Learning Like Humans: Analogical Concept Learning for Generalized Category Discovery

Supplementary Material

7. More Implementation details

7.1. Dataset Details and Class Splits

In this work, we evaluate our proposed framework on six datasets, encompassing both general and fine-grained image classification tasks. Table 5 summarizes the distribution of known and unknown classes for each dataset, which are used to assess the generalized category discovery (GCD) performance. Below, we provide detailed information about each dataset and the corresponding class split protocol.

CIFAR100. CIFAR100 [32] is a widely used image classification dataset containing 60,000 images across 100 categories. For GCD evaluation, we use 80 classes as known categories for training and the remaining 20 classes as unknown categories for testing.

ImageNet100. ImageNet100 [33] is a subset of ImageNet, containing 100 categories. Following standard protocols, we randomly select 50 classes as known categories and the other 50 as unknown categories.

CUB-200-2011. CUB-200-2011 [34] is a fine-grained bird classification dataset with 200 categories. We follow the standard fine-grained GCD protocol [38] and split the dataset into 100 known classes and 100 unknown classes.

Stanford Cars. Stanford Cars [35] is a fine-grained car classification dataset containing 196 classes. We use 98 classes as known categories and 98 as unknown categories, consistent with previous fine-grained GCD works.

FGVC Aircraft. FGVC Aircraft [36] is another fine-grained dataset with 100 aircraft categories. We use 50 classes as known categories and 50 classes as unknown, following the SSB protocol.

Herbarium19. Herbarium19 [37] is a large-scale fine-grained dataset with 683 plant species. Due to its high intra-class variance and fine-grained nature, we split the dataset into 341 known classes and 342 unknown classes to test the robustness of GCD methods on challenging datasets.

Class Split Protocol. For the general datasets (CIFAR100 and ImageNet100), we adopt a random class split protocol using a fixed seed for reproducibility. For fine-grained datasets, including CUB-200-2011, Stanford Cars, and FGVC Aircraft, we follow the SSB split protocol [38], which ensures a balanced distribution of known and unknown classes while maintaining intra-class variance. Herbarium19 is split into nearly equal known and unknown classes to test the model's ability to handle fine-grained datasets with high intra-class variation.

Dataset	Known Classes	Unknown Classes
CIFAR100 [32]	80	20
ImageNet100 [33]	50	50
CUB-200-2011 [34]	100	100
Stanford Cars [35]	98	98
FGVC-Aircraft [36]	50	50
Herbarium19 [37]	341	342

Table 5. Distribution of known and unknown classes.

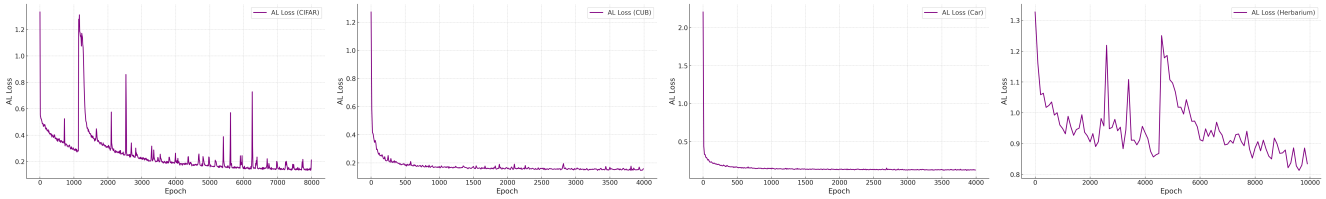
8. Training Visualization and Analysis

8.1. ATCG Training

The Analogical Textual Concept Generator (ATCG) training process is crucial for generating meaningful textual embeddings for novel categories. The ATCG training loss for four datasets—CIFAR100, CUB-200, Stanford Cars, and Herbarium19—is shown in Fig. 5. The loss curves provide insights into the effectiveness of analogical learning across datasets with varying semantic structures.

