2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

FALCO: Foundation Model Guided Active Learning for Cost-Effective Off-Road Freespace Detection

Shuai Wang, Chenxin Li, Yintong Chen, Yaobo Jia, Hongze Li, Chen Min, Jilin Mei, Huijing Zhao

***Abstract*— Freespace detection in unstructured off-road environments is critical for safe autonomous navigation but remains highly challenging due to ambiguous boundaries, diverse terrains, and long-tail safety-critical cases. Constructing large annotated datasets in such environments is prohibitively costly, which makes active learning essential to maximize model robustness under limited annotation budgets. However, conventional uncertainty or diversity-based strategies are unreliable in these complex settings, often failing to capture rare yet important scenarios. To address this, we propose FALCO, a foundation model guided active learning framework for cost-effective off-road freespace detection. FALCO integrates three complementary criteria: prediction deviation from a vision foundation model, model uncertainty, and semantic evaluation from a vision-language model to form a reliable sample criticality score. In addition, we introduce a semantic grid based sampling strategy that balances coverage across scene conditions while prioritizing challenging cases. Extensive experiments show that FALCO substantially improves robustness on rare and difficult scenarios, achieving significant gains in low-percentile IoU compared to state-of-the-art baselines, while maintaining competitive overall performance.**

I. INTRODUCTION

Freespace detection is a fundamental component of autonomous driving, playing an important role in trajectory planning. Although substantial progress has been achieved in structured environments such as highways and urban roads, extending autonomous driving to unstructured scenarios (e.g., farmlands, forests, deserts) presents considerable challenges. These challenges arise from the diversity of environment types and elements, and the ambiguity of road-related semantics such as road conditions and scene structures [1]. Deep learning models have shown strong potential for unstructured freespace detection, but their performance heavily depends on large-scale, finely annotated datasets. However, obtaining such datasets in unstructured environments is prohibitively costly, as road boundaries are often ambiguous and geometrically complex, making manual labeling particularly difficult. At the same time, the data distribution shows a pronounced long-tail effect, where rare yet safety-critical cases are difficult to capture and annotate, raising a key question: *how can the most informative samples be selected under a limited labeling budget to achieve robust and generalizable performance?*

Active learning is a common paradigm for addressing this question. As shown in Fig. 1, the goal of active learning is to find effective ways to select the most informative samples for manual labeling from a pool of unlabeled data points in order to improve predictive model performance [2]. Much of the existing research in this field centers on sample selection strategies based on uncertainty and diversity [3]. However, both face challenges when applied to unstructured autonomous driving data. Due to the high complexity of the environment and the lack of effective prior knowledge, autonomous driving systems in unstructured scenarios face increased levels of uncertainty and unpredictability [1], which makes it difficult for existing methods to accurately estimate data uncertainty. Diversity-based methods typically rely on low-level features or embedding-space distributions that do not always align with semantic distinctions. This misalignment can introduce semantic bias—for instance, samples with large variations in texture or illumination may be selected, while rare yet safety-critical semantic categories, such as narrow trails partially blocked by fallen trees, remain underrepresented. These challenges call for more robust sample selection approaches that can capture meaningful diversity and prioritize critical scenarios.

Foundation models (FMs) offer a promising direction: vision foundation models (VFM) such as SAM [4] provide generalized perception of obstacles, terrain boundaries, and regions, while vision-language foundation models (VLM) contribute strong semantic understanding and cross-modal reasoning, enabling more informative and semantically aware sample selection [5]. Building on these capabilities, we aim to investigate whether foundation models can facilitate more effective selection of samples for annotation, thereby reducing labeling costs while improving model robustness in unstructured environments.

In this paper, we present FALCO, an active learning method for off-road freespace detection in autonomous driving. Under a limited annotation budget, FALCO selects informative unlabeled samples for annotation and model fine-tuning. It first estimates sample criticality from three complementary perspectives: semantic analysis of predictions with vision-language foundation models, prediction deviation measured by the vision foundation model SAM, and model uncertainty estimation. Building on this, we incorporate semantic partitioning with vision-language models, which categorize scene conditions, such as road surface, illumination, and vegetation coverage, that affect freespace

This work was supported by the National Natural Science Foundation of China under Grants U22A2061 and 92582205. S. Wang, C. Li, Y. Jia, H. Li and H. Zhao are with the State Key Laboratory of General Artificial Intelligence, Peking University, and also with the School of Intelligence Science and Technology, Peking University; Y. Chen is with the School of Automation, Beijing Institute of Technology; C. Min and J. Mei are with the Research Center for Intelligent Computing Systems, SKLP, Institute of Computing Technology, Chinese Academy of Sciences. Correspondence: S.Wang, wangshuai_cis@pku.edu.cn, and H.Zhao, zhaohj@pku.edu.cn.

