

RELATIONAL SEMANTIC GROUNDING FOR TEST-TIME DISCOVERY

Tianle Liu¹, Gongbo Cheng¹, Chenggong Ni², Minming Gu¹, Tao Zhou³,
Fuyuan Hu^{*1}, Fan Lyu^{*4}

¹School of Electronics and Information Engineering, Suzhou University of Science and Technology

²School of Intelligence Science and Technology, Nanjing University

³School of Computer Science Engineering, North Minzu University

⁴Computer Vision Center, Universitat Autònoma de Barcelona

tianleliu@post.usts.edu.cn

fanlyu@cvc.uab.cat fuyuanhu@mail.usts.edu.cn

ABSTRACT

Test-Time Discovery (TTD) aims to recognize known classes and detect novel ones from a continuous unlabeled test stream after deployment. Existing TTD methods mainly rely on visual clustering or prototype comparison, and can only assign anonymous labels to discovered classes, leaving them difficult to interpret in open-world applications. In this paper, we propose SEGA, a Semantic Grounding with Attributes framework for TTD. SEGA leverages large language models to generate attribute-rich descriptions for known classes and constructs an attribute-anchored visual-semantic space during training. At test time, we introduce Relational Semantic Grounding (RSG), which grounds unknown samples with similar and dissimilar known-class anchors, prompts the LLM to generate candidate concepts, and reranks them in the learned semantic space. The selected concept is further used for confidence-guided test-time semantic stabilization. By coupling attribute-level semantic alignment with relational concept grounding, SEGA transforms anonymous test-time clusters into interpretable semantic concepts. Experiments show that SEGA improves TTD baselines with the promise of semantic grounding for open-world. Our code will be available at <https://github.com/LeTianL-TT/SEGA-for-TTD>.

1 INTRODUCTION

Novel Class Discovery (NCD) (Pu et al., 2023; Wu et al., 2023) aims to discover unseen categories by transferring knowledge from labeled known classes to an unlabeled data pool. However, existing NCD methods assume that the unlabeled data are available in advance, making discovery an offline process. This assumption becomes restrictive in open-world deployment, where new categories may continuously emerge in applications. Test-Time Discovery (TTD) (Lyu et al., 2025b) addresses this setting by requiring the model to recognize known classes, discover novel ones, and update its category space from a continual unlabeled test stream during inference. In traditional NCD, unknown classes are discovered offline and represented by semantic-agnostic cluster IDs (such as “#1”), making immediate semantic interpretation less necessary. TTD, however, discovers novel classes during deployment, where the discovered classes should be understandable and reusable for humans or subsequent inference rather than merely separated as new clusters. Existing TTD methods (Lyu et al., 2025b; Liu et al., 2025) largely inherit the NCD-style treatment, assigning semantic-agnostic IDs to unknown classes rather than grounding them as meaningful concepts, which limits their practical value (Fig. 1a). Some semantic-guided NCD methods (Wang et al., 2024; Zheng et al., 2024b) introduce textual information through open-vocabulary recognition or fixed lexicon retrieval (Fig. 1b). However, these strategies assume that candidate novel categories can be covered by a predefined vocabulary, which is misaligned with open-world TTD where possible novel classes are unknown before deployment. A simple alternative is to use vision-language models with class-name

*Corresponding authors.

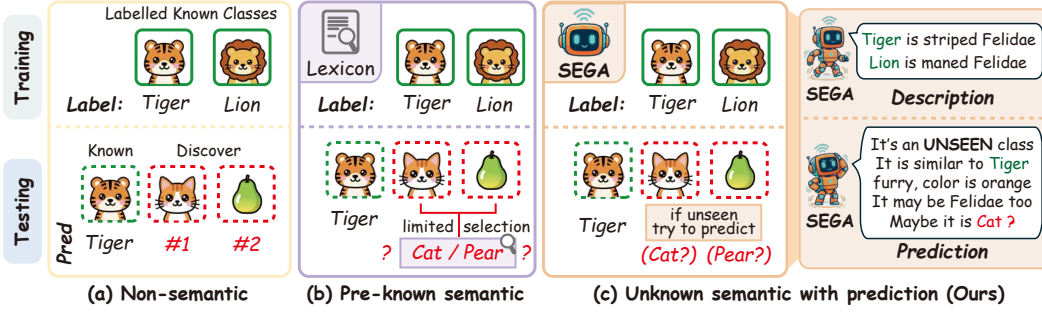


Figure 1: Comparison among different discovery method. (a) Non-semantic discovery, where unknown samples are detected or clustered online but are only assigned anonymous placeholder labels. (b) Pre-known semantic discovery handles unknown sample by matching visual features to a predefined category vocabulary, but it is unrealistic and its discovery ability is limited. (c) Our method aims to more meaningful and open-ended test-time discovery where unknown classes are grounded with visual and semantic relations to known classes, and interpreted through concept generation.

prompts such as “a photo of a [class]”, but such prompts provide only coarse semantic anchors and lack fine-grained attributes for relating unknown classes to known ones (Tan et al., 2026a; Lyu et al., 2026a;b). LLMs, with their rich conceptual and attribute-level knowledge, offer a more flexible source of semantic cues for open-ended discovery (Du et al., 2024; Fu et al., 2025). This suggests that attribute-rich descriptions generated by LLMs can provide a semantic basis for TTD, allowing known-class attributes to serve as transferable bridges for grounding novel classes at test time.

In this paper, we propose **Semantic Grounding with Attributes (SEGA)**, a framework that transforms TTD from anonymous clustering into semantic concept generation (Fig. 1c). During training, SEGA first uses LLM-generated attribute descriptions to enrich known classes and learns an attribute-aware visual-semantic space through cross-modal alignment. At test time, when an unknown-class candidate is detected, we introduce Relational Semantic Grounding (RSG) to describe it by its relations to known-class anchors and visual attributes, and then query an LLM to generate candidate category names. Since free-form LLM generation can produce overly broad or implausible names, SEGA reranks the candidates according to their plausibility in the learned semantic space. The selected concept is further converted into a pseudo description for confidence-guided test-time training, which stabilizes subsequent discovery and reduces the effect of noisy semantic predictions. By coupling attribute-level semantic alignment, relational concept generation, and semantic stabilization, SEGA provides an interpretable TTD pipeline that grounds discovered unknown classes as meaningful concepts rather than anonymous tokens. Experiments on four TTD benchmarks show that SEGA improves discovery performance over recent TTD baselines, with up to 4.55% gain in agreement metrics. In summary, our contributions are:

- (1) We identify and address the semantic grounding problem in TTD, where discovered unknown classes should be represented as meaningful concepts rather than anonymous cluster IDs, and propose SEGA as a unified framework for semantic test-time discovery.
- (2) We introduce RSG, which combines LLM-generated known-class attributes, relational semantic anchors, and semantic reranking to generate reliable candidate names for unknown classes.
- (3) We evaluate SEGA on four TTD benchmarks, showing improvements over prior TTD methods in discovery performance, known-class stability, and semantic quality of the generated class names.

2 RELATED WORK

From Semantics-Agnostic to Semantic-Guided Category Discovery. (Han et al., 2020; Zhang et al., 2023; Zhong et al., 2021) Novel Class Discovery (NCD) aims to identify new categories from unlabeled data by transferring knowledge from known classes. GCD (Vaze et al., 2022) and PromptCCD (Cendra et al., 2024) replace the traditional k-means stage with contrastive learning prompting for purer prototypes. On-the-fly Category Discovery (OCD) (Du et al., 2023; Zheng et al., 2024a) pushes NCD toward real-time operation by hashing images and assigning them according to Hamming-distance. Later, some work start to explore incorporating textual information into category discovery to address the semantics gap. SNCD (Wang et al., 2024) cluster unlabeled images but also

semantically match them by constructing semantic similarity matrices. Some CLIP-guided GCD works (Zheng et al., 2024b; Wang et al., 2025) leverage pre-trained vision-language alignment to guide discovery with semantic priors. And the open-vocabulary paradigm (Wu et al., 2024; Zhu & Chen, 2024) has further expanded the semantic horizon. However, these approaches operate in offline or static settings and depend on predefined or retrieved textual descriptions.

