A Advanced Data Mining Approaches

This section details the advanced data mining strategies we developed to improve detection of rare and safety-critical objects in autonomous driving scenarios.

A.1 Random-Target and Random-Target+ Approaches

We developed several specialized variants of our data mining framework to enhance coverage of rare and safety-critical objects:

- **Random-Target (Bicycle, Motorcycle):** This is the main method of this paper, and it focuses on two vulnerable road user categories that are critical for safety. By targeting both bicycles and motorcycles, it achieves enhanced performance in detecting these challenging categories while maintaining the same overall mAP improvement.
- **Random-Target (Bicycle, Construction Vehicle):** This setting targets bicycle and construction vehicle, which were the two categories with the lowest mAP in the Random 10% experiment (Table 1).
- **Random-Target+ (Bicycle, Construction Vehicle):** This approach incorporates potential target objects into the caption generation pipeline regardless of their outlier status, indicated by the red arrow in Figure 1. This ensures comprehensive extraction of near target objects such as construction vehicles and trucks. The approach is particularly effective for classes with high visual diversity, where standard outlier detection might miss important instances. The 'truck' class in YOLOX corresponds to all of 'construction_vehicle', 'truck', and 'trailer' in nuScenes, and by including all of these near target objects in the caption generation process, this approach aims to capture most of the construction vehicles in nuscenes images.

A.2 Performance Analysis of Advanced Approaches

Performance results for our various Random-Target approaches are shown in Table 1. Notable observations include:

- **Random-Target (Bicycle, Construction Vehicle):** Achieved exceptional performance in bicycle detection (11.9%), demonstrating the benefit of selective filtering for bicycle objects. This approach shows

Table 1: Extended performance comparison on the nuScenes test set in terms of Average Precision (AP) for our advanced mining strategies. The Random-Target (B,CV) refers to bicycle and construction vehicle targeting, while Random-Target (B,M) refers to bicycle and motorcycle targeting. Asterisks (*) denotes cases where the accuracy improved by more than 0.5 AP compared to the Random 20% setting.

Data	mAP	Car	Truck	Bus	Tra	CV	Ped	Mot	Bic	Traf	Barrier
Random 20%	43.3	81.1	45.2	59.7	28.2	8.8	73.2	31.1	7.5	46.4	52.4
Random-Rare	44.0*	81.4	46.1*	59.9	30.5*	9.4*	73.0	31.2*	7.8	48.5*	52.4
Random-Target (B,CV)	44.7*	81.3	45.4*	58.1	32.3*	8.5	73.1	32.7*	11.9*	49.4*	54.2*
Random-Target (B,M)	44.7*	81.2	45.5*	58.1	31.2*	9.3*	72.9	34.6*	11.2*	48.7*	54.1*
Random-Target+	44.6*	81.5	46.7*	57.5	34.0*	9.3*	73.0	33.3*	9.1*	48.7*	52.4

strong trailer detection (32.3%) but is less effective for construction vehicle detection compared to the Random-Rare approach.

- **Random-Target (Bicycle, Motorcycle):** By targeting two vulnerable road user categories, this approach achieved significant improvements in motorcycle detection (34.6% vs 31.1% for Random 20%) while maintaining strong bicycle detection performance (11.2% vs 7.5% for Random 20%). It also showed good performance in construction vehicle detection (9.3%), suggesting that improved motorcycle detection indirectly benefits the detection of other rare categories.

- **Random-Target+ (Bicycle, Construction Vehicle):** Achieved strong performance in truck detection (46.7%) and trailer detection (34.0%), outperforming the Random-Target approaches in these categories. This validates the effectiveness of including potential target objects regardless of outlier status for these visually diverse categories.

All variants of our approach demonstrate significant improvements over the Random 20% baseline across most categories, with about the same 1.4 (or 1.3) mAP improvement (44.7% vs 43.3%). This confirms the effectiveness and flexibility of our concept-based data mining framework, which can be tailored to target different safety-critical object categories depending on specific application needs.