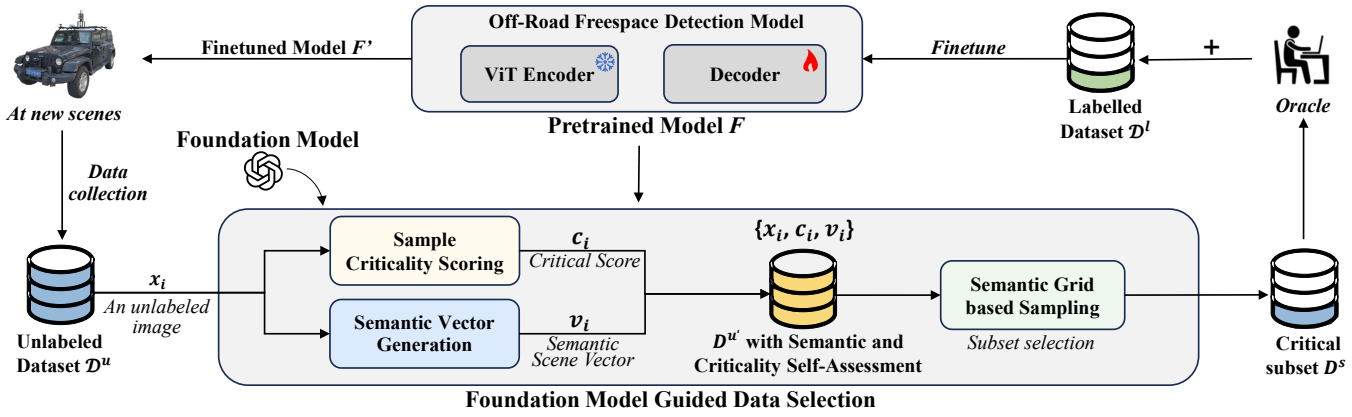

Fig. 1: Pipeline of the proposed FALCO framework. Unlabeled data collected in a new scene are evaluated by a foundation model guided data selection algorithm to select critical subset. The subset is then labeled and used to finetune the pretrained model to have a better performance.

detection. Finally, we design a data selection algorithm that jointly considers sample criticality and semantic diversity, constructing a set of key samples that balance scene coverage with the inclusion of critical cases for further annotation and fine-tuning. Experimental results demonstrate that our method significantly improves model robustness compared to baseline approaches. Our contributions are summarized as follows:

- We propose FALCO, a foundation model guided active learning framework for off-road drivable area detection. Leveraging the semantic understanding capability of vision-language models, FALCO partitions data at the semantic level and integrates this with sample criticality scoring. Our method ensures broad semantic coverage while prioritizing critical cases, thereby mitigating the long-tail effect and enhancing the representativeness of selected subsets.
- We introduce a novel data evaluation framework for assessing sample criticality by integrating three perspectives: semantic analysis from vision–language foundation model, prediction deviation from vision foundation model, and uncertainty estimation from pretrained model. This framework provides a more reliable assessment of sample criticality and improves alignment with challenging samples.
- We conduct extensive experiments on the ORAD-3D[6] dataset, demonstrating that our approach consistently outperforms state-of-the-art baselines. In particular, FALCO achieves substantial gains in robustness on long-tail and challenging samples, highlighting its effectiveness for off-road freespace detection under diverse and complex conditions.

II. RELATED WORKS

A. Off-road Freespace Detection

Freespace detection, i.e., pixel-wise labeling of traversable regions, is typically formulated as semantic segmentation. In off-road environments, methods developed for urban streets are often inadequate due to the absence of lane markings, pavements, or other artificial boundaries. Existing approaches can be broadly categorized into LiDAR-based, vision-based, and fusion-based.