Test-Time Discovery (TTD) requires models to detect and learn novel classes in real-time as samples arrive sequentially. The pioneering HM (Lyu et al., 2025b) proposes a training-free approach using locality-sensitive hashing for efficient sample retrieval and prototype comparison. DAA (Liu et al., 2025) introduces a lightweight trainable adapter to amplify feature-level discrepancies between known and unknown classes. While these methods achieve impressive performance in visual feature discrimination, discovered classes are assigned placeholder labels without any interpretable meaning.

LLM for Visual Knowledge. LLM-generated descriptions for visual recognition are widely used in different work recently (Zhang et al., 2025; Lei et al., 2025; Fu et al., 2025). Detailed visual descriptions are revealed to outperform fixed templates and can improve task like zero-shot, open-vocabulary and attribute-concept annotation (Ding et al., 2026; Mei et al., 2026; Tan et al., 2026b). The dynamic, streaming nature of TTD with its real-time constraints and irreversible decisions remains unexplored in the context of LLM-assisted semantic discovery (Lyu et al., 2025a; Ni et al., 2025; Tan et al., 2026c). Our work bridges this gap by enabling test-time LLM reasoning for semantic interpretation, transforming TTD from visual clustering to conceptual understanding.

3 METHOD

3.1 TTD AND A CLIP-BASED BASELINE

TTD aims to recognize known classes and discover unknown ones in a continuous unlabeled stream after model deployment. Formally, let $\mathcal{D}_{\text{train}} = \{(x_i, y_i)\}_{i=1}^N$ denote the training set, where $x_i \in \mathcal{X}$ is input image, and $y_i \in \mathcal{Y}_{\text{kn}}$ is its number label from the known class set $\mathcal{Y}_{\text{kn}}$. During testing, the model encounters a test stream $\mathcal{D}_{\text{test}} = \mathcal{D}_{\text{kn}} \cup \mathcal{D}_{\text{un}}$, where $\mathcal{D}_{\text{un}}$ may contain unknown classes that were never seen during training. TTD requires the model to satisfy two objectives simultaneously. First, *novel class discovery and learning*, where inputs $x \in \mathcal{D}_{\text{test}}^{\text{un}}$ must be identified as belonging to unseen classes and assigned new labels to distinguish them from known-class samples. Second, *unified classification*, where all inputs $x \in \mathcal{D}_{\text{test}}$ are accurately classified into either a known class $y \in \mathcal{Y}_{\text{kn}}$ or a novel class $y \in \mathcal{Y}_{\text{seen}}$, with $\mathcal{Y}_{\text{seen}}$ denoting the set of dynamically discovered novel classes.

Inspired by recent vision-language discovery methods (Zheng et al., 2024b; Wang et al., 2025), we adopt CLIP (Radford et al., 2021) as the backbone. To preserve linguistic knowledge, we freeze the pretrained text encoder and fine-tune the image encoder, with lightweight learnable projection layers attached to both modalities. The projection layers map image and text features into a shared attribute-aware space and are updated during training and test-time adaptation, enabling flexible alignment without modifying the frozen text encoder.

Standard CLIP-based methods often use class-name prompts such as “a photo of [CLASS]”, which provide only coarse semantic anchors. As shown in Fig. 2, such template-based text embeddings are dominated by class names and may entangle semantically related but visually distinct categories. In the following subsection, we instead enrich each known class with LLM-generated attribute descriptions, following the intuition that attributes provide transferable semantic cues across categories (Du et al., 2024; Zheng et al., 2024b). Attribute descriptions capture fine-grained cues such as shape, color, texture, and parts, helping distinguish similar classes while preserving their semantic relations.

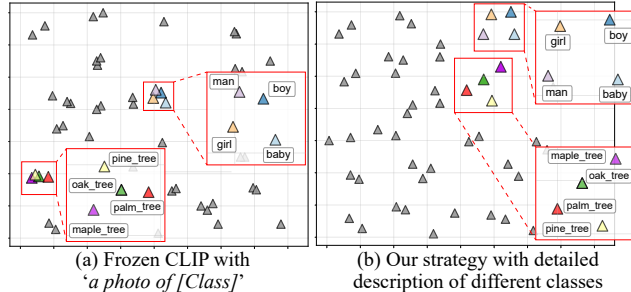


Figure 2: T-SNE visualization of the CLIP-based textual feature for different kinds of text prompt.

3.2 ATTRIBUTE-ANCHORED SEMANTIC SPACE CONSTRUCTION

Based on the above analysis, we first construct an attribute-anchored visual-semantic space using known-class training data to support semantic grounding at test time. For each known class $c \in \mathcal{Y}_{\text{kn}}$, we query an LLM, e.g., GPT-4o-mini, to generate an attribute-rich description t_c that summarizes discriminative visual properties such as color, shape, texture, and parts¹. This yields a set of textual anchors $\mathcal{T} = \{t_c\}_{c \in \mathcal{Y}_{\text{kn}}}$, which provide semantic references for known classes². The text encoder is frozen to preserve its linguistic knowledge, while the image encoder and lightweight projection layers are trainable. Given an image x and the textual anchor t_c of its class, we obtain

$$\mathbf{z}_v(x) = P_v(E_v(x); W_v) \quad \mathbf{z}_t(t_c) = P_t(E_t(t_c); W_t), \quad (1)$$

where P_v and P_t map both modalities into a shared attribute space $\mathcal{Z}$.

Semantic anchor calibration. The LLM-generated descriptions are not used as ordinary class labels. Actually, they act as semantic anchors $\mathbf{a}_c = \mathbf{z}_t(t_c)$, and all known-class anchors form an anchored textual manifold $\mathcal{M}_t = \{\mathbf{a}_c \mid c \in \mathcal{Y}_{\text{kn}}\}$ that define how known classes are arranged in the attribute space. However, directly using descriptions as anchors can be unstable, since pretrained models are not inherently arranged for rich descriptions. We therefore *calibrate* these textual anchors so that descriptions of the same-class textual neighbors $\mathcal{N}_t^+(i) = \{j \in \mathcal{B} : y_j = y_i\}$ collapse into compact semantic regions, while anchors from different classes $\mathcal{N}_t^-(i) = \{j \in \mathcal{B} : y_j \neq y_i\}$ occupy separable territories. A well-calibrated anchor space therefore provides a reliable coordinate system for subsequent relational inference, rather than merely serving as a set of text classifiers. Specifically, given a test batch $\mathcal{B}$, we ensure that the text embedding space preserves class-level discriminability with attribute-rich descriptions by

$$\mathcal{L}_{\text{text}} = -\frac{1}{|\mathcal{B}|} \sum_{i=1}^{|\mathcal{B}|} \log \frac{\sum_{j: y_j = y_i} \exp[\mathbf{z}_t(t_i)^\top \mathbf{z}_t(t_j)]}{\sum_{c \in \mathcal{B}} \exp[\mathbf{z}_t(t_i)^\top \mathbf{z}_t(t_c)]}. \quad (2)$$

Visual neighborhood structuring. While $\mathcal{M}_t$ provide the semantic coordinate system, the visual stream also needs a stable neighborhood structure before it can support discovery. Let $\mathbf{v}_i = \mathbf{z}_i(x_i)$ denote the visual feature of image x_i and if visual manifolds $\mathcal{M}_v = \{\mathbf{v}_i\}$ are scattered or overly fitted to known-class decision boundaries, an unseen sample may be absorbed by a known class or isolated as noise. We therefore organize $\mathcal{M}_v$ with a dual objective by

$$\mathcal{L}_{\text{img}} = -\frac{1}{|\mathcal{B}|} \sum_{i=1}^{|\mathcal{B}|} \left(\log \frac{\exp[\mathbf{z}_i(x_i)^\top \tilde{\mathbf{z}}_i(x_i)]}{\sum_{c \in \mathcal{B}} \exp[\mathbf{z}_i(x_i)^\top \mathbf{z}_i(x_c)]} + \log \frac{\sum_{j: y_j = y_i} \exp[\mathbf{z}_i(x_i)^\top \mathbf{z}_i(x_j)]}{\sum_{c \in \mathcal{B}} \exp[\mathbf{z}_i(x_i)^\top \mathbf{z}_i(x_c)]} \right), \quad (3)$$

where $\tilde{\mathbf{z}}_i$ denotes the feature of an augmented view of x_i . In this way, the visual representation is made robust to visual variations and the image encoder learns a visual manifold that is suitable for both recognizing known classes and detecting deviations caused by emerging unknown classes.