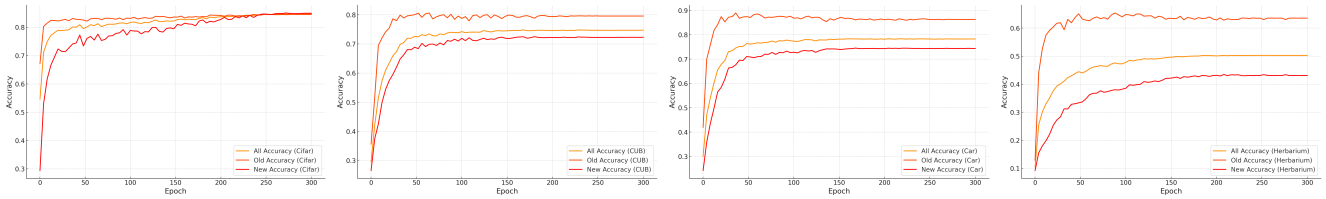
For datasets like CUB-200 and Stanford Cars, the ATCG loss decreases smoothly and stabilizes quickly. This is attributed to the compositional nature of their class names, which facilitates analogical reasoning. For example, in CUB-200, species names such as *Yellow Warbler* and *Black-throated Blue Warbler* can be combined linearly to form plausible new class names like *Blue Warbler*. Similarly, in Stanford Cars, structured names such as *BMW M3 Coupe* and *BMW M5 Sedan* offer clear semantic relationships, enabling effective analogical learning. In contrast, the loss curves for CIFAR100 and Herbarium19 exhibit greater fluctuations and slower convergence. This is due to the lack of compositional semantics in their class names. CIFAR100 features abstract class names such as *airplane* and *frog*, which do not lend themselves to meaningful combinations. Similarly, Herbarium19 contains botanical names like *Acer rubrum* and *Quercus alba*, which are domain-specific and lack intuitive semantic relationships. These challenges force the model to rely more heavily on visual features, making the training process less efficient.

The AL loss in the ATCG training phase reflects the pseudo-GCD process, where the model iteratively refines textual embeddings to separate novel categories. While datasets like CUB-200 and Stanford Cars benefit from strong semantic cues, the robustness of our approach is demonstrated in CIFAR100 and Herbarium19, where ana-



(a) CIFAR100. (b) CUB. (c) Stanford Cars. (d) Herbarium19.

Figure 5. ATCG Training Loss Across Different Datasets.



(a) CIFAR100. (b) CUB. (c) Stanford Cars. (d) Herbarium19.

Figure 6. GCD Training Accuracy Across Different Datasets.

logical learning operates effectively even in the absence of clear semantic relationships.

8.2. GCD Training

The Generalized Category Discovery (GCD) training stage focuses on the model's ability to achieve balanced performance across old and novel classes. The accuracy trends for GCD training on CIFAR100, CUB-200, Stanford Cars, and Herbarium19 are presented in Fig. 6. These trends highlight the model's adaptability to different datasets and the role of analogical reasoning in novel class discovery.

For CUB-200 and Stanford Cars, the model demonstrates rapid convergence, with novel class accuracy improving significantly within the first 50 epochs. The compositional nature of the class names allows the model to leverage semantic analogies effectively, resulting in strong performance across both old and novel classes. For instance, semantic relationships between bird species names in CUB-200 or car model names in Stanford Cars provide the model with a foundation for discovering novel categories.

In CIFAR100 and Herbarium19, the accuracy trends reveal the challenges posed by unstructured or domain-specific class names. The model achieves steady improvements in old and novel class accuracy, but the convergence is slower compared to CUB-200 and Stanford Cars. This is because the abstract or specialized class names in CIFAR100 and Herbarium19 offer limited semantic information for analogical reasoning. Nevertheless, the model demonstrates robustness by relying on the fusion of visual and textual features to achieve balanced performance.