LiDAR-based methods exploit geometric cues from point clouds for ground or terrain estimation. For example, GndNet [7] combines PointNet with pillar feature encoding for ground segmentation, while Forkel et al. [8] formulate terrain estimation as recursive Gaussian state estimation with maximum a posteriori optimization. The high cost of LiDAR sensors motivates vision-based alternatives. Vision-based methods typically rely on CNNs or transformers [9], [10], and recent studies further explore contrastive learning [11], [12] and self-supervised learning [13], [14]. To improve efficiency, Sun et al. proposed ROD [15], which combines a pretrained ViT with a lightweight decoder to achieve state-of-the-art performance while maintaining real-time inference speed. Fusion-based methods combine LiDAR geometry with image texture to improve accuracy [16], but their computational cost can limit real-time deployment.

B. Active Learning

Active learning seeks to maximize model performance under a limited annotation budget by selecting the most informative samples for labeling. Although query-synthesizing and stream-based settings have also been studied [17], most prior work adopts a pool-based protocol, where a subset is selected from an unlabeled data pool for annotation. Existing pool-based methods mainly rely on two criteria: uncertainty [18], [19] and diversity [2], [20].

Uncertainty-based methods prioritize samples that the current model finds difficult to predict. Early approaches use heuristic indicators such as posterior probability [21], entropy [22], and classification margin [23]. Later studies estimate uncertainty through predicted training loss [24] or expected influence on model performance [19], [25]. More recent extensions further incorporate task-specific cues: [26] integrates explainability and region-level uncertainty for semantic segmentation in driving scenes, while [12] combines active learning with contrastive representation learning for fine-grained off-road semantic segmentation and adopts risk-based frame selection under weak supervision. Despite

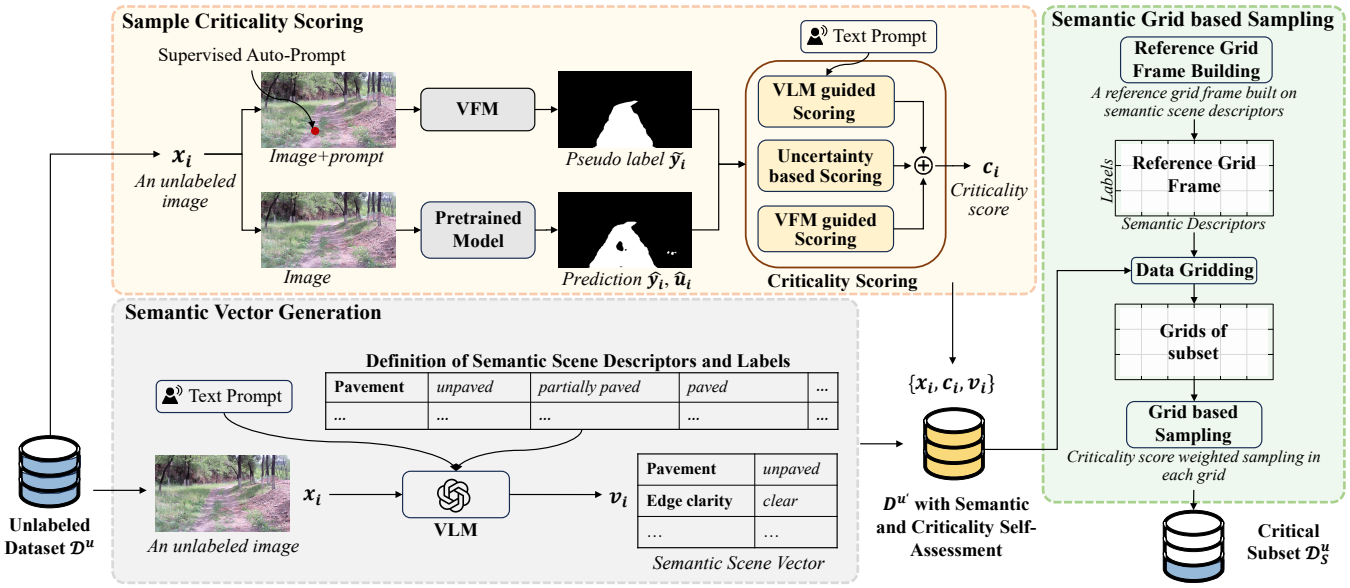

Fig. 2: Detailed process of the proposed active learning method. Unlabeled data are assessed via criticality scoring and semantic vector generation, and a semantic grid based sampling strategy selects critical subset for annotation.

their effectiveness, uncertainty-based methods become less reliable in complex and highly diverse unstructured environments, where prediction-based estimates may be inaccurate. In contrast, our method combines vision-language foundation models, vision foundation models, and model predictions to evaluate sample criticality more reliably.