Cross-modal coordinate binding. The calibrated textual manifold $\mathcal{M}_t$ and the structured visual manifold $\mathcal{M}_v$ are useful only if they share the same coordinate system. To bridge this gap, we build image-anchor correspondences $\mathcal{G} = \{(\mathbf{v}_i, \mathbf{a}_{y_i})\}_{i \in \mathcal{B}}$, where each visual feature is paired with the semantic anchor of its class. Therefore, we further introduce a cross-modal binding objective along $\mathcal{G}$ connecting each image to its corresponding attribute-rich textual anchor without which the image encoder may rely on spurious visual correlations. This binding encourages visual features to inherit the semantic directions encoded by the LLM descriptions, such as shape, texture, color, and part-level cues. Once bound, visual representations are no longer determined purely by appearance and unknown sample can then be described through its position relative to known semantic anchors because they share the same semantic coordinates in the attribute space. This binding is achieved via:

$$\mathcal{L}_{\text{cross}} = -\frac{1}{2|\mathcal{B}|} \sum_{i=1}^{|\mathcal{B}|} \left(\log \frac{\exp[\mathbf{z}_i(x_i)^\top \mathbf{z}_t(t_i)]}{\sum_{c \in \mathcal{B}} \exp[\mathbf{z}_i(x_i)^\top \mathbf{z}_t(t_c)]} + \log \frac{\exp[\mathbf{z}_t(t_i)^\top \mathbf{z}_i(x_i)]}{\sum_{c \in \mathcal{B}} \exp[\mathbf{z}_t(t_i)^\top \mathbf{z}_i(x_c)]} \right). \quad (4)$$

¹ Such as "Describe [CLASS] in detail, including visual attributes like color, shape...".

² "A zebra has black and white stripes, a horse-like shape..."

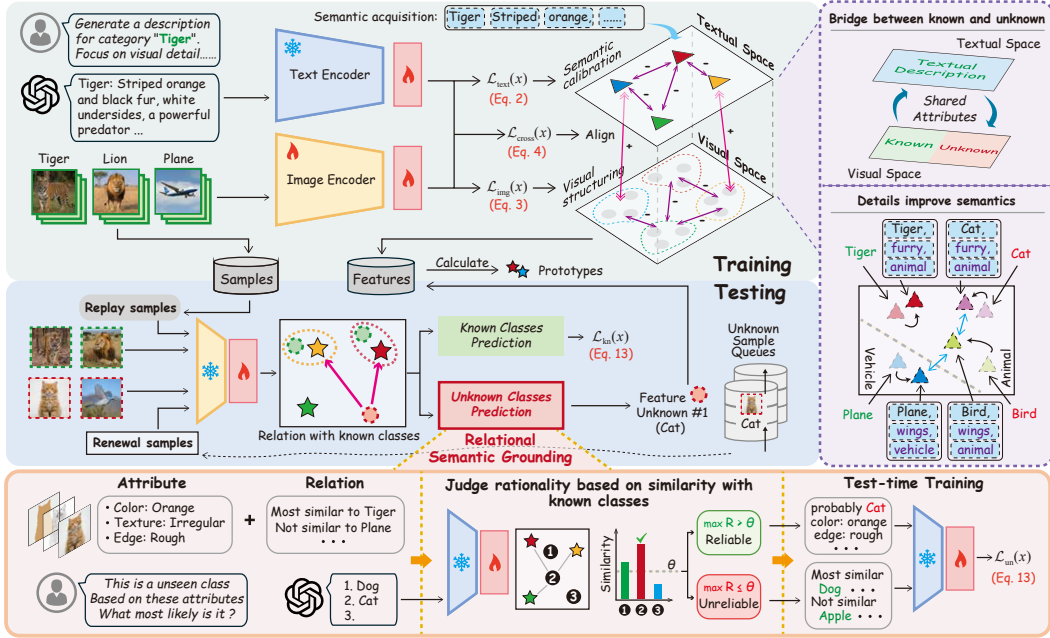


Figure 3: Method Schema. During training, LLM-generated attribute descriptions are used to construct a semantic space and guide cross-modal alignment between images and text. During testing, we discover unknown classes, retrieve their semantic relations to known class anchors, and query the LLM to generate candidate category names. The candidates are then grounded and filtered by the RSG module, and the selected prediction is used to guide confidence-aware test-time updates.

The three objectives play complementary roles which together prepare the model for testing where known classes become stable semantic references, while unknown samples can be interpreted by their relational position in the anchored visual-semantic space. The total training loss is:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{text}} + \mathcal{L}_{\text{img}} + \mathcal{L}_{\text{cross}}. \quad (5)$$

3.3 RELATIONAL CONCEPT-LEVEL SEMANTIC INFERENCE

When encountering the test sample x with a known or unknown class, the prediction procedure involves calculating the cosine similarity $s_k = \text{sim}(\mathbf{z}_i(x), \mathbf{p}_c)$ between the image feature $\mathbf{z}_i(x)$ and the prototypes $\mathbf{p}$ of known classes to determine their classification. Based on the max similarity scores, we apply a confidence threshold γ to determine whether a test sample x belongs to a seen class or an unseen class.

$$\hat{y} = \begin{cases} \arg \max_{c \in (\mathcal{Y}_{\text{kn}} \cup \mathcal{Y}_{\text{un}})} (s_k), & \text{if } \max_c s_k > \gamma, \\ \text{new discovered class}, & \text{otherwise.} \end{cases} \quad (6)$$

Prompt construction and candidate generation. If the sample is classified as an unseen class $\hat{y} \in \mathcal{Y}_{\text{unseen}}$, two sets Top- k most similar classes $\mathcal{S}_{\text{sim}}$ and Top- k least similar classes $\mathcal{S}_{\text{dis}}$ are extracted. These similarity patterns serve as relational anchors, characterizing the unknown class by what it resembles and what it differs from in the known semantic space.

$$\mathcal{S}_{\text{sim}} = \{(c, s_k) \mid c \in \text{top-}k \text{ of } s_k\} \quad \mathcal{S}_{\text{dis}} = \{(c, s_k) \mid c \in \text{bottom-}k \text{ of } s_k\} \quad (7)$$

Relational similarity alone may be insufficient for fine-grained naming. To provide additional discriminative cues, we extract interpretable visual attributes $\mathcal{A}$ from the image x , which include color statistics (dominant hue, saturation, temperature), texture properties (roughness, regularity), lighting conditions (brightness, uniformity), and structural characteristics (edge density, composition). These attributes are computed using only low-level operations on the raw images.

$$\mathcal{A} = \Phi_{\text{attr}}(\mathbf{x}) = \{\mathcal{A}_{\text{color}}, \mathcal{A}_{\text{texture}}, \mathcal{A}_{\text{lighting}}, \mathcal{A}_{\text{structure}}\} \quad (8)$$