Across all datasets, the GCD training process reflects the complementary strengths of analogical textual embeddings and visual features. While compositional class names

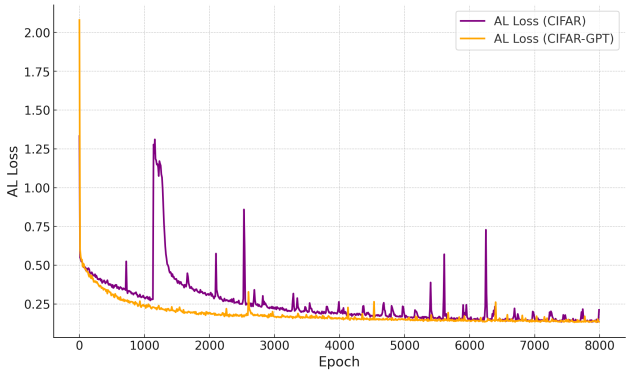


Figure 7. ATCG Training Loss with and without additional descriptions.

enhance the efficiency of novel class discovery in datasets like CUB-200 and Stanford Cars, the ability to adapt to less structured datasets like CIFAR100 and Herbarium19 highlights the versatility of our framework. These analyses underscore the importance of semantic structures in facilitating analogical learning and reveal the robustness of our approach in diverse scenarios.

9. More Ablation Study

9.1. Cross-Backbone Generalization

To assess generality, we plug AL-GCD into SelEx and vary the *visual* backbone among DINOv1, DINOv2, and CLIP, while **all settings use CLIP's text encoder**. Training splits and schedules are kept identical across settings for a fair comparison. To keep the feature dimensions consistent, when using DINOv1 or DINOv2 as the visual backbone, we

Method	DINOv1			DINOv2			CLIP			AVG		
	ALL	Old	New	ALL	Old	New	ALL	Old	New	ALL	Old	New
GCD (CVPR 22)	51.3	56.6	48.7	71.9	71.2	72.3	51.1	56.2	48.6	58.1	61.3	56.5
SimGCD (ICCV 23)	60.3	65.6	57.7	74.9	78.5	73.1	69.6	75.8	66.5	68.3	73.3	65.8
RLCD (ICML 25)	70.0	**79.1**	65.4	78.7	79.5	78.3	-	-	-	-	-	-
SelEx (ECCV 24)	73.6	75.3	72.8	87.4	85.1	88.5	74.2	69.5	76.5	78.4	76.6	79.3
+ AL-GCD	**76.0**	75.4	**76.3**	**90.1**	**86.3**	**92.0**	**85.6**	**77.3**	**89.7**	**83.9**	**79.7**	**86.0**
Improvement	***+2.4***	***+0.1***	***+3.5***	***+2.7***	***+1.2***	***+3.5***	***+11.4***	***+7.8***	***+13.2***	***+5.5***	***+3.0***	***+6.7***

Table 6. Comparison across different backbones (DINOv1, DINOv2, CLIP) and average on CUB200.

Additional Descriptions	CIFAR100		
	ALL	Old	Novel
✗	84.7	84.5	84.9
✓	**85.0**	**84.6**	**85.8**

Table 7. Comparison of CIFAR100 results with and without additional descriptions. Best values are bolded.

append a linear projection layer after the CLIP text encoder. For context, we report representative baselines alongside our results, including GCD [1], SimGCD [5], RLCD [41], and SelEx [40]. Results are summarized in Table 6.

ATCG improves performance under all three visual backbones, though the gain profile differs across models. With DINOv1, we see a modest boost of **+2.4** in overall accuracy, driven mainly by a **+3.5** increase on novel classes while known classes change little. Switching to DINOv2 yields slightly stronger gains: **+2.7** overall, with **+1.2** on known and **+3.5** on novel classes. The benefit becomes much more pronounced when the visual backbone is CLIP, where ATCG lifts performance by **+11.4** overall, including **+7.8** on known and **+13.2** on novel classes. Averaged across the three backbones, ATCG delivers **+5.5** overall, **+3.0** on known classes, and **+6.7** on novel classes.