Diversity-based methods aim to select a subset that better reflects the distribution of the unlabeled pool. Sener et al. [2] formulate sample selection as a k-Center problem and propose the CoreSet algorithm. FreeSel [27] extracts semantic patterns from a pretrained model and performs distance-based selection for one-pass fine-grained sampling. ActiveFT [28] optimizes sample selection in continuous space to balance representativeness and diversity. While diversity-based approaches guarantee feature-space coverage, proximity in feature space does not necessarily correspond to semantic similarity, meaning that rare but critical scenarios may still be overlooked.

III. METHODOLOGY

A. Problem Formulation

We formally define the active learning task for freespace detection in unstructured scenarios. As is demonstrated in Fig. 1, a freespace detection model $F(\cdot|\omega_0)$ with pretrained weights ω_0 is given. The model F takes an RGB image x as input and predicts pixel-wise logits over two classes(traversable vs. non-traversable): $z = F(x|\omega_0) \in \mathbb{R}^{H\times W\times 2}$. The predicted label mask is then obtained by $\hat{y}(h,w) = arg\max_{c\in\{0,1\}} z(h,w,c)$, $\hat{y} \in \{0,1\}^{H\times W}$. We also have access to a large data pool $D^u = \{x_i\}_{i\in[N]}$, where $[N] = \{1,2,...,N\}$.

The goal is to design a sampling strategy $S = \{n_j \in [N]\}_{j\in[B]}$ to select a subset $D_S^u = \{x_{n_j}\}_{j\in[B]} \subset D_u$ from D_u, where B is the annotation budget size for supervised finetuning. The model then has access to the labels $\{y_{n_j}\}_{j\in[B]}$ of this subset through oracle, obtaining a labeled data pool $D_S^l = \{x_{n_j}, y_{n_j}\}_{j\in[B]}$. Subsequently, the model F is supervisedly finetuned on D_S^l and its parameters are updated to ω_S. The objective of active learning is to find an optimal sampling strategy S_{opt} that selects an informative subset, thereby reducing $F(\cdot|\omega_S)$'s prediction error.

B. Sample Criticality Scoring

To comprehensively evaluate sample criticality, we design three types of scoring strategies: (1) VFM guided scoring, (2) uncertainty based scoring, and (3) VLM guided scoring. These scores are then integrated into a composite score that quantifies the criticality of each sample.

1) VFM guided Scoring: We leverage the generalization capability of VFMs by generating pseudo labels $\tilde{y}$ with it and using them as a reference to evaluate the quality of the model's predictions. Since the data are typically collected as continuous trajectories, we follow the standard paradigm of semi-supervised video object segmentation, as exemplified by SAM2 [4]. Specifically, a few coarse point prompts $\mathcal{P}$ are provided on the first frame of each trajectory, and pseudo labels are then propagated across the entire sequence using the model's video tracking capability:

$$\tilde{y}_i^{(1)} = \text{SAM2}\left(x_i^{(1)}, \mathcal{P}\right), \quad (1)$$

$$\tilde{y}_i^{(t)} = \text{SAM2}\left(x_i^{(t)} \mid \tilde{y}_i^{(1)}, \ldots, \tilde{y}_i^{(t-1)}\right), t = 2, \ldots, T. \quad (2)$$