Then we construct a prompt $\mathcal{P} = (\mathcal{S}_{\text{sim}}, \mathcal{S}_{\text{dis}}, \mathcal{A})$ that integrates both relational and attribute information to instruct the LLM to generate M candidate category names for the discovered class:

$$\mathcal{C}_{\text{LLM}} = \{\hat{c}_1, \hat{c}_2, \dots, \hat{c}_M\} = \text{LLM}(\mathcal{P}(\mathcal{S}_{\text{sim}}, \mathcal{S}_{\text{dis}}, \mathcal{A})) \quad (9)$$

Semantic reranking with text encoder. Some of these candidates are plausible but potentially noisy. To ground the LLM’s free-form generation in the actual visual-semantic space, we propose a semantic reranking mechanism. We encode both the LLM-generated candidates and the top- k most similar known classes with our text encoder. And the relevance score R for each candidate $\hat{c}$ is computed as the average similarity with the known reference classes.

$$R(\hat{c}) = \frac{1}{|\mathcal{S}_{\text{sim}}|} \sum_{c \in \mathcal{S}_{\text{sim}}} \mathbf{z}_t(\hat{c}) \cdot \mathbf{z}_t(y_c), \quad \hat{c} \in \mathcal{C}_{\text{LLM}} \quad (10)$$

The top-ranked candidate serves as the semantic label $\hat{y}_t$ for the discovered class, and this class will become a seen class $\mathcal{Y}_{\text{seen}} \leftarrow \hat{y}_t$, and the model will treat this sample as the first instance of this new seen class. The image feature $\mathbf{z}_i(x)$ will be used as the initial prototype for this class, denoted as $\mathbf{p}_{\hat{y}_t}$. Subsequent samples will be compared to all known and seen prototypes, including this new $\mathbf{p}_{\hat{y}_t}$.

$$\hat{y}_t = \arg \max_{\hat{c} \in \mathcal{C}_{\text{LLM}}} R(\hat{c}) \quad (11)$$

3.4 CONFIDENCE-GUIDED TEST-TIME SEMANTIC STABILIZATION

We acknowledge that precisely predicting the ground-truth label of a never-before-seen class is inherently challenging due to the open-world nature of the task. By introducing LLMs, we take an initial step toward addressing the lack of semantic interpretability which is a fundamental challenge intrinsic to TTD. However, we believe that while accurate label prediction is the ultimate goal for downstream targets, it is not entirely strict for auxiliary classification. What matters more is that the discovered semantic label $\hat{y}_t$ occupies a reasonable position in the text embedding space.

Therefore, we propose a reliability strategy for test-time training guided by the semantic reranking score $R(\hat{c})$. A high R indicates that the top candidate $\hat{y}_t$ is semantically consistent with known reference classes, suggesting a stable prediction. Conversely, a low R implies ambiguous generation. For a sample x with a stable prediction, we construct an attribute-rich textual description $\tilde{t}_{\hat{y}_t}$ with the predicted label $\hat{y}_t$. While a sample with an uncertain prediction, we avoid using a potentially incorrect label. Instead, we construct a relational textual prompt based solely on the similarity set $\mathcal{S}_{\text{sim}}$ and dissimilarity set $\mathcal{S}_{\text{dis}}$.

$$\tilde{t}_{\hat{y}_t} = \begin{cases} \mathcal{P}(\hat{y}_t, \mathcal{A}, \mathcal{S}_{\text{sim}}, \mathcal{S}_{\text{dis}}), & \text{if } \max_{\hat{c} \in \mathcal{C}_{\text{LLM}}} R(\hat{c}) > \theta, \\ \mathcal{P}(\mathcal{S}_{\text{sim}}, \mathcal{S}_{\text{dis}}), & \text{otherwise.} \end{cases} \quad (12)$$

For a test batch $\mathcal{B}$, the first type of loss applies to samples with the pseudo labels:

$$\begin{aligned} \mathcal{L}_{\text{kn}}(x, t) &= \mathbb{E}_{\hat{y} \in \mathcal{Y}_{\text{kn}}} \left[\mathcal{L}_{\text{text}}(t | \hat{y}) + \mathcal{L}_{\text{img}}(x | \hat{y}) + \mathcal{L}_{\text{cross}}((x, t) | \hat{y}) \right], \\ \mathcal{L}_{\text{un}}(x, \tilde{t}_{\hat{y}_t}) &= \mathbb{E}_{\hat{y} \in \mathcal{Y}_{\text{un}}} \left[\mathcal{L}_{\text{text}}(\tilde{t}_{\hat{y}_t} | \hat{y}) + \mathcal{L}_{\text{img}}(x | \hat{y}) + \mathcal{L}_{\text{cross}}((x, \tilde{t}_{\hat{y}_t}) | \hat{y}) \right]. \end{aligned} \quad (13)$$

The second kind of loss follows DAA (Liu et al., 2025) to conduct replay for known classes and renewal for seen classes by retraining on samples stored in a memory bank. Since both initial predictions and early-stage classifications are inevitably error-prone, we incorporate a self-correction mechanism as DAA does to, which in the TTD setting can effectively refine subsequent classification performance. Specifically, we retrieve the samples regularly, and these samples would be reprocessed, re-evaluated, and update SEGA again. The overall testing phase loss is shown below:

$$\mathcal{L}_{\text{test}} = \lambda_1 \cdot \mathbb{E}_{\hat{y} \in \mathcal{Y}_{\text{kn}}} \left[\mathcal{L}_{\text{kn}}(x, t | \hat{y}) + \mathcal{L}_{\text{kn}}^{\text{replay}}(x) \right] + \lambda_2 \cdot \mathbb{E}_{\hat{y} \in \mathcal{Y}_{\text{seen}}} \left[\mathcal{L}_{\text{un}}((x, \tilde{t}_{\hat{y}_t}) | \hat{y}) + \mathcal{L}_{\text{un}}^{\text{renew}}(x | \hat{y}) \right], \quad (14)$$

where λ_1 and λ_2 are hyperparameters that balance the contributions of each loss term.

Table 1: Comparisons on four datasets. Real-time evaluation reflects the accumulated performance across all test batches, while post evaluation reassesses all test samples after testing. **Bold** and underlining denote the best and second-best results.

