 2022). Encoder updates change the image and text embeddings, refining target assignments within this textual semantic space.

At each training iteration, we sample a labeled source mini-batch $\mathcal{B}_s = \{(x_i^s, y_i^s)\}_{i=1}^{B_s}$ and an unlabeled target mini-batch with two views $\mathcal{B}_t = \{(x_i^{t,1}, x_i^{t,2})\}_{i=1}^{B_t}$. Source predictions are computed over the seen anchors $\mathcal{A}_s$, while target predictions are computed over the full target semantic space $\mathcal{A}_t$. We denote the source prediction by $\mathbf{p}_i^s$, and the two target-view predictions by $\mathbf{p}_i^{t,1}$ and $\mathbf{p}_i^{t,2}$.

The source cross-entropy loss $\mathcal{L}_{ce}$ preserves seen-class discrimination. For target samples, ViSTA applies information maximization on $\mathbf{p}_i^{t,1}$:

$$\mathcal{L}_{im} = \mathcal{L}_{ent} + \mathcal{L}_{div}, \tag{10}$$

where $\mathcal{L}_{ent}$ minimizes the entropy of individual target predictions and $\mathcal{L}_{div}$ encourages a diverse batch-level mean prediction over $\mathcal{A}_t$ (Grandvalet & Bengio, 2004; Liang et al., 2020). This combination sharpens target assignments while reducing collapse to a few textual anchors.

To enforce assignment stability under augmentation, ViSTA further uses prediction-level contrastive consistency. The PCL loss $\mathcal{L}_{pcl}$ treats $\mathbf{p}_i^{t,1}$ and $\mathbf{p}_i^{t,2}$ from the same target image as a positive pair, and predictions from other target images as negatives (Chen et al., 2020; van den Oord et al., 2018). Unlike feature-level contrastive learning, this loss operates on semantic assignment distributions over $\mathcal{A}_t$, encouraging two views of the same target sample to be grounded into consistent textual semantic targets. Its explicit InfoNCE form is provided in Appendix A.4.

The overall Stage 2 objective is

$$\mathcal{L} = \lambda_{ce}\mathcal{L}_{ce} + \lambda_{ent}\mathcal{L}_{ent} + \lambda_{div}\mathcal{L}_{div} + \lambda_{pcl}\mathcal{L}_{pcl}. \tag{11}$$

Through this objective, Stage 2 preserves source seen-class recognition while stabilizing target assignments in the textual semantic space constructed by Stage 1.

4 Experiments

4.1 Experimental Setup

Datasets and Protocol. We evaluate ViSTA on four domain-shifted category discovery benchmarks: Office (Saenko et al., 2010), OfficeHome (Venkateswara et al., 2017), DomainNet (Peng et al., 2019), and VisDA (Peng et al., 2017). All experiments follow the DS-GCD protocol: the source domain provides labeled samples from seen classes, while the target domain is unlabeled and contains both seen and unseen classes. For each benchmark, we evaluate three seen/unseen category splits, including 70%, 50%, and 30% seen-class ratios. For Office, we use Amazon, DSLR, and Webcam, resulting in six ordered transfer tasks. For OfficeHome, we use Art, Clipart, Product, and Real-World, resulting in twelve ordered transfer tasks. For DomainNet, we use Painting, Real, and Sketch, resulting in six ordered transfer tasks. For VisDA, we evaluate the standard synthetic-to-real transfer task. Results are averaged over all transfer tasks of each benchmark under each seen/unseen split.

Evaluation Metrics. We report Seen Accuracy, Unseen Accuracy, and H-Score. Seen Accuracy measures classification accuracy on target samples from seen classes. Unseen Accuracy is computed after matching predicted unseen clusters to ground-truth unseen classes using the Hungarian algorithm (Kuhn, 1955). H-Score is the harmonic mean of Seen Accuracy and Unseen Accuracy, and serves as the primary metric for evaluating the trade-off between preserving seen-class recognition and discovering unseen categories. Formal metric definitions are provided in Appendix B.1.

Implementation. ViSTA is implemented with CLIP ViT-B/16 (Radford et al., 2021; Dosovitskiy et al., 2021). In Stage 1, we construct the target textual semantic space using source seen-class names and WordNet nouns. The target category budget K_0 is set to the total number of target categories under each split. Stage 1 uses visual-structure-guided silhouette scoring with dual candidate filtering. In Stage 2, we apply LoRA to the visual and textual branches of CLIP and optimize the model. Implementation hyperparameters are provided in Appendix B.2.