This process requires only minimal human effort: for a trajectory containing several hundred frames, fewer than five coarse point annotations are typically sufficient for initialization, with only occasional correction in practice, to generate pseudo labels for the entire sequence. As a result, the annotation cost is negligible compared with precise manual labeling. Although the pseudo labels generated

by VFMs on unlabeled data exhibit lower average pixel-wise accuracy than the pretrained model's own predictions (see Table I), they preserve strong semantic and structural consistency with the true distribution of drivable regions. Prior work on learning with noisy labels [29] has shown that such structurally aligned but imperfect labels can still provide reliable guidance. Building on this property, we treat VFM pseudo labels as a weak supervisory signal and compute the Intersection-over-Union (IoU) between the model prediction $\hat{y}_i$ and the pseudo label $\tilde{y}_i$:

$$s_i^{vis} = \mathrm{IoU}(\hat{y}_i, \tilde{y}_i) = \frac{|\hat{y}_i \cap \tilde{y}_i|}{|\hat{y}_i \cup \tilde{y}_i|}, \tag{3}$$

where s_i^{vis} serves as a proxy for prediction quality and is further used to assess the criticality of each sample.

2) Uncertainty based Scoring: We further use model predictive confidence map as a proxy for uncertainty-related indicator of sample criticality. The intuition is that samples on which the model is less confident are more likely to reveal potential weaknesses and thus provide higher value for model improvement. For each pixel in the input image x_i, we define the pixel-wise confidence u_i as the maximum softmax probability over two classes:

$$u_i(h,w) = \max_{c\in\{0,1\}} \frac{\exp\big(z_i(h,w,c)\big)}{\sum_{c'=0}^{1}\exp\big(z_i(h,w,c')\big)}. \tag{4}$$

Unlike prior work that aggregates uncertainty across the entire image [30] or within generic foreground regions, we propose a task-specific variant tailored for freespace detection. Specifically, we compute the mean confidence only within the predicted drivable area:

$$s_i^{uc} = \frac{1}{|\{(h,w) : \hat{y}_i(h,w) = 1\}|} \sum_{(h,w):\hat{y}_i(h,w)=1} u_i(h,w). \tag{5}$$

thereby avoiding irrelevant background noise and focusing on the region of interest for autonomous navigation. This semantically constrained aggregation provides a more reliable measure of sample-level uncertainty, reflecting the model's confidence in its predictions.

3) Vision-language Model based Scoring: Building on the first two scoring strategies, we further incorporate semantic evaluation with GPT-4V as a vision–language model. Given each image data and the corresponding prediction of the pretrained model, GPT-4V is guided by a predefined structured textual prompt to assess the semantic reliability of freespace predictions along three complementary dimensions that are commonly used by human experts when evaluating drivable area segmentation:

- **Inclusion**(s_i^{inc}): whether the predicted drivable region sufficiently covers all clearly drivable surfaces.
- **Exclusion**(s_i^{exc}): whether non-drivable areas (e.g., vegetation, obstacles) are excluded from the prediction.
- **Consistency**(s_i^{con}): whether the overall mask is spatially coherent and logically consistent with the off-road scene.

The three semantic dimensions are complementary rather than hierarchical. To avoid arbitrarily prioritizing one dimension over the others, we assign equal weights and compute their average:

$$s_i^{vlm} = \frac{1}{3}\left(s_i^{inc} + s_i^{exc} + s_i^{con}\right). \tag{6}$$

We interpret s_{vlm} as a semantic reliability score that captures high-level consistency and plausibility of the predicted freespace. Unlike IoU or uncertainty, which mainly reflect pixel-level discrepancies, s_{vlm} emphasizes semantic mistakes such as missing entire drivable paths or incorrectly including obstacles.

4) Integrate Sample Criticality Scoring: Finally, we derive the sample criticality score by integrating the three perspectives described above. To ensure fair comparability across heterogeneous metrics, each score is min-max normalized to $[0, 1]$. We then design a weighted fusion function to integrate the three scores:

$$c_i = \alpha s_i^{vis} + \beta s_i^{uc} + \gamma s_i^{vlm}, \alpha + \beta + \gamma = 1, \tag{7}$$

where the fusion weights (α, β, γ) are tuned on labeled pretraining data via grid search with cross-validation [31], ensuring balanced integration of the three complementary signals while mitigating overfitting.

C. Semantic Vector Generation

We leverage the vision-language understanding capability of GPT-4V to automatically distinguish semantic scenarios that may influence freespace prediction. Unlike feature-based methods that struggle to capture high-level semantics, GPT-4V enables automated scene annotation by leveraging its strong vision-language understanding capability.