	Method	Real-time Evaluation			Post Evaluation			
		KA↑	TA↑	CA↑	KA↑	TA↑	CA↑	KF↓
CIFAR100D	L2P (Wang et al., 2022b)	59.93±2.15	8.57±2.49	43.10±4.30	50.53±7.25	9.60±1.50	27.39±2.11	27.85±7.21
	DP (Wang et al., 2022a)	66.09±1.01	8.80±1.69	48.34±6.78	56.19±2.00	8.68±2.06	28.93±2.25	29.06±1.99
	GMP (Cendra et al., 2024)	72.77±1.20	7.37±0.88	42.26±4.54	67.21±2.53	13.04±2.59	27.10±1.77	17.69±2.53
	PHE (Zheng et al., 2024a)	72.35±1.12	16.74±1.05	37.34±1.31	70.12±1.07	14.55± 1.01	24.65±1.35	2.12 ±0.92
	HM (Lyu et al., 2025b)	79.17± 0.13	21.13± 0.62	56.37 ±1.42	80.73±1.59	31.03±1.24	34.81±1.22	3.41±1.49
	DAA (Liu et al., 2025)	80.81 ±0.44	31.94±0.98	45.38± 0.88	80.27 ± 0.60	35.63±1.11	37.31 ± 1.20	4.10±0.60
	Ours	83.49 ±0.64	34.88 ±0.92	47.94± <u>1.04</u>	82.78 ± 0.89	38.64 ±1.20	39.15 ±1.24	<u>3.04</u> ± 0.54
CUB200D	L2P (Wang et al., 2022b)	46.22±1.53	9.01±0.87	55.37±7.79	31.97±3.35	4.75±0.73	24.48±1.65	42.29±3.14
	DP (Wang et al., 2022a)	53.69±1.24	43.68±2.20	45.68±5.88	63.37±3.27	45.94±0.91	7.92±1.10	5.85±3.11
	GMP (Cendra et al., 2024)	62.97±1.33	46.44±1.87	47.99±4.34	58.11±3.00	48.02±1.20	10.31±1.45	5.46±2.77
	PHE (Zheng et al., 2024a)	57.29±1.23	32.68±1.66	57.22±2.75	54.77±1.75	27.92± 0.43	21.36±1.36	2.18 ±0.71
	HM (Lyu et al., 2025b)	66.20± <u>0.55</u>	<u>58.30</u> ±2.37	43.33±6.10	64.42± <u>0.65</u>	65.28 ±1.78	37.25±4.90	4.07±0.47
	DAA (Liu et al., 2025)	68.09 ± 0.33	54.49±1.03	58.34 ± <u>1.35</u>	<u>66.26</u> ± 0.39	64.13±2.23	<u>44.65</u> ± 0.88	3.60± 0.39
	Ours	68.68 ±1.21	58.30 ± 0.62	<u>57.72</u> ± 1.01	67.73 ±0.94	64.25±1.63	46.00 ±1.22	<u>2.83</u> ±0.62
T-ImageNetD	L2P (Wang et al., 2022b)	46.25±1.41	7.79±2.92	53.55±6.47	29.50±3.77	10.38±2.42	23.28± 0.79	47.97±3.77
	DP (Wang et al., 2022a)	46.51± 0.58	6.41±0.93	58.27±6.10	28.53±3.33	9.63±2.10	26.80±1.38	47.57±3.32
	GMP (Cendra et al., 2024)	62.47±1.40	6.25±1.72	58.02±4.29	63.95±2.04	15.64±2.63	26.31±2.33	16.86±2.04
	PHE (Zheng et al., 2024a)	63.28±1.29	13.54±1.20	70.33±3.80	58.39± 1.14	12.80± 1.05	19.42±1.45	1.88± 0.35
	HM (Lyu et al., 2025b)	75.31±1.31	16.04± <u>0.76</u>	73.81 ±2.67	74.94±2.20	16.23± <u>1.24</u>	<u>37.43</u> ±1.30	1.15 ±2.18
	DAA (Liu et al., 2025)	<u>76.38</u> ±0.82	24.10±0.80	71.70±2.02	<u>75.50</u> ±1.96	<u>23.17</u> ±1.35	36.41±1.12	2.39±1.56
	Ours	77.37 ±1.53	26.01 ± 0.75	<u>72.72</u> ± 1.74	76.20 ± <u>1.93</u>	25.77 ±1.55	37.69 ±1.27	<u>1.65</u> ± <u>0.46</u>
AircraftD	L2P (Wang et al., 2022b)	43.21±1.08	20.11±1.20	35.52±1.93	40.17±2.34	21.00±1.21	27.21± <u>1.33</u>	15.04±2.10
	DP (Wang et al., 2022a)	40.18±1.10	21.88±1.23	34.80±1.79	42.53±1.46	20.22± 0.96	28.15±1.74	12.68±1.74
	GMP (Cendra et al., 2024)	41.26± <u>1.04</u>	21.12±0.67	35.10±2.50	32.56±1.75	21.25± <u>1.02</u>	26.34±1.48	22.65±2.25
	PHE (Zheng et al., 2024a)	45.82± 1.02	22.56±1.05	41.72±2.13	47.28±2.88	25.09±1.76	24.38±1.67	7.93±1.13
	HM (Lyu et al., 2025b)	<u>50.35</u> ±1.08	<u>28.01</u> ±0.85	40.23±2.22	<u>49.77</u> ±1.94	29.80±1.25	28.74±1.53	<u>5.44</u> ±1.56
	DAA (Liu et al., 2025)	49.37±1.46	27.79± 0.31	40.51± 1.05	48.87±1.73	28.37±1.69	28.99±1.69	5.61± <u>1.04</u>
	Ours	51.27 ±1.51	28.53 ± <u>0.32</u>	<u>41.84</u> ± <u>1.38</u>	50.84 ± 1.16	30.26 ±1.17	30.44 ± 1.23	4.42 ± 0.74

4 EXPERIMENT

4.1 EXPERIMENTAL DETAILS

Datasets. We conduct our experiments based on four benchmark datasets. CIFAR100 (C100) (Krizhevsky et al., 2009) is a commonly used classic dataset. Tiny ImageNet (Le & Yang, 2015) is a subset of ImageNet (Deng et al., 2009). Caltech-UCSD Birds-200-2011 (CUB) (Wah et al., 2011) and FGVC-Aircraft (Maji et al., 2013) are complex fine-grained datasets. All these datasets are split into known and unknown classes (7:3). The model is trained on the known training set and tested on the mixture of known and unknown test sets. Because the dataset is used for discovery, we follow the names of four transformed datasets from HM (Lyu et al., 2025b): CIFAR100D, CUB-200D, Tiny-ImageNetD, and AircraftD.

Implementation details. In our implementation, we build our method on CLIP (Radford et al., 2021) as the image and text encoder, which employs a ViT-B/16 backbone (Khan et al., 2022) following the pretraining procedure of GCD and TTD work. A linear projection layer is added after the encoders for each modality and only the last layers of image encoders in the backbone are fine-tuned on the known classes, using SGD optimizer and cosine decay learning rate scheduler with an initial learning rate of 0.1 and minimum learning rate of 0.0001, and weight decay of 0.00005. All input images are resized to 224×224 and augmented to match the pretrained backbone settings.

Evaluation metrics. Following (Lyu et al., 2025b), we use the following metrics. For unknown classes, we use two agreement metrics: (1) *True-label Agreement ratio (TA)*. This metric measures the maximum proportion of samples from a given true class that are predicted as the same class; and (2) *Cluster Agreement ratio (CA)*. This metric measures the maximum proportion of samples from a given predicted cluster that have the same true label. For known classes, we also employ two key metrics to comprehensively assess the model’s performance: *Known Accuracy (KA)* and *Known Forgetting (KF)*. KA measures the traditional classification accuracy of the model on known classes while KF quantifies the degree of performance degradation on known classes over time. To quantitatively assess the quality of LLM-generated unknown class names, we introduce the *Semantic Score (S²)*. This metric evaluates how semantically close the predicted class names are to the ground-

truth labels in vector space via a pretrained Sentence-BERT encoder (Reimers & Gurevych, 2019),

$$S^2(\hat{c}, y) = \frac{f_{\text{BERT}}(\hat{c}) \cdot f_{\text{BERT}}(y)}{\|f_{\text{BERT}}(\hat{c})\|_2 \|f_{\text{BERT}}(y)\|_2}.$$

4.2 MAJOR COMPARISONS

In this paper, we first compare our methods with naive thresholding-based training-free methods. When exceeding the threshold, the naive method will be considered to have discovered a new class. And we compare with some training-required methods including L2P (Wang et al., 2022b), DP (Wang et al., 2022a), and GMP (Cendra et al., 2024) as these methods update prompts like TTA and CL methods. We also compare with the recent PHE (Zheng et al., 2024a) method for OCD and HM (Lyu et al., 2025b), DAA (Liu et al., 2025) for TTD.