A.2.1 Construction Vehicle Detection: Analysis and Challenges

Construction vehicles present an interesting case study in our experiments. While the Random-Rare approach achieved slightly better performance (9.4%) compared to the Random-Target (Bicycle, Construction Vehicle) approach (8.5%), this can be explained by several factors:

- **Inherent Rarity:** Construction vehicles are inherently rare in the nuScenes dataset, making them challenging targets for any data mining approach.
- **High Visual Diversity:** The construction vehicle category encompasses a wide variety of vehicle types (excavators, bulldozers, cement mixers, etc.) with significantly different visual appearances, making it difficult for VLMs to consistently identify all variants.
- **Visual Similarity to Trucks:** Many construction vehicles share visual characteristics with trucks, leading to potential confusion in VLM-based identification. Our analysis of concept similarity scores revealed frequent overlap between truck and construction vehicle categories.

Interestingly, the Random-Target (Bicycle, Motorcycle) approach achieved 9.3% AP for construction vehicles despite not explicitly targeting this category. This suggests that, due to the high intra-class diversity of construction vehicles, a more generic rare-object targeting strategy can outperform explicitly targeting this category when the VLM's recognition capability is not yet reliable. Furthermore, since the architecture allows for flexible replacement of the VLM component (validated with BLIP, Qwen-based models, and OpenAI-based VLMs), the AP for construction vehicles is expected to improve as VLM accuracy improves. This highlights the robustness and generalizability of our method.

B Implementation Details of the Concept-based Data Mining Framework

This section provides information about the implementation of our concept-based data mining framework, focusing on the algorithmic innovations that enable effective rare object detection.

B.1 Multiple Outlier Detection Approaches

We tried a diverse set of outlier detection algorithms to enhance the robustness of rare object identification:

- **Isolation Forest:** We utilize Isolation Forest with a contamination parameter of 0.20, which efficiently identifies non-linear anomalies by isolating observations through recursive partitioning.

- **Train-test split for outlier detection** In our early experiments, we explored training the Isolation Forest and t-SNE model on 10% of the initial nuScenes dataset and then using it to detect outliers from the remaining 90%. This approach, visualized in Figure 2, provided valuable insights into the distribution of rare objects and helped establish our current methodology.

- **Local Outlier Factor (LOF):** This density-based method identifies outliers by comparing the local density of a sample with the densities of its neighbors, effectively detecting outliers in regions of varying density.

- **t-SNE Distance-based Detection:** After dimensionality reduction using t-SNE, we implement a distance-based outlier detection approach that identifies points with anomalously large distances to their nearest neighbors.

- **Ensemble Combination:** combines the results from multiple algorithms using either union or intersection operations, significantly improving the precision of outlier detection. While we initially experimented with IOF+LOF combinations, our comparative analysis revealed that the t-SNE+IOF combination yielded the most effective results for our specific task, particularly in identifying visually distinct rare objects relevant to autonomous driving. This combination was therefore selected for our final implementation.

B.2 Class-Aware Outlier Detection

Class imbalance presents a significant challenge in autonomous driving datasets. To address this issue, the proposed method employs class-aware outlier detection, which processes each class independently. This targeted approach accounts for class-specific characteristics, proving particularly effective when dealing with highly skewed distributions typical in autonomous driving data. The contamination parameters for each class are fine-tuned according to their frequency and importance in the dataset, providing greater flexibility in controlling outlier detection sensitivity across different object categories.

B.3 Multi-modal Feature Extraction

Central to the framework's capacity for identifying semantically meaningful rare objects is the integration of visual and linguistic features through several complementary techniques:

- **CLIP Embeddings:** The system leverages CLIP (ViT-B/32) to extract image embeddings, yielding robust visual representations with rich semantic understanding derived from large-scale vision-language pretraining.
- **Caption Generation:** A flexible captioning mechanism alternates between Qwen2-VL and BLIP models depending on specific requirements and availability, maximizing the quality of generated descriptions.
- **Feature Integration:** By combining visual and textual information, the approach creates a comprehensive semantic representation of each object, capturing nuances that might be missed by single-modality systems.
- **Weighted Similarity Calculation:** An adjustable weighting parameter balances textual and visual similarity scores, allowing fine-tuning based on detection requirements for different object categories.

C Exploration and Development Process

This section details key insights from our exploration process and challenges we overcame in developing the concept-based data mining framework.