These results suggest that ATCG complements purely visual representations by injecting analogical textual concepts, and the strongest synergy appears when the visual backbone is already language-aligned, as in CLIP. Meanwhile, under DINO, known-class accuracy remains largely stable while novel-class accuracy improves steadily, indicating that the added textual guidance mainly benefits categories that are harder to separate using visual features alone.

9.2. Impact of Additional Descriptions

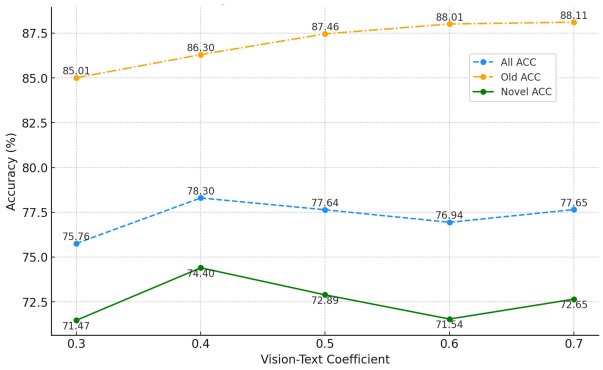
To explore the impact of enriched textual descriptions, we introduced GPT-4-generated class descriptions for CIFAR100, such as *A photo of a butterfly: A delicate insect with large, colorful wings and a slender body, often seen fluttering in gardens*. These descriptions were incorporated into the ATCG training phase to enhance the semantic understanding of each class.

The inclusion of these descriptions led to notable improvements in training stability, as illustrated in Fig. 7. The ATCG loss curve for the model trained with additional descriptions converges more smoothly and exhibits reduced fluctuations compared to the baseline without them. This demonstrates that the enriched textual information allows the model to better align visual and semantic features, facilitating more effective analogical learning. The detailed descriptions provide semantic context that bridges the gap between visually distinct but semantically related classes.

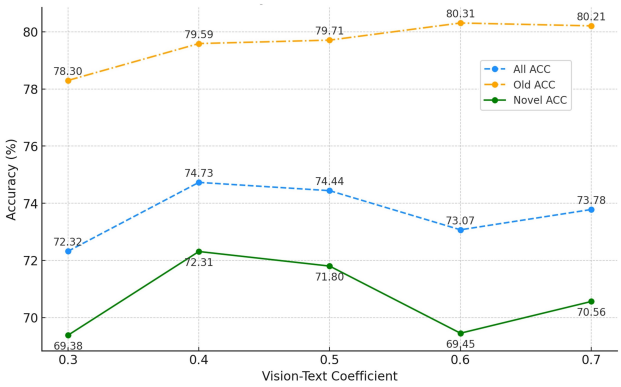
As shown in Table 7, the addition of GPT-4 descriptions resulted in consistent accuracy improvements across all metrics. Specifically, the overall accuracy increased from 84.7% to 85.0%, while the novel class accuracy rose significantly from 84.9% to 85.8%. This suggests that the model benefits greatly from the enriched textual information when discovering novel categories, as the detailed descriptions provide a stronger basis for analogical reasoning. Moreover, the old class accuracy also improved, indicating that the additional descriptions enhance the model's capacity to generalize across both seen and unseen categories.

The effectiveness of these descriptions is particularly valuable for datasets like CIFAR100, where the original class names (e.g., *apple*, *bicycle*, and *chair*) lack compositional semantics. By introducing detailed and descriptive features, the model is able to infer richer semantic relationships, leading to more robust representations. This finding highlights the importance of textual enrichment in facilitating analogical learning, especially in scenarios where the dataset's semantic structure is less intuitive.

In summary, the results demonstrate that incorporating GPT-4-generated descriptions significantly enhances the model's analogical learning capability. This approach not only improves training efficiency but also boosts overall and novel class performance, showcasing the potential of semantic enrichment in addressing challenges posed by abstract or non-compositional datasets.