Baselines. We compare ViSTA with GCD, ORCA, SIMPLE, SAN, DANCE, OVANet, UniOT, GLC, DCC, NCDDA, and CROW (Vaze et al., 2022; Cao et al., 2022; Zang et al., 2023; Saito et al., 2020; Saito & Saenko, 2021; Chang et al., 2022; Qu et al., 2023; Li et al., 2021; Yu et al., 2022; Wen & Brbic, 2024). Results marked with † use CLIP ViT-L/14 and are taken from CROW; other methods use CLIP ViT-B/16. On VisDA, we add SpectralGCD-SC (Caselli et al., 2026), a DS-GCD adaptation with a source-conditioned mixed-batch prior and a frozen LAION CLIP ViT-H/14 teacher. Appendix D specifies its objective and training protocol.

4.2 Results

Table 1 compares performance across the four benchmarks. ViSTA achieves the best average H-Score of 75.76%, exceeding GLC by 4.12 points, and ranks first in 10 of the 12 benchmark-split

settings. With the same CLIP ViT-B/16 backbone, it improves over CROW by 10.44 points on average.

Table 1: H-Score (%) on four DS-GCD benchmarks, averaged over transfer tasks under each seen/unseen split. †: CLIP ViT-L/14 results from CROW; other methods use CLIP ViT-B/16. SpectralGCD-SC is adapted to DS-GCD with a source-conditioned prior and a frozen ViT-H/14 teacher. – denotes unavailable results; Avg covers all 12 settings. Best and second-best results are in **bold** and underline.

Method	OFFICE			OFFICEHOME			VISDA			DOMAINNET			Avg
	21/10	16/15	10/21	45/20	33/32	20/45	8/4	6/6	4/8	240/105	173/172	105/240	
SIMPLE†	64.9	66.8	78.3	62.3	66.0	65.4	55.4	50.8	50.9	53.2	55.9	57.8	60.64
GCD†	62.6	58.5	58.0	48.7	47.5	48.1	31.2	32.5	25.2	35.0	41.3	41.3	44.16
ORCA†	57.9	63.4	62.3	48.9	48.9	49.9	31.3	33.6	33.9	28.9	31.5	33.7	43.68
DCC†	72.8	74.9	75.2	63.6	65.0	64.7	60.7	57.3	56.7	45.5	47.5	47.7	60.97
NCDDA†	80.3	81.2	81.7	63.2	64.3	65.0	57.2	60.7	59.3	50.1	52.6	55.7	64.28
SAN†	**80.5**	80.2	**82.0**	64.3	65.0	67.2	61.2	63.5	61.5	53.2	54.8	55.0	65.70
DANCE	78.6	76.8	81.4	76.8	73.7	71.6	68.3	67.8	61.9	58.6	60.2	53.9	69.13
OVANET	62.0	64.8	68.2	64.3	60.5	65.5	40.8	41.5	60.9	47.7	47.1	43.2	55.54
UNIOT	68.2	71.7	81.2	74.4	72.1	70.9	53.5	62.0	53.2	45.2	52.1	58.3	63.57
GLC	80.3	79.1	80.9	76.1	73.6	74.4	72.6	72.0	74.3	57.2	58.9	60.3	71.64
CROW	71.6	74.7	79.7	65.6	63.3	67.7	71.8	68.6	67.0	50.3	51.3	52.3	65.32
SpectralGCD-SC	–	–	–	–	–	–	44.9	52.4	67.5	–	–	–	–
ViSTA	78.4	**83.8**	81.8	**81.8**	**80.1**	**78.6**	**84.6**	**76.9**	**77.8**	**59.2**	**62.1**	**64.0**	**75.76**

The largest gain is on VisDA 8/4, where ViSTA exceeds the second-best result by 12.0 points. It also leads on all OfficeHome and DomainNet splits, although the DomainNet gains are smaller. On Office, it ranks first on 16/15 and second on 10/21, while SAN and NCDDA perform better on 21/10.

Table 2 breaks down the 50% split into Seen and Unseen Accuracy. ViSTA obtains the highest H-Score on all four benchmarks, combining competitive seen-category recognition with improved unseen discovery.

Table 2: Seen Accuracy (%), Unseen Accuracy (%), and H-Score (%) under the 50% shared-class split. †: CLIP ViT-L/14 results from CROW; unmarked methods: CLIP ViT-B/16. SpectralGCD-SC additionally uses a frozen ViT-H/14 teacher. Best and second-best values are highlighted in **bold** and underline, respectively.