C.1 Key Insights and Challenges

Several important insights emerged during our development process:
Class Imbalance Insights: We discovered that raw instance counts do not necessarily predict detection performance. Despite construction vehicles having more instances than motorcycles in the nuScenes dataset, their detection performance was significantly lower. This is due to the higher visual diversity within the construction vehicle category compared to motorcycles.
Visual-Semantic Alignment: Ensuring consistency between visual features and semantic concepts was challenging. We developed a specialized mapping function to bridge YOLO's detection categories with nuScenes classes to improve concept matching accuracy.

C.2 Ablation Studies and Performance Analysis

Table 2 presents a summary of our key experimental configurations and their corresponding performance metrics.

Table 2: Ablation study of different data mining configurations. The training set for the final method (our proposed approach) was constructed by randomly selecting samples from the entire pool of discoveries made using IOF t-SNE combined detection unified with near target (possible target classes), along with all samples for which the top concept was either "construction vehicle" or "bicycle."

Configuration	**mAP**	**Notes**
Random-only baseline (20%)	43.4	Baseline performance
IOF+LOF with random sampling	43.7	Modest improvement
BLIP for CV and bicycle detection	44.0	Better concept identification
Final method	44.7	Best performance

Best Performing Configuration: Our most successful approach (achieving 44.7 mAP) first identified potential classes (e.g., truck, bicycle) using YOLOv8, then applied the t-SNE+IOF combination for outlier detection, which proved more effective than other combinations like IOF+LOF. Specifically, the Isolation Forest was applied to detect outliers within target categories, while t-SNE visualization enhanced the identification of visually distinct rare objects. Qwen2-VL was used to generate captions for these detected objects, and concept similarity matching identified the most relevant rare objects. This was complemented with construction vehicle neighborhood mining for additional focused improvement.

D Model and Implementation Details

The following table provides technical specifications of the models and key parameters used in our final implementation:

E Training Curve Comparison

As shown in Figure 3, our data mining strategy enables faster convergence during training. The loss decreases more rapidly both in terms of epochs and actual elapsed training time, indicating improved sample efficiency and potential for reduced computational cost. This is an additional practical advantage, especially for large-scale or resource-constrained training scenarios.

Table 3: Model specifications and key parameters

Component	**Specification**
Object Detection	YOLOv8-l (80 classes, 640×640 input)
VLM for Captioning	Qwen2-VL-2B-Instruct (float16 precision)
Feature Extractor	CLIP-ViT-B/32 (LAION-2B trained)
Isolation Forest	contamination=0.20, n_estimators=100
LOF	n_neighbors=20, contamination=0.20
t-SNE	perplexity=30, n_components=2
Cosine Similarity	text weight=0.5, image weight=0.5
Hardware	NVIDIA A100 40GB GPU
Processing Time	3.5 hours for full pipeline on nuScenes train split

F Training Set Examples

This method constructs a training set that includes construction vehicles, bicycles, and other rare objects. A portion of the training set is shown in Figure 4.