To capture aspects most relevant to freespace detection, we design semantic dimensions covering three categories: (1) road attributes, including road pavement, wetness, gravel, and vegetation on the road; (2) visual perception conditions, such as edge clarity, light, and weather; and (3) driving-related attributes, including road curvature and obstacles. Each dimension is discretized into multiple label categories, with each category representing a specific condition or degree of variation.

Formally, we define K semantic dimensions $\mathcal{K} = \{1, 2, \ldots, K\}$, where each dimension k is associated with a label set $\mathcal{V}_k = \{v_{k,1}, ..., v_{k,Q_k}\}$ of size Q_k. A structured text prompt guides GPT-4V to analyze each image and assign labels across all dimensions, with outputs enforced in JSON format to facilitate automatic parsing. By parsing these results, we obtain a semantic vector v_i for each image, which is subsequently used for semantic-aware sample selection.

D. Semantic Grid based Sampling

To ensure semantic coverage in data selection, we first construct a semantic grid where each cell corresponds to a predefined semantic label (e.g., sunny weather or partially paved surface). Each sample can be projected into multiple cells according to its semantic annotations, and the dataset

is thus divided into $T = \sum_{k\in\mathcal{K}} Q_k$ semantic subsets. The budget allocated to each semantic subset $D_{k,q}$ is

$$b_{k,q} = \min\left(|D_{k,q}|, \left\lfloor \frac{B}{T} \right\rfloor\right), \tag{8}$$

where $D_{k,q}$ denotes the set of samples belonging to the semantic dimension k and label q.

Within each subset, sampling is guided by criticality scores. Since lower scores indicate poorer model predictions and higher sample criticality, we invert the score $(1 - c_i)$ when computing selection probabilities. The selection probability of each sample i is defined as:

$$\rho_i = \frac{\exp\left((1-c_i)/\tau\right)}{\sum\limits_{(x_m,c_m)\in D_{k,q}} \exp\left((1-c_m)/\tau\right)}, \tag{9}$$

where τ is a temperature parameter controlling the balance between high and low-scored samples. Sampling proceeds without replacement, and all selected data are merged into the critical subset. If a subset has fewer samples than allocated, the remaining budget is filled by drawing additional samples from $\mathcal{D}^{u'} \setminus \mathcal{D}_S^u$ using the same probability rule. The overall procedure is summarized in Algorithm 1.

Algorithm 1 Semantic Grid based Sampling

Require: Unlabeled dataset $\mathcal{D}^{u'} = \{(x_i, c_i, v_i)\}_{i=1}^N$ with scores and semantic labels; total budget B; temperature τ; semantic dimension set $\mathcal{K} = \{1, 2, \ldots, K\}$; label value set $\mathcal{V}_k = \{v_{k,1}, \ldots, v_{k,Q_k}\}$ for each k

Ensure: Selected subset $\mathcal{D}_S^u$

```
Initialize D_S^u ← ∅, b_used ← 0, T ← Σ_{k∈K} Q_k
for all k ∈ K do
    for all q ∈ {1, ..., Q_k} do
        Construct subset D_{k,q} with condition (k, q)
        Compute budget b_{k,q} using Eq. (8)
        Compute sampling probability ρ_i using Eq. (9)
        Sample b_{k,q} data from D_{k,q} with ρ
        Update D_S^u and b_used
    end for
end for
if b_used < B then
    Compute sampling probabilities ρ_i using Eq. (9)
    Sample B − b_used data from D^{u'} \ D_S^u with ρ
    Update D_S^u
end if
return D_S^u
```

IV. EXPERIMENTS

A. Experimental Design

1) Baseline: We compare our method with two baselines and three state-of-the-art active learning methods. The baselines include Random and K-Means, where K-Means clusters the feature pool following the implementation in [28]. The active learning methods include CoreSet [2], FreeSel [27] and ActiveFT [28], all evaluated under the settings reported in their respective papers or released codes. Methods are evaluated by the prediction accuracy of a model finetuned on its selected subset, using the IoU metric as in M2F2-Net [32]. In the subsequent experiments, mIoU denotes the mean IoU over the dataset, while $\text{mIoU}_{k\%}$ refers to the mean IoU computed over the worst-performing k% of samples.