Method	OFFICE (16/15)			OFFICEHOME (33/32)			VISDA (6/6)			DOMAINNET (173/172)		
	Seen	Unseen	H	Seen	Unseen	H	Seen	Unseen	H	Seen	Unseen	H
SIMPLE†	62.3	72.0	66.8	69.2	63.1	66.0	68.1	40.5	50.8	69.1	47.0	55.9
GCD†	54.1	63.8	58.5	46.6	48.5	47.5	43.8	25.8	32.5	42.3	40.3	41.3
ORCA†	69.6	58.2	63.4	67.1	38.4	48.9	68.3	22.3	33.6	62.3	21.1	31.5
DCC†	78.0	72.1	74.9	70.3	60.5	65.0	75.3	46.2	57.3	50.2	45.1	47.5
NCDDA†	88.9	74.7	81.2	71.2	58.6	64.3	70.4	53.3	60.7	68.9	42.5	52.6
SAN†	89.4	72.7	80.2	72.0	59.2	65.0	74.5	55.3	63.5	67.3	46.2	54.8
DANCE	81.7	72.6	76.8	80.9	67.7	73.7	**92.4**	53.6	67.8	**75.0**	50.3	60.2
OVANET	65.7	64.0	64.8	63.3	57.9	60.5	41.2	41.8	41.5	55.1	41.1	47.1
UNIOT	73.6	69.9	71.7	78.3	66.9	72.1	78.5	51.2	62.0	60.2	45.9	52.1
GLC	**93.1**	68.7	79.1	79.1	68.9	73.6	81.1	64.8	72.0	66.0	53.1	58.9
CROW	79.2	70.7	74.7	65.4	61.3	63.3	76.3	62.3	68.6	61.1	44.3	51.3
SpectralGCD-SC	–	–	–	–	–	–	57.9	47.8	52.4	–	–	–
ViSTA	88.8	**79.3**	**83.8**	**82.7**	**77.6**	**80.1**	85.6	**69.8**	**76.9**	74.0	**53.6**	**62.1**

4.3 ANALYSIS

Stage-2 Loss Design. Table 3 examines the Stage-2 objective on VisDA (6/6). CE-only and PCL-only favor seen recognition. Combining IM and PCL improves over either alone, and adding source CE yields the highest H-Score among these variants. The complementary objectives help balance seen recognition and unseen discovery.

Table 3: Stage-2 loss ablations on VisDA (6/6). IM denotes the information-maximization objective $\mathcal{L}_{ent} + \mathcal{L}_{div}$.

Setting	CE	IM	PCL	Seen	Unseen	H-Score
CE only	✓			<u>91.0</u>	55.1	68.6
IM only		✓		79.4	61.6	69.4
PCL only			✓	**93.1**	56.5	70.3
CE + IM	✓	✓		83.7	61.8	71.1
IM + PCL		✓	✓	84.8	<u>65.9</u>	<u>74.1</u>
ViSTA (default)	✓	✓	✓	85.6	**69.8**	**76.9**

Semantic Filtering. Table 4 compares candidate filtering strategies before greedy search. Combining entropy and frequency filtering achieves the highest Seen Accuracy, Unseen Accuracy, and H-Score. The two filters remove different weak candidates: entropy filters nouns with diffuse visual responses, while frequency filters nouns with insufficient target-sample support. Together, they narrow the vocabulary to candidates that can support a coherent target partition.

Table 4: Stage-1 filtering ablations on VisDA (6/6). The results show the effect of different candidate filtering strategies before semantic target search.

Filtering Mode	Seen	Unseen	H-Score
Frequency only	84.8	62.9	72.2
Entropy only	83.7	68.6	75.4
Dual filtering	**85.6**	**69.8**	**76.9**

Search Hyperparameters. Figure 3 varies the semantic neighborhood size K_s, entropy threshold γ_{ent}, and frequency threshold γ_{fre} on VisDA (6/6). Increasing K_s expands the candidate set and improves search quality, with diminishing returns at larger values. A stricter entropy threshold removes ambiguous nouns but may discard useful alternatives; the frequency threshold controls the target support required for inclusion. These results indicate that candidate coverage and filtering selectivity both affect the search.

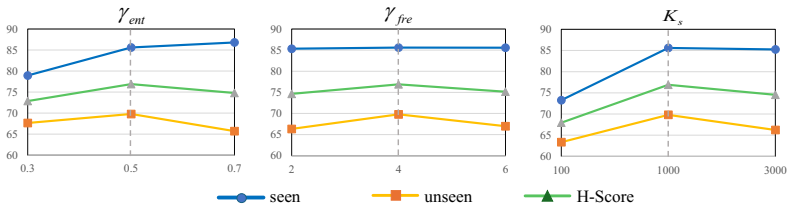


Figure 3: Sensitivity of Stage-1 hyperparameters on VisDA (6/6).