Comparisons on TA & CA. Table 1 presents a comprehensive comparison of our method against various TTD approaches across four benchmarks. Training-based methods often suffer from severe performance degradation, indicating that directly updating the model on streaming unknown classes destabilizes known-class recognition. In contrast, training-free method like HM achieve better known-class stability but exhibit only moderate TA, reflecting limited discovery capability. While DAA incorporates a lightweight module that can be updated which still only based on non-semantic discovery and performance remains limited. Notably, a good TTD method must balance and improve both TA and CA; a high value in only one metric often masks poor overall discovery quality. Our proposed SEGA framework achieves the best overall performance, demonstrating that bridging visual clusters to interpretable semantic concepts via LLM-enriched text alignment improves both discovery accuracy and knowledge retention.

Comparison on Semantic Score of predictions. To quantitatively evaluate the semantic quality of our discovered labels, we compute the S^2 and Figure 4 reports averaged results over multiple runs on CIFAR100D and CUB200D. First, replacing the simple “a photo of [CLASS]” template with our attribute-rich LLM descriptions method consistently improves Semantic Scores across all metrics, confirming that fine-grained attribute alignment enhances semantic discriminability. Moreover, when LLM is provided with some attribute assistance, the score of the predicted results is higher, whether it is the template sentence method or our method. Notably, the maximum score among the top-5 candidates rises over Top-1, showing that the correct label often appears within the first few LLM suggestions but may not always be ranked first. However, as we enlarge the candidate pool to top-10, the maximum score only marginally increases, while the average score over top-10 drops significantly which reflects that LLMs sometimes generate implausible or hallucinated candidates when asked for too many options, and these low-quality entries pull down the average.

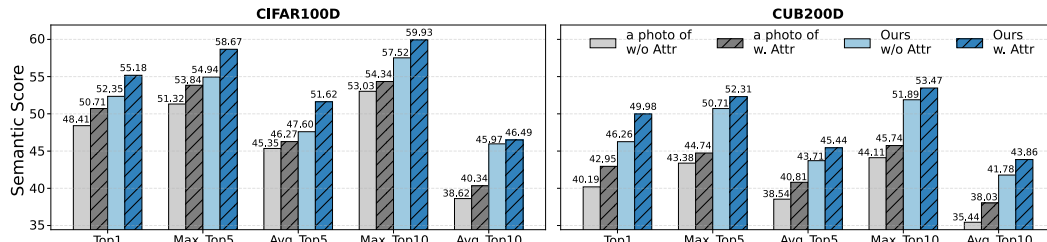


Figure 4: Comparisons of average Semantic Score (S^2) on predictions of unknown classes.

4.3 ANALYSIS ON SEGA

Examples of prediction procedure and performance. Fig. 5 showcases representative discovery cases from our framework. When the candidate set includes the ground-truth label in some cases, the correct label is successfully selected with a high reranking score. Even when exact prediction is not achieved, the top candidate remains semantically close. And from the visualization of the text embedding space, the predicted label consistently lies closer to the ground-truth embedding than to any known-class prototype, while still maintaining a clear separation from visually similar known categories. Consequently, even a non-exact prediction provides a useful anchor for test-time training which respects both semantic relatedness and discriminability guiding the adaptation for unknown classes without causing catastrophic confusion.

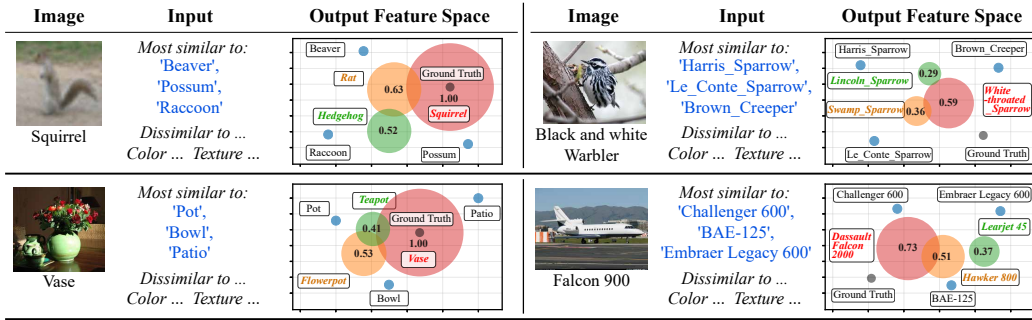


Figure 5: Some examples of our prediction procedure and textual feature distribution of output. (Size of bubble and value in bubble is S^2 of the output.)

Table 2: Comparisons of clustering metrics of post evaluation on four datasets.

Method	CIFAR100D				CUB200D				T-ImageNetD				AircraftD			
	HCA	ARI	NMI	VM	HCA	ARI	NMI	VM	HCA	ARI	NMI	VM	HCA	ARI	NMI	VM
HM (Lyu et al., 2025b)	63.69	48.35	73.08	73.08	50.46	27.81	70.25	70.25	56.64	40.35	72.70	72.70	33.30	21.33	46.46	46.46
DAA (Liu et al., 2025)	65.62	53.96	75.20	75.21	58.19	32.43	75.29	75.29	58.91	42.50	73.53	73.53	32.08	21.20	46.44	46.44
Ours	66.73	53.47	75.51	75.51	58.58	33.66	75.07	75.07	58.55	43.50	74.59	74.59	33.74	22.58	47.12	47.12

Comparison of compatibility metrics. TA and CA directly reflect online discovery quality and class consistency as the primary metrics, while we also provide some clustering metrics that traditional NCD methods use, including Hungarian Cluster Accuracy (HCA) (Meilă, 2003), Adjusted Rand Index (ARI) (Rand, 1971), Normalized Mutual Information (NMI) (McDaid et al., 2011) and V-Measure (Rosenberg & Hirschberg, 2007). The results in Table 2 demonstrate that our method also has better clustering performance on post evaluation.

Ablation study. As shown in Table 3, when none of these components is used which relies solely on a frozen CLIP model, the performance is poor, demonstrating that the pre-trained CLIP model struggles to handle the open-world TTD task due to its limited fine-grained semantic discriminability. Adding only Test-time Training (T) leads to modest improvements, indicating that zero-shot test-time adaptation alone is insufficient without a well-structured semantic space. In contrast, employing only Semantic Pre-construction (C) brings a substantial performance gain, highlighting the importance of constructing an attribute-anchored semantic space before deployment. Finally, the full model combining all three components achieves the best performance, confirming that semantic pre-construction provides a solid foundation, test-time training enables dynamic adaptation, and semantic reranking effectively refines the performance.

Reranking threshold. Table 4 shows the impact of the confidence threshold θ , which determines whether the top-ranked LLM-generated candidate is considered reliable enough to be used as a pseudo label for test-time training. Setting $\theta = 0$ which treats all predictions as reliable leads to the inclusion of many erroneous and noisy pseudo labels, which misguides subsequent test-time training and causes performance degradation. Conversely, setting $\theta = 1$ makes the method degenerate to using only relational cues similar/dissimilar known classes without explicit label and attribute assignment; while this conservative strategy avoids noise, it provides limited guidance and yields suboptimal results. The best performance is achieved at $\theta = 0.8$, where reliable predictions are adopted as pseudo labels to fully leverage semantic guidance, while unreliable ones are discarded and fall back to relation-only prompts, striking an effective balance between noise tolerance and semantic informativeness.

Table 3: Ablation study. C: Semantic Pre-construction, T: Test-time Training, S: Semantic Reranking, Sum: TA+CA.

C	T	S	Real-time Eval			Post Eval			
			TA	CA	Sum	TA	CA	Sum	KF
			20.57	31.16	51.73	23.54	29.26	52.80	8.81
		✓	21.46	40.69	62.15	24.00	33.20	57.20	11.89
	✓		25.07	34.59	59.65	25.27	31.53	56.80	6.54
	✓	✓	31.58	42.20	73.78	32.10	35.99	68.09	4.34
	✓	✓	34.74	47.77	82.51	38.66	39.21	77.87	3.28

Table 4: Comparisons of different reranking threshold. Sum: TA+CA.