2) Dataset: For freespace detection, we use two datasets: ORFD [33] and ORAD-3D. We use ROD [15] as the base detector, pretraining it on ORFD and finetuning it on subsets selected from ORAD-3D, which contains over 100 trajectories, follows the same format as ORFD, and is about four times larger. For active learning, ORAD-3D is split into training, validation, and test sets with a ratio of 8:1:1.

Table I compares the prediction quality of the pretrained model F and the pseudo labels generated by SAM2 on the two datasets. While F performs well on ORFD, its accuracy drops markedly on ORAD-3D, highlighting the need for finetuning in new scenes. SAM2 also underperforms F on both datasets, indicating that its pseudo labels are unsuitable as direct training targets.

TABLE I: Comparison of mean IoU on different datasets.

Dataset	mIoU of F	mIoU of SAM2
ORFD	0.938	0.859
ORAD-3D	0.873	0.844

We analyze the IoU distribution of F's predictions on both datasets in Fig. 3. The results reveal a clear long-tail effect: while most frames are relatively easy with high IoU, a smaller set of challenging samples concentrates in the lower tail of the distribution ($\text{mIoU}_{1\%} < 0.5$ in ORAD-3D). These hard cases are particularly critical for driving safety, as the model performs poorly on them, highlighting the importance of actively selecting and finetuning on long-tail samples.

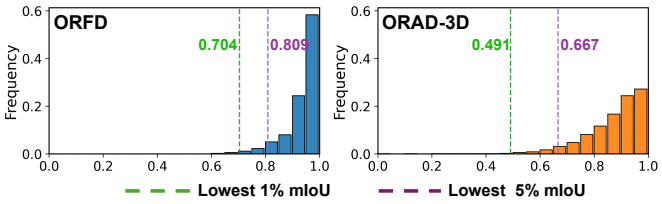

Fig. 3: Distribution of IoU of pretrained model on ORFD and ORAD-3D datasets.

3) Implementation: We finetune both attention and MLP blocks of ROD model using LoRA [34] with rank $r = 4$, using the same loss and optimizer as in [15]. During VFM guided pseudo-label propagation, small errors are tolerated because the masks still preserve the overall structure of the drivable region and are used only as weak reference signals for scoring. In challenging cases, such as turning or abrupt surface material changes, sparse point prompts are used to correct occasional drift by re-initializing propagation. Such correction is needed less than once per trajectory on average.

The fusion weights used to combine the three criticality scores are set to $\alpha = 0.35$, $\beta = 0.35$, and $\gamma = 0.3$, obtained by a lightweight cross-validated grid search using the pretrained model outputs on the held-out ORFD split, which is unseen during pretraining and disjoint from the ORAD-3D.

TABLE II: Comparison to other methods on the testing set of ORAD-3D. We report the average of multiple trials. Best and second best results are highlighted in **bold** and underline.

Method	1%			5%			10%		
	mIoU	$\text{mIoU}_{1\%}$	$\text{mIoU}_{5\%}$	mIoU	$\text{mIoU}_{1\%}$	$\text{mIoU}_{5\%}$	mIoU	$\text{mIoU}_{1\%}$	$\text{mIoU}_{5\%}$
Random	90.11%	42.37%	63.75%	93.15%	48.91%	72.12%	94.31%	60.41%	77.28%
KMeans	90.02%	38.11%	60.87%	93.09%	53.75%	72.48%	94.33%	60.34%	77.48%
Coreset	90.01%	39.94%	62.32%	93.16%	51.13%	72.54%	94.31%	60.43%	77.50%
FreeSel	90.24%	41.99%	62.43%	**93.35%**	57.68%	74.91%	94.41%	67.78%	79.64%
ActiveFT	**90.34%**	45.83%	64.40%	93.28%	57.40%	74.89%	**94.57%**	68.88%	80.18%
FALCO (Ours)	90.14%	**51.45%**	**65.77%**	93.26%	**61.80%**	**75.42%**	94.43%	**72.44%**	**80.80%**

Since the search is performed on existing outputs, it requires neither additional annotation nor repeated model training and introduces only negligible overhead. In addition, the temperature parameter is set to $\tau = 0.5$ to balance sampling between high- and low-criticality samples.