(a) Ablation study of the α on Stanford Cars.



(b) Ablation study of the α on CUB-200.

Figure 8. Ablation study of the α

9.3. The Influence of the Vision-Text Coefficient.

The vision-text coefficient α controls the relative contributions of visual and textual information in generating fused embeddings. This hyperparameter is critical in balancing the model's performance across old and novel classes. We conduct an ablation study on two datasets, Stanford Cars and CUB-200, to investigate its impact.

For **Stanford Cars** (Fig. 8a), increasing α from 0.3 to 0.7 leads to a consistent improvement in old class accuracy, peaking at 88.11% when $\alpha = 0.7$. However, novel class accuracy decreases, with the highest value of 74.40% achieved at $\alpha = 0.4$. This trend suggests that emphasizing visual features benefits old class performance, as visual information is more aligned with the rich and well-learned representations of the old categories. However, for novel classes, textual embeddings generated through analogical reasoning play a more significant role. The highest all-class accuracy of 78.30% is observed at $\alpha = 0.4$, demonstrating the effectiveness of balancing visual and textual contributions.

For **CUB-200** (Fig. 8b), a similar pattern emerges. Old class accuracy reaches its maximum at 80.31% when $\alpha = 0.6$, while novel class accuracy peaks at 72.31% when $\alpha = 0.4$. All-class accuracy follows a similar trend, achieving the best performance of 74.73% at $\alpha = 0.4$. This dataset, characterized by fine-grained visual distinctions, highlights the importance of textual embeddings in separating novel classes. Textual concepts enhance the semantic understanding of novel categories, providing an edge in discovering previously unseen classes.

These results reveal a trade-off between old and novel class performance as α varies. Smaller values of α prioritize textual embeddings, which are essential for novel class separability, while larger α values give more weight to visual embeddings, benefiting old class accuracy. A balanced coefficient, such as $\alpha = 0.4$, achieves robust performance across all metrics by effectively leveraging the complementary strengths of visual and textual modalities.

In summary, the vision-text coefficient α plays a pivotal role in determining the effectiveness of our analogical learning framework. The ablation study demonstrates that an optimal balance between visual and textual contributions is key to achieving superior performance in generalized category discovery tasks, particularly for fine-grained datasets.

Table 8. Efficiency of ATCG training on different datasets measured with a single NVIDIA RTX 3090 GPU.

Dataset	Memory (GB)	Training Time / Epoch (s)
CUB	3.7	0.22
Herbarium19	7.4	0.87

10. Efficiency of ATCG Training

To demonstrate the practicality and scalability of our proposed ATCG framework, we report its computational efficiency in terms of memory usage and training time per epoch on two representative datasets: CUB and Herbarium19. All experiments are conducted on a single NVIDIA RTX 3090 GPU using a batch size of 256. As shown in Table 8, the training cost of ATCG is very low. On the CUB dataset, each epoch takes only **0.22 seconds**, and the full training process of 4000 epochs completes in approximately **14 minutes**. On the larger and more complex Herbarium19 dataset, each epoch takes only **0.87 seconds**, which is still below 1 second per epoch. These results indicate that ATCG is not only effective in discovering categories but also computationally efficient. The lightweight training process makes it suitable for deployment in resource-constrained environments and facilitates fast model updates in dynamic or continual learning settings.

11. Limitations and Future Work

While AL-GCD achieves consistent gains across six benchmarks, several limitations remain. *First*, our integration varies the *visual* backbone (DINOv1/v2, CLIP) but keeps

the *text* side fixed to CLIP; performance may thus depend on CLIP's vocabulary and linguistic priors, especially in domains with specialist nomenclature. Future work will explore lightweight text adapters and domain/multilingual descriptors. *Second*, our study focuses on image classification and inherits K-estimation from clustering baselines; extending analogical concept generation to detection/segmentation, video, continual/open-world discovery, and jointly estimating K in the analogical space are promising directions.