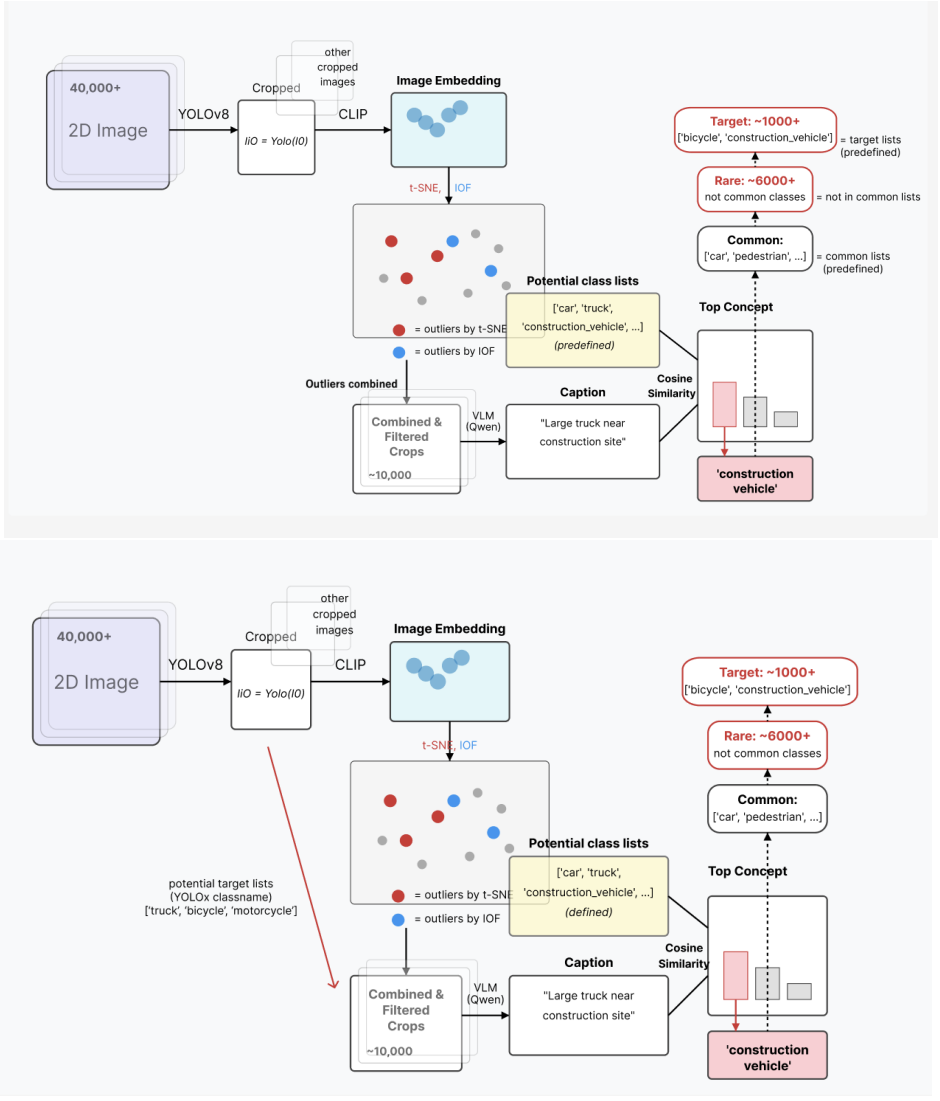

Figure 1: Comparison of Random-Target (Top) and Random-Target+ (Bottom) approaches. Random-Target+ (Bottom) incorporates potential target objects into the caption generation pipeline regardless of their outlier status, indicated by the red arrow, to ensure comprehensive extraction of target objects such as trucks. Random-Target (Top) employs a more selective approach to target object inclusion.