Two-Stage Complementarity. We use a 2×2 study on VisDA (6/6) to separate Stage 1 search (S) from Stage 2 adaptation (A). All variants share the vocabulary, filtering, initialization, backbone, and evaluation. V00/V01 use the pre-search anchors $\mathcal{A}_{\text{init}}$; V10/V11 use the anchors $\mathcal{A}_{\text{search}}$ obtained from that initialization. V01 and V11 share the same adaptation configuration.

Table 5: Controlled 2×2 stage-isolation study on VisDA(6/6). S and A denote Stage 1 greedy anchor search and Stage 2 LoRA adaptation, respectively. Adaptation variants are evaluated at the final iteration.

Variant	S	A	Anchor	Seen	Unseen	H-Score
V00	–	–	$\mathcal{A}_{\text{init}}$	71.6	25.0	37.1
V10	✓	–	$\mathcal{A}_{\text{search}}$	**90.7**	53.8	67.5
V01	–	✓	$\mathcal{A}_{\text{init}}$	81.2	53.5	64.5
V11	✓	✓	$\mathcal{A}_{\text{search}}$	85.6	**69.8**	**76.9**

Table 5 shows that search and adaptation contribute at different points. Search raises H by 30.4 points with frozen features and by 12.4 points when followed by adaptation. Starting from the searched anchors, adaptation increases Unseen Accuracy from 53.8 to 69.8 and yields the highest H-Score. Together, these results support the two-stage design: target geometry identifies useful textual anchors, and encoder adaptation improves category separation within their textual semantic space.

Textual Anchors and Category Budget. Table 6 tests category budget and literal-name dependence on frozen Stage-1 partitions for VisDA (6/6). With six source anchors fixed, budgets of 10 and 14 lower H by 6.17 and 2.03 points, respectively, relative to 12. Excluding target-private names and direct WordNet synonyms reduces the filtered pool from 584 to 486 candidates, yet leaves the selected anchors and assignments unchanged. The anchor *rollerblading* covers 89.52% of skateboard samples with 62.47% purity. These anchors can therefore organize categories through related vocabulary. Appendix C gives the exclusion protocol and anchor-level analysis.

Table 6: Frozen Stage-1 sensitivity on VisDA (6/6), seed 2022. The true category count is 12; ΔH is relative to the $K = 12$ reference.

Condition	Seen	Unseen	H	ΔH
$K = 10$	91.91	46.08	61.39	-6.17
$K = 12$ (reference)	90.77	53.80	67.56	+0.00
$K = 14$	88.31	52.10	65.53	-2.03
$K = 12$, literal-name exclusion	90.77	53.80	67.56	+0.00

5 Conclusion and Limitations

We presented ViSTA, a two-stage Visual Structure Guided Textual Semantic Space Adaptation framework for Domain-Shifted Generalized Category Discovery. Target visual geometry guides textual-anchor selection, and encoder adaptation improves assignments within the resulting textual semantic space. Experiments on four benchmarks show improved balance between seen-category recognition and novel-category discovery. ViSTA uses a predefined vocabulary and a given category budget; extending it to adaptive category-number estimation and richer semantic candidates is a direction for future work.

Reproducibility Statement

Appendix A gives the filtering, scoring, candidate construction, and consistency definitions. Appendix B provides metrics and hyperparameters. Appendices C–D document the supplementary experiments, including seeds, shared inputs, and evaluation endpoints. Supplementary tables are generated from archived metric and contingency records.

AI Use Statement

Generative AI tools assisted with language editing, title and abstract revision, preparation of figures and their illustrative object images, as well as adaptation of the manuscript to the ICLR 2027 LaTeX template. The authors remain responsible for the content of this work.

REFERENCES

Kaidi Cao, Maria Brbic, and Jure Leskovec. Open-world semi-supervised learning. In *International Conference on Learning Representations*, 2022.

Lorenzo Caselli, Marco Mistretta, Simone Magistri, and Andrew Bagdanov. SpectralGCD: Spectral concept selection and cross-modal representation learning for generalized category discovery. In *International Conference on Learning Representations*, 2026.

Wanxing Chang, Ye Shi, Hoang Tuan, and Jingya Wang. Unified optimal transport framework for universal domain adaptation. In *Advances in Neural Information Processing Systems*, volume 35, pp. 29512–29524, 2022.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 1597–1607. PMLR, 2020.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

Wei Feng and Zongyuan Ge. Generalized category discovery under domain shift: A frequency domain perspective. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. NeurIPS 2025 poster.