B. Comparative experiments on ORAD-3D

Comparisons on the ORAD-3D dataset across different sampling ratios are detailed in Table II. Although our method does not always achieve the highest average mIoU, its performance remains very close to the strongest baseline. The main advantage of our approach lies in its robustness on challenging samples. Due to the long-tail distribution of challenging samples, model performance on these cases is often diluted in the overall mIoU. To better capture robustness on such samples, we report the mIoU over the lowest-k% of predictions, providing a focused evaluation of model performance of the most difficult cases. The results indicate that our method outperforms other methods, achieving a 5.6% improvement in $\text{mIoU}_{1\%}$ at selection ratio 1%, and a 4.1% improvement in $\text{mIoU}_{1\%}$ at selection ratio 5%.

TABLE III: Performance of Individual and Combined Criticality Scorers on ORAD-3D

Score	Corr.	$\text{Corr.}_{\text{Norm.}}$	$\text{Corr.}_{\text{Chal.}}$	Rec.@10%
s^{vis}	0.7552	0.7027	0.7091	0.5280
s^{uc}	0.6765	0.6141	0.3925	0.4511
s^{vlm}	0.5351	0.4375	0.4735	0.4111
$s^{vis} + s^{uc}$	**0.7969**	**0.7452**	0.6942	**0.5543**
$s^{vis} + s^{uc} + s^{vlm}$	0.7963	0.7373	**0.7306**	0.5442

Notes. Corr. denotes Pearson correlation between criticality scores and IoU ($\text{Corr.}_{\text{Norm.}}$ for samples with IoU ≥ 0.5, $\text{Corr.}_{\text{Chal.}}$ for IoU < 0.5). Rec.@k% is the recall of ground-truth lowest-k% IoU samples within the top-k% ranked by score. Correlations > 0.7 indicate strong alignment; values around 0.5 indicate moderate alignment.

C. Data Critical Scoring Experiment

Table III presents the performance of different scorers. Among the individual scores, the visual score s^{vis} achieves the strongest correlation and recall, while s^{uc} also shows good correlation but is less effective on challenging cases. The VLM based score s^{vlm} performs weaker overall but introduces complementary semantic information. When combining scores, $s^{vis} + s^{uc}$ attains the strongest performance in terms of overall correlation and recall. However, adding the VLM based score ($s^{vis} + s^{uc} + s^{vlm}$) further improves performance on challenging cases, where semantic reasoning provides additional benefits. Since robustness on rare and difficult samples is particularly critical for off-road driving, we adopt the combination of all three scores as our final criticality score. This choice ensures that overall performance remains comparable to the best two-score combination, while offering a clear advantage in long-tail scenarios. This design not only pursues optimal overall relevance but also considers long-tail robustness.

D. Scene Semantic Understanding Experiment

To illustrate the capability of VLMs in semantic understanding, we provide recognition results of several representative semantic dimensions under diverse scene conditions, as shown in Fig. 4. These examples highlight the effectiveness of the proposed semantic vector generation module in capturing multi-dimensional semantics under diverse conditions.

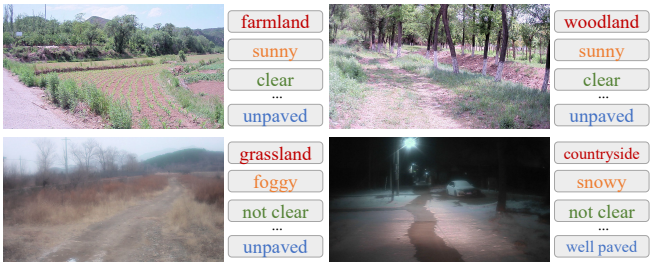

Fig. 4: Examples of semantic vector generation, each image is annotated across multiple semantic dimensions, including scene, weather, edge clarity, and road pavement.

E. Critical Subset Selection Experiment

To validate the effectiveness of the proposed Semantic Grid based Sampling algorithm, we analyze the semantic distribution of the selected subset from ORAD-3D. As illustrated in Fig. 6, taking edge clarity and obstacle level as examples, our method increases the proportion of long-tail semantic categories (e.g., blocked, moderate, severe in