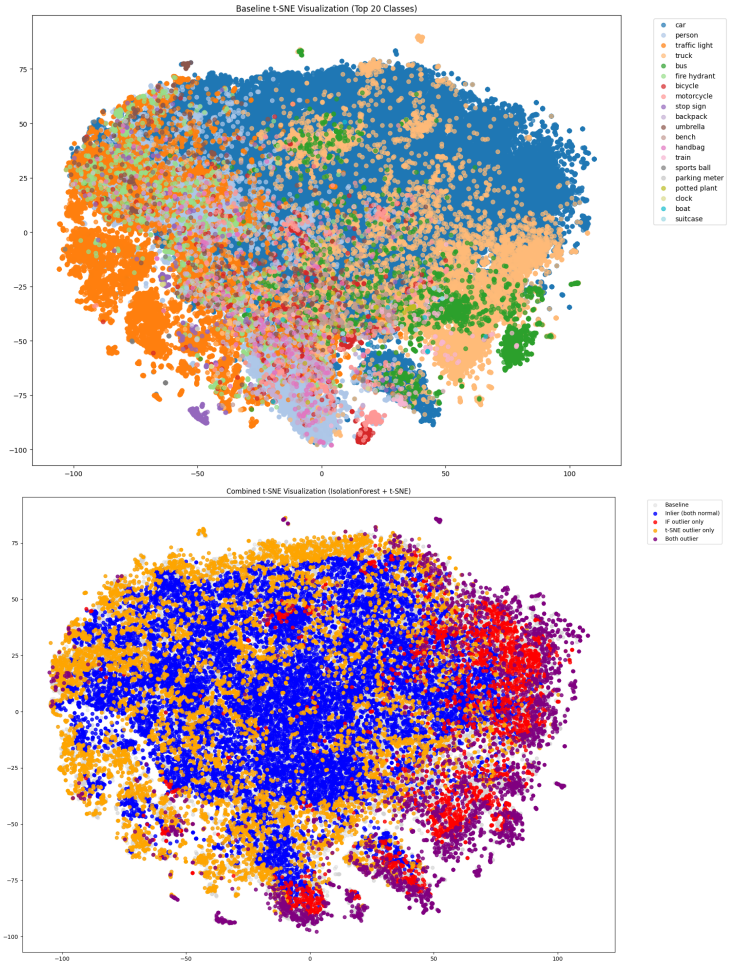

Figure 2: t-SNE visualization of object embeddings from our early experiment where we trained the Isolation Forest model on 10% of the initial nuScenes dataset and attempted to detect outliers from the remaining 90%. (*Top*) Embeddings colored by object category. (*Bottom*) Isolation Forest outlier detection overlaid on the t-SNE map: detected outliers are highlighted, while inliers retain their category-based colors. This approach helped establish the foundation for our current outlier detection methodology.

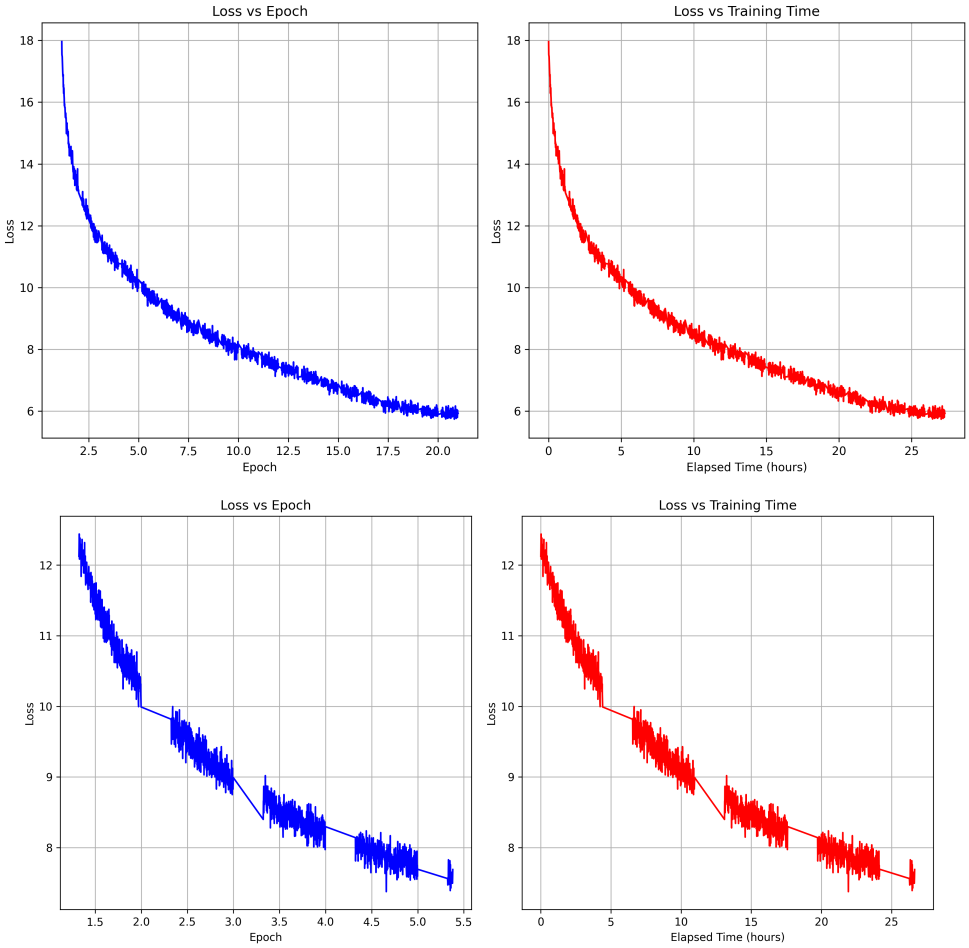

Figure 3: Training loss curves comparision for the proposed data mining strategy (Top) and baseline using whole training dataset (Bottom). (**Left**) Loss vs Epoch. (**Right**) Loss vs Elapsed Training Time (hours). The plots show that not only does the final loss reach a lower value, but the convergence is also faster compared to the baseline. The bottom plot shows training results (for about first 25 hours) using the entire training set without data mining. When aligned by elapsed training time, the progress in terms of epochs is significantly slower, and the loss reduction is more gradual. This highlights the improved training efficiency enabled by our approach.

Figure 4: Training set examples including construction vehicles, bicycles, and other rare objects.