Enrico Fini, Enver Sangineto, Stéphane Lathuilière, Zhun Zhong, Moin Nabi, and Elisa Ricci. A unified objective for novel class discovery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9284–9292, 2021.

Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 1180–1189. PMLR, 2015.

Yves Grandvalet and Yoshua Bengio. Semi-supervised learning by entropy minimization. In *Advances in Neural Information Processing Systems*, volume 17, pp. 529–536, 2004.

Kai Han, Andrea Vedaldi, and Andrew Zisserman. Learning to discover novel visual categories via deep transfer clustering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

Kai Han, Sylvestre-Alvise Rebuffi, Sebastien Ehrhardt, Andrea Vedaldi, and Andrew Zisserman. Automatically discovering and learning new visual categories with ranking statistics. In *International Conference on Learning Representations*, 2020.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

Harold W. Kuhn. The hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2(1–2):83–97, 1955. doi: 10.1002/nav.3800020109.

Guangrui Li, Guoliang Kang, Yi Zhu, Yunchao Wei, and Yi Yang. Domain consensus clustering for universal domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9757–9766, 2021.

Jian Liang, Dapeng Hu, and Jiashi Feng. Do we really need to access the source data? Source hypothesis transfer for unsupervised domain adaptation. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 6028–6039. PMLR, 2020.

Stuart Lloyd. Least squares quantization in PCM. *IEEE Transactions on Information Theory*, 28(2): 129–137, 1982. doi: 10.1109/TIT.1982.1056489.

Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I. Jordan. Deep transfer learning with joint adaptation networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 2208–2217. PMLR, 2017.

George A. Miller. WordNet: A lexical database for english. *Communications of the ACM*, 38(11): 39–41, 1995. doi: 10.1145/219717.219748.

Rabah Ouldnoughi, Chia-Wen Kuo, and Zsolt Kira. CLIP-GCD: Simple language guided Generalized Category Discovery. *arXiv preprint arXiv:2305.10420*, 2023.

Xingchao Peng, Ben Usman, Neela Kaushik, Dequan Wang, Judy Hoffman, and Kate Saenko. VisDA: The visual domain adaptation challenge. *arXiv preprint arXiv:1710.06924*, 2017.

Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1406–1415, 2019.

Sanqing Qu, Tianpei Zou, Florian Röhrbein, Cewu Lu, Guang Chen, Dacheng Tao, and Changjun Jiang. Upcycling models under domain and category shift. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20019–20028, 2023.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021.

Sai Bhargav Rongali, Sarthak Mehrotra, Ankit Jha, Mohamad Hassan N C, Shirsha Bose, Tanisha Gupta, Mainak Singha, and Biplab Banerjee. CDAD-Net: Bridging domain gaps in Generalized Category Discovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 2616–2626, 2024.

Peter J. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65, 1987. doi: 10.1016/0377-0427(87)90125-7.

Kate Saenko, Brian Kulis, Mario Fritz, and Trevor Darrell. Adapting visual category models to new domains. In *European Conference on Computer Vision*, pp. 213–226. Springer, 2010.

Kuniaki Saito and Kate Saenko. OVANet: One-vs-all network for universal domain adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9000–9009, 2021.

Kuniaki Saito, Donghyun Kim, Stan Sclaroff, and Kate Saenko. Universal domain adaptation through self supervision. In *Advances in Neural Information Processing Systems*, volume 33, pp. 16282–16292, 2020.

Colin Troisemaine, Vincent Lemaire, Stéphane Gosselin, Alexandre Reiffers-Masson, Joachim Flocon-Cholet, and Sandrine Vaton. Novel class discovery: An introduction and key concepts. *arXiv preprint arXiv:2302.12028*, 2023.

Aäron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Sagar Vaze, Kai Han, Andrea Vedaldi, and Andrew Zisserman. Generalized category discovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7492–7501, 2022.

Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5018–5027, 2017.

Enguang Wang, Zhimao Peng, Zhengyuan Xie, Fei Yang, Xialei Liu, and Ming-Ming Cheng. GET: Unlocking the multi-modal potential of CLIP for Generalized Category Discovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20296–20306, 2025a.

Hongjun Wang, Sagar Vaze, and Kai Han. HiLo: A learning framework for Generalized Category Discovery robust to domain shifts. In *International Conference on Learning Representations*, 2025b.