θ	Real-time Eval			Post Eval			
	TA	CA	Sum	TA	CA	Sum	KF
0	30.88	46.59	77.46	31.40	37.57	68.96	5.39
0.6	31.43	47.46	78.89	33.98	37.35	71.33	4.32
0.7	32.35	45.50	77.85	35.27	35.93	71.19	4.06
0.8	34.74	47.77	82.51	38.66	39.21	77.87	3.28
0.9	33.93	46.60	79.63	38.25	38.15	76.40	3.61
1	33.15	45.90	79.05	37.63	37.76	75.39	3.82

Post visualization using T-SNE. In Fig. 6, we employ t-SNE (Van der Maaten & Hinton, 2008) to visualize the distribution of test samples. HM produces scattered and overlapping clusters, where known and unknown classes are poorly separated. DAA improves separation by amplifying feature discrepancies, yet the resulting clusters remain largely anonymous and categories are either mixed together or arbitrarily scattered, offering no interpretable organization. Our SEGA yields notably clearer and more structured clusters. Semantically similar classes like Bus and Truck, Apple and Orange, Rose and Sunflower are embedded close to each other, while still maintaining clear inter-class boundaries not only improving discrimination but also supporting understandable semantics.

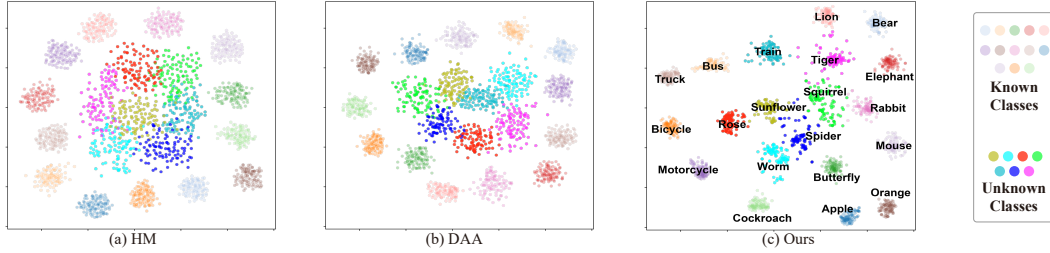


Figure 6: T-SNE visualization in feature space on CIFAR100D with known and unknown classes.

5 CONCLUSION

We presented SEGA, a framework that transforms Test-Time Discovery from anonymous clustering into interpretable semantic concept generation. By leveraging LLM-generated attribute descriptions to construct an attribute-anchored visual-semantic space, and introducing Relational Semantic Grounding with confidence-guided test-time stabilization, SEGA enables unknown classes to be named and understood rather than merely separated. Experiments on four benchmarks show consistent improvements over prior TTD methods in both discovery accuracy and semantic coherence. While LLM-generated candidates may occasionally be noisy or hallucinated, the semantic anchoring in a discriminative feature space provides a practical and robust foundation for open-world discovery. Future work will focus on more principled uncertainty estimation and adaptive filtering to further suppress spurious predictions, and to more dynamic, open-ended environments.

REFERENCES

- Eneko Agirre, Daniel Cer, Mona Diab, and Aitor Gonzalez-Agirre. SemEval-2012 task 6: A pilot on semantic textual similarity. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics and the Sixth International Workshop on Semantic Evaluation*, pp. 385–393, 2012.
- Malik Boudiaf, Romain Mueller, Ismail Ben Ayed, and Luca Bertinetto. Parameter-free online test-time adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.
- Dhanajit Brahma and Piyush Rai. A probabilistic framework for lifelong test-time adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.
- Chentao Cao, Zhun Zhong, Zhanke Zhou, Tongliang Liu, Yang Liu, Kun Zhang, and Bo Han. Noisy test-time adaptation in vision-language models. In *The Thirteenth International Conference on Learning Representations*, 2025.
- Fernando Julio Cendra, Bingchen Zhao, and Kai Han. Promptccd: Learning gaussian mixture prompt pool for continual category discovery. In *European Conference on Computer Vision*, 2024.
- Daniel Cer, Mona Diab, Eneko Agirre, Inigo Lopez-Gazpio, and Lucia Specia. SemEval-2017 task 1: Semantic textual similarity: Multilingual and cross-lingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation*, pp. 1–14, 2017.

- Liang Chen, Yong Zhang, Yibing Song, Ying Shan, and Lingqiao Liu. Improved test-time adaptation for domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.
- Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2009.
- Luqiong Ding, Jiayao Tan, Chenggong Ni, Fuyuan Hu, and Fan Lyu. Vd-lora: Adaptive reuse of low-rank directions for continual learning. In *European Conference on Computer Vision*, 2026.
- Mario Döbler, Robert A Marsden, and Bin Yang. Robust mean teacher for continual and gradual test-time adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.
- Ruoyi Du, Dongliang Chang, Kongming Liang, Timothy Hospedales, Yi-Zhe Song, and Zhanyu Ma. On-the-fly category discovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.
- Yingjun Du, Wenfang Sun, and Cees G Snoek. Ipo: Interpretable prompt optimization for vision-language models. *Advances in Neural Information Processing Systems*, 2024.
- Shenghao Fu, Qize Yang, Qijie Mo, Junkai Yan, Xihan Wei, Jingke Meng, Xiaohua Xie, and Wei-Shi Zheng. Llmdet: Learning strong open-vocabulary object detectors under the supervision of large language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025.
- Yossi Gandelsman, Yu Sun, Xinlei Chen, and Alexei Efros. Test-time training with masked autoencoders. *Advances in Neural Information Processing Systems*, 2022.
- Zhengqing Gao, Xu-Yao Zhang, and Cheng-Lin Liu. Unified entropy optimization for open-set test-time adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.
- Jiawei Gu, Ziyue Qiao, and Zechao Li. Activation shape matters: Ood detection with norm-entropy fusion. In *Proceedings of ACM International Conference on Multimedia*, 2025.
- 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*, 2019.
- Kai Han, Sylvestre-Alvise Rebuffi, Sébastien Ehrhardt, Andrea Vedaldi, and Andrew Zisserman. Automatically discovering and learning new visual categories with ranking statistics. In *International Conference on Learning Representations*, 2020.
- X. Y. Han, Vardan Papayan, and David L. Donoho. Neural collapse under MSE loss: Proximity to and dynamics on the central path. 2022.
- Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *International Conference on Machine Learning*, 2019.
- Salman Khan, Muzammal Naseer, Munawar Hayat, Syed Waqas Zamir, Fahad Shahbaz Khan, and Mubarak Shah. Transformers in vision: A survey. *ACM computing surveys (CSUR)*, 2022.
- Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschiot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in Neural Information Processing Systems*, 2020.
- James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 2017.

- Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.
- Yann Le and Xuan Yang. Tiny imagenet visual recognition challenge. *CS 231N*, 2015.
- Ting Lei, Shaofeng Yin, Qingchao Chen, Yuxin Peng, and Yang Liu. Open-vocabulary hoi detection with interaction-aware prompt and concept calibration. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025.
- Jian Liang, R. He, and Tien-Ping Tan. A comprehensive survey on test-time adaptation under distribution shifts. *International Journal of Computer Vision*, 2023.
- Huiwei Lin, Baoquan Zhang, Shanshan Feng, Xutao Li, and Yunming Ye. Pcr: Proxy-based contrastive replay for online class-incremental continual learning. *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.
- Jiashuo Liu, Zheyang Shen, Yue He, Xingxuan Zhang, Renzhe Xu, Han Yu, and Peng Cui. Towards out-of-distribution generalization: A survey. *arXiv preprint arXiv:2108.13624*, 2021a.
- Tianle Liu, Fan Lyu, Chenggong Ni, Zhang Zhang, Fuyuan Hu, and Liang Wang. Daa: Amplifying unknown discrepancy for test-time discovery. *Advances in Neural Information Processing Systems*, 2025.
- Yuejiang Liu, Parth Kothari, Bastien Van Delft, Baptiste Bellot-Gurlet, Taylor Mordan, and Alexandre Alahi. Ttt++: When does self-supervised test-time training fail or thrive? *Advances in Neural Information Processing Systems*, 2021b.
- Fan Lyu, Gongbo Cheng, Daofeng Liu, Linglan Zhao, Zhang Zhang, Fuyuan Hu, and Liang Wang. Mitigating catastrophic forgetting in online continual learning with dual-margin contrastive replay. *IEEE Transactions on Circuits and Systems for Video Technology*, 2025a.
- Fan Lyu, Tianle Liu, Zhang Zhang, Fuyuan Hu, and Liang Wang. Test-time discovery via hashing memory. *arXiv preprint arXiv:2503.10699*, 2025b.
- Fan Lyu, Gongbo Cheng, Daofeng Liu, Linglan Zhao, Zhang Zhang, Fuyuan Hu, and Liang Wang. Mitigating catastrophic forgetting in online continual learning with dual-margin contrastive replay. *IEEE Transactions on Circuits and Systems for Video Technology*, 36(4):4433–4449, 2026a. doi: 10.1109/TCSVT.2025.3639309.
- Fan Lyu, Linglan Zhao, Chengyan Liu, Yinying Mei, Zhang Zhang, Baoqing Yu, Tao Zhang, Fuyuan Hu, and Liang Wang. Gain: Global-atomic interaction graph for few-shot class-incremental learning. *IEEE Transactions on Circuits and Systems for Video Technology*, 36(7):10019–10035, 2026b. doi: 10.1109/TCSVT.2026.3673928.
- Subhansu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013.
- Anqi Mao, Mehryar Mohri, and Yutao Zhong. Cross-entropy loss functions: Theoretical analysis and applications. In *International Conference on Machine Learning*, 2023.
- Aaron F McDaid, Derek Greene, and Neil Hurley. Normalized mutual information to evaluate overlapping community finding algorithms. *arXiv preprint arXiv:1110.2515*, 2011.
- Yinying Mei, Linglan Zhao, Fan Lyu, Tianle Liu, Fenglei Xu, and Fuyuan Hu. Expansive geometry stabilization for exemplar-free continual learning with pre-trained models. In *IEEE International Conference on Multimedia and Expo (ICME)*, 2026.
- Marina Meilă. Comparing clusterings by the variation of information. In *Proceedings of Learning Theory and Kernel Machines*, 2003.
- Chenggong Ni, Fan Lyu, Jiayao Tan, Fuyuan Hu, Rui Yao, and Tao Zhou. Maintaining consistent inter-class topology in continual test-time adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

- Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Yaofo Chen, Shijian Zheng, Peilin Zhao, and Mingkui Tan. Efficient test-time model adaptation without forgetting. In *International Conference on Machine Learning*, 2022.
- Jaewoo Park, Jacky Chen Long Chai, Jaeho Yoon, and Andrew Beng Jin Teoh. Understanding the feature norm for out-of-distribution detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.
- Nan Pu, Zhun Zhong, and Nicu Sebe. Dynamic conceptional contrastive learning for generalized category discovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.
- Daiqing Qi, Handong Zhao, and Sheng Li. Better generative replay for continual federated learning. *International Conference on Learning Representations*, 2023.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 2021.
- William M Rand. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, 1971.
- Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the conference on Empirical Methods in Natural Language Processing and International Joint Conference on Natural Language Processing*, 2019.
- Andrew Rosenberg and Julia Hirschberg. V-measure: A conditional entropy-based external cluster evaluation measure. In *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, 2007.
- Matteo Sodano, Federico Magistri, Lucas Nunes, Jens Behley, and Cyrill Stachniss. Open-world semantic segmentation including class similarity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.
- Yu Sun, Xiaolong Wang, Zhuang Liu, John Miller, Alexei Efros, and Moritz Hardt. Test-time training with self-supervision for generalization under distribution shifts. In *International Conference on Machine Learning*, 2020.
- Jiayao Tan, Tianle Liu, Fuyuan Hu, Wei Feng, and Liang Wan. E-logic prompt: Unified energy-logic framework for continual visual question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2026a.
- Jiayao Tan, Fan Lyu, Tianle Liu, Fuyuan Hu, and Wei Feng. Towards dynamic modality alignment in multimodal continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026b.
- Jiayao Tan, Fan Lyu, Chenggong Ni, Fuyuan Hu, Wei Feng, and Rui Yao. Exposing mixture and annotating confusion for active universal test-time adaptation. In *International Conference on Learning Representations*, 2026c.
- Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 2008.
- 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*, 2022.
- Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. Technical report, California Institute of Technology, 2011.
- Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno A. Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. In *International Conference on Learning Representations*, 2021.

- 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 Computer Vision and Pattern Recognition Conference*, 2025.
- Weishuai Wang, Ting Lei, Qingchao Chen, and Yang Liu. Semantic-guided novel category discovery. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024.
- Zifeng Wang, Zizhao Zhang, Sayna Ebrahimi, Ruoxi Sun, Han Zhang, Chen-Yu Lee, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer Dy, et al. Dualprompt: Complementary prompting for rehearsal-free continual learning. In *European Conference on Computer Vision*, 2022a.
- Zifeng Wang, Zizhao Zhang, Chen-Yu Lee, Han Zhang, Ruoxi Sun, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer Dy, and Tomas Pfister. Learning to prompt for continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022b.
- Jianzong Wu, Xiangtai Li, Shilin Xu, Haobo Yuan, Henghui Ding, Yibo Yang, Xia Li, Jiangning Zhang, Yunhai Tong, Xudong Jiang, et al. Towards open vocabulary learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.
- Yanan Wu, Zhixiang Chi, Yang Wang, and Songhe Feng. Metagcd: Learning to continually learn in generalized category discovery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.
- Jingkang Yang, Kaiyang Zhou, Yixuan Li, and Ziwei Liu. Generalized out-of-distribution detection: A survey. *International Journal of Computer Vision*, pp. 5635–5662, 2024.
- Longhui Yuan, Binhui Xie, and Shuang Li. Robust test-time adaptation in dynamic scenarios. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.
- Hongyi Zhang, Moustapha Cissé, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. 2018.
- Jianyang Zhang, Qianli Luo, Guowu Yang, Wenjing Yang, Weide Liu, Guosheng Lin, and Fengmao Lv. Attribute-formed class-specific concept space: Endowing language bottleneck model with better interpretability and scalability. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025.
- Sheng Zhang, Salman Khan, Zhiqiang Shen, Muzammal Naseer, Guangyi Chen, and Fahad Shahbaz Khan. Promptcal: Contrastive affinity learning via auxiliary prompts for generalized novel category discovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.
- Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. BERTScore: Evaluating text generation with BERT. In *International Conference on Learning Representations*, 2020.
- Haiyang Zheng, Nan Pu, Wenjing Li, Nicu Sebe, and Zhun Zhong. Prototypical hash encoding for on-the-fly fine-grained category discovery. *Advances in Neural Information Processing Systems*, 2024a.
- 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*, 2024b.
- Zhun Zhong, Enrico Fini, Subhankar Roy, Zhiming Luo, Elisa Ricci, and Nicu Sebe. Neighborhood contrastive learning for novel class discovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.
- Chaoyang Zhu and Long Chen. A survey on open-vocabulary detection and segmentation: Past, present, and future. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.
- Lei Zhu, Yanyu Xu, Yong Liu, Rick Siow Mong Goh, and Xinxing Xu. Ad 2 mix: Adversarial and adaptive mixup for unsupervised domain adaptation. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2025.