Shuo Wen and Maria Brbic. Cross-domain open-world discovery. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 52744–52761. PMLR, 2024.

Xin Wen, Bingchen Zhao, and Xiaojuan Qi. Parametric classification for Generalized Category Discovery: A baseline study. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 16590–16600, 2023.

Kaichao You, Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I. Jordan. Universal domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2720–2729, 2019.

Qing Yu, Daiki Ikami, Go Irie, and Kiyoharu Aizawa. Self-labeling framework for novel category discovery over domains. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 3161–3169, 2022.

Zelin Zang, Lei Shang, Senqiao Yang, Fei Wang, Baigui Sun, Xuansong Xie, and Stan Z. Li. Boosting novel category discovery over domains with soft contrastive learning and all in one classifier. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11858–11867, 2023.

Haiyang Zheng, Nan Pu, Wenjing Li, Nicu Sebe, and Zhun Zhong. Textual knowledge matters: Cross-modality co-teaching for generalized visual class discovery. In *European Conference on Computer Vision*, pp. 41–58. Springer, 2024.

A Additional Method Details

A.1 Dual Filtering

For completeness, we provide the detailed definitions used in the visual-aware candidate filtering step. Given the response distribution $q_{k|m}$ defined in Section 3.4, the normalized entropy of noun v_m is

$$H_e(v_m) = -\frac{1}{\log K_0} \sum_{k=1}^{K_0} q_{k|m} \log q_{k|m}. \tag{12}$$

The entropy-filtered pool is

$$\mathcal{V}_e = \{v_m \in \mathcal{V} \mid H_e(v_m) \le \gamma_{ent}\}. \tag{13}$$

Let $\mathbf{U}_e$ denote the text feature matrix of $\mathcal{V}_e$. The activation frequency of noun $v_m \in \mathcal{V}_e$ is

$$F(v_m) = \sum_i \mathbb{I}\left[v_m \in \mathrm{TopK}_{K_{fre}}\left(Z_i^\top \mathbf{U}_e\right)\right]. \tag{14}$$

Following the implementation, the frequency threshold is

$$\delta_{fre} = \left\lfloor \frac{N}{|\mathcal{V}_e|} K_{fre} \gamma_{fre} \right\rfloor, \tag{15}$$

where N is the number of image features used for filtering. The final filtered pool is

$$\mathcal{V}' = \{v_m \in \mathcal{V}_e \mid F(v_m) \ge \delta_{fre}\}. \tag{16}$$

A.2 Silhouette Scoring

Given the target assignment $\pi_{\mathcal{S}}(x_i^t)$ induced by a semantic space $\mathcal{S}$, let $\mathcal{C}_{\pi_i}$ be the set of target samples assigned to the same anchor as x_i^t. We compute

$$\begin{aligned} a_i &= \frac{1}{|\mathcal{C}_{\pi_i}| - 1} \sum_{x_j^t \in \mathcal{C}_{\pi_i}, j \ne i} d(Z_i^t, Z_j^t), \\ b_i &= \min_{c \ne \pi_i} \frac{1}{|\mathcal{C}_c|} \sum_{x_j^t \in \mathcal{C}_c} d(Z_i^t, Z_j^t). \end{aligned} \tag{17}$$

The silhouette score is

$$\mathrm{Sil}(\mathcal{S}) = \frac{1}{N_t} \sum_{i=1}^{N_t} \frac{b_i - a_i}{\max(a_i, b_i)}. \tag{18}$$

In implementation, $d(\cdot, \cdot)$ is computed in the normalized target visual feature space.

A.3 Local Candidate Construction

For each slot s_i, ViSTA retrieves semantically nearby nouns from the filtered pool:

$$\mathcal{N}_i = \mathrm{TopK}_{K_s}\left(\mathcal{V}', v \mapsto T_{s_i}^\top T_v\right). \tag{19}$$

To reduce redundant semantic anchors, candidates that are too similar to existing active slots are removed:

$$\mathcal{C}_i = \{s_i\} \cup \left\{ v \in \mathcal{N}_i \mid \max_{s_j \in \mathcal{S}_{-i}} T_v^\top T_{s_j} \le \gamma_{sim} \right\}. \tag{20}$$

Here, $\mathcal{S}_{-i} = \mathcal{S} \backslash \{s_i\}$. Including s_i in $\mathcal{C}_i$ allows the greedy update to keep the current slot unchanged.

A.4 Prediction-Level Contrastive Consistency

Let

$$\mathcal{Q} = \{\mathbf{p}_j^{t,1}\}_{j=1}^{B_t} \cup \{\mathbf{p}_j^{t,2}\}_{j=1}^{B_t} \tag{21}$$

denote all target prediction distributions in a mini-batch. Following the implementation, $\mathbf{p}_i^{t,2}$ is used as the anchor and $\mathbf{p}_i^{t,1}$ as its positive. The PCL loss is

$$\mathcal{L}_{pcl} = -\frac{1}{B_t}\sum_{i=1}^{B_t} \log \frac{\exp\left((\mathbf{p}_i^{t,2})^\top \mathbf{p}_i^{t,1}/\tau_p\right)}{\sum_{\mathbf{q}\in\mathcal{Q}\setminus\{\mathbf{p}_i^{t,2}\}} \exp\left((\mathbf{p}_i^{t,2})^\top \mathbf{q}/\tau_p\right)}. \tag{22}$$

This loss is applied after softmax prediction over $\mathcal{A}_t$, and therefore regularizes semantic assignment distributions rather than visual features.

B Implementation Details

B.1 Evaluation Metrics

Let $\mathcal{D}_t^{seen}$ and $\mathcal{D}_t^{unseen}$ denote target samples from seen and unseen classes, respectively. Seen Accuracy is computed on $\mathcal{D}_t^{seen}$ by direct comparison between predicted labels and ground-truth seen labels:

$$\mathrm{Acc}_{seen} = \frac{1}{|\mathcal{D}_t^{seen}|}\sum_{x_i\in\mathcal{D}_t^{seen}} \mathbb{I}[\hat{y}_i = y_i]. \tag{23}$$

For unseen classes, predictions are evaluated as clustering assignments. We compute a one-to-one matching between predicted unseen clusters and ground-truth unseen labels using the Hungarian algorithm, and then report the matched clustering accuracy as Unseen Accuracy. The H-Score is

$$\text{H-Score} = \frac{2\cdot \mathrm{Acc}_{seen}\cdot \mathrm{Acc}_{unseen}}{\mathrm{Acc}_{seen} + \mathrm{Acc}_{unseen}}. \tag{24}$$

Seen IDs are fixed, and Hungarian matching is restricted to novel output slots. Unseen samples assigned to seen slots remain in the Unseen denominator. The experiments in Appendix C use a 10^{-5} offset in the H-Score denominator on the fraction scale.

B.2 Hyperparameters

Table 7 summarizes the main implementation settings used by ViSTA. Stage 1 constructs the textual semantic space using frozen CLIP features. Stage 2 keeps the selected anchor vocabulary fixed and adapts both CLIP encoders with LoRA. The experiments are conducted on an A40 GPU.

Table 7: Implementation hyperparameters of ViSTA.

Component	Setting
Stage 1 search rounds	$n_{\text{inner}} = 3$
Stage 1 temperature	0.01
$\gamma_{ent}, K_{fre}, \gamma_{fre}, \gamma_{sim}$	$0.5, 5, 4, 0.9$
Stage 2 adaptation	LoRA on CLIP image and text branches
LoRA rank / alpha	$r = 8, \alpha = 8$
Prediction temperature	0.05
PCL temperature	$\tau_p = 0.3$
Loss weights	$\lambda_{ce} = 1.0, \lambda_{ent} = 1.0, \lambda_{div} = 2.0, \lambda_{pcl} = 1.0$
Optimizer	SGD with momentum 0.9
Learning rate / weight decay	5×10^{-4} / 5×10^{-4}
Batch size / validation batch size	64 / 128
Training iterations	6000
Learning-rate schedule	Cosine schedule with 50 warm-up iterations

C Vocabulary and Anchor Analysis

We examine how the category budget and candidate vocabulary affect the textual semantic space, then inspect the partitions induced by individual anchors. These frozen Stage-1 experiments use VisDA (6/6), CLIP ViT-B/16, and seed 2022.

C.1 Category Budget and Literal-Name Exclusion

Table 6 varies the category budget while keeping the six source anchors fixed. Both reducing and increasing the budget from the true total of 12 lowers H, with the larger loss under underestimation. This suggests that the number of available slots affects the balance between category separation and fragmentation.

To test dependence on literal class names, we exclude target-private names and direct WordNet synonyms from the noun pool and cached text features. Of 26 specified terms, 14 occur in the vocabulary. Removing them reduces the filtered pool from 584 to 486 candidates, yet leaves the final anchors and contingency counts unchanged. The selected textual anchors therefore organize these target categories through related vocabulary, even when their literal names are excluded.

C.2 What the Anchors Represent

Table 8 reports all six novel anchors. Purity measures the fraction of an anchor's samples belonging to its majority class; coverage measures the fraction of that class assigned to the anchor. For example, *rollerblading* captures most skateboard samples despite using a different name. Conversely, *bat boy* and *tennis player* both primarily capture person samples. The words provide useful assignment coordinates, but do not always correspond one-to-one to true categories.

Table 8: Novel anchors in the frozen $K = 12$ partition. Purity and coverage use the anchor majority class, with anchor size and true-class population as their respective denominators.

Anchor	Majority class	Size	Purity (%)	Coverage (%)
houseplant	plant	4342	95.32	90.99
rollerblading	skateboard	3269	62.47	89.52
passenger train	train	4369	92.01	94.90
bat boy	person	1782	76.09	33.90
tennis player	person	1245	69.56	21.65
garbage truck	truck	3235	81.82	47.71

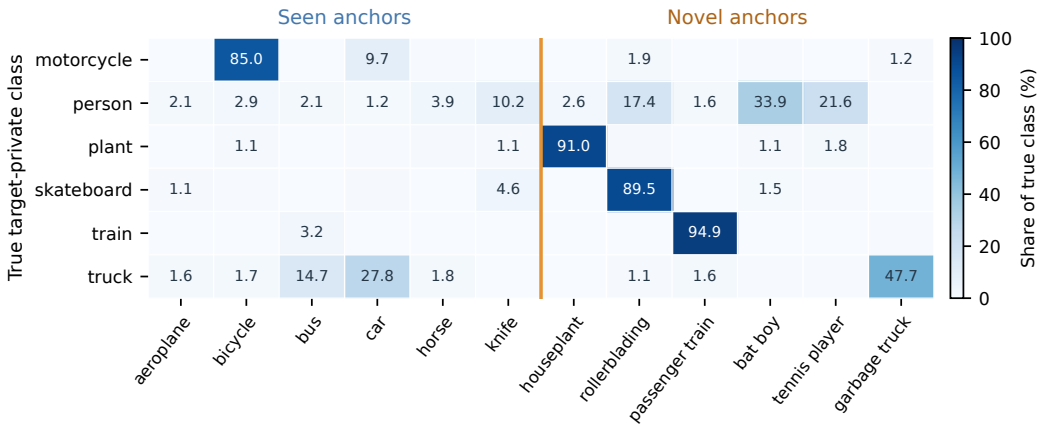


Figure 4: Frozen Stage-1 assignments for the six target-private classes. Rows are normalized by true-class population; columns show the selected anchors. The orange divider separates seen and novel slots. Cells with at least 1% mass are labeled.

Figure 4 reveals a second source of error: 36.57% of unseen samples enter seen slots. Motorcycle samples, for example, are mostly assigned to *bicycle*. Such errors cannot be corrected by matching labels within novel slots. This motivates adapting the image–text assignment function after selecting the vocabulary.

D Adapted GCD Baselines

D.1 SpectralGCD and the Mixed-Batch Class Prior

We adapt SpectralGCD (Caselli et al., 2026) to DS-GCD by training on labeled source and unlabeled target data. It uses a CLIP ViT-B/16 student and an additional frozen LAION CLIP ViT-H/14 teacher for concept selection. Training uses seed 2022, 200 epochs, source/target batches of 64 each, supervised weight 0.35, and maximum-entropy weight 2. We evaluate the final checkpoint with the metrics in Appendix B.1.

A uniform prior over the mixed batch can conflict with source supervision. On the 4/8 split, source samples occupy half the batch, whereas the uniform prior assigns only one third of its mass to seen classes. We compare this objective with a source-conditioned prior:

$$q_c = \frac{n_c^s + B_t/K_0}{B_s + B_t}, \tag{25}$$

where n_c^s is the number of source samples of class c in the batch and is zero for novel classes. This modified regularizer combines the observed source counts with a uniform target contribution and requires no target labels.

Table 9: SpectralGCD (Adapted) on VisDA after 200 epochs, seed 2022. Both objectives use a CLIP ViT-B/16 student and frozen LAION ViT-H/14 teacher. Values are percentages.

	Uniform prior			Source-conditioned prior		
Seen/unseen	Seen	Unseen	H	Seen	Unseen	H
4/8	0.01	57.98	0.02	62.88	72.78	67.47
6/6	16.16	74.73	26.57	57.93	47.82	52.39
8/4	20.84	60.60	31.01	64.72	34.39	44.91

The source-conditioned prior improves Seen Accuracy and H on all three splits. Unseen Accuracy increases on 4/8 but decreases on 6/6 and 8/4 (Table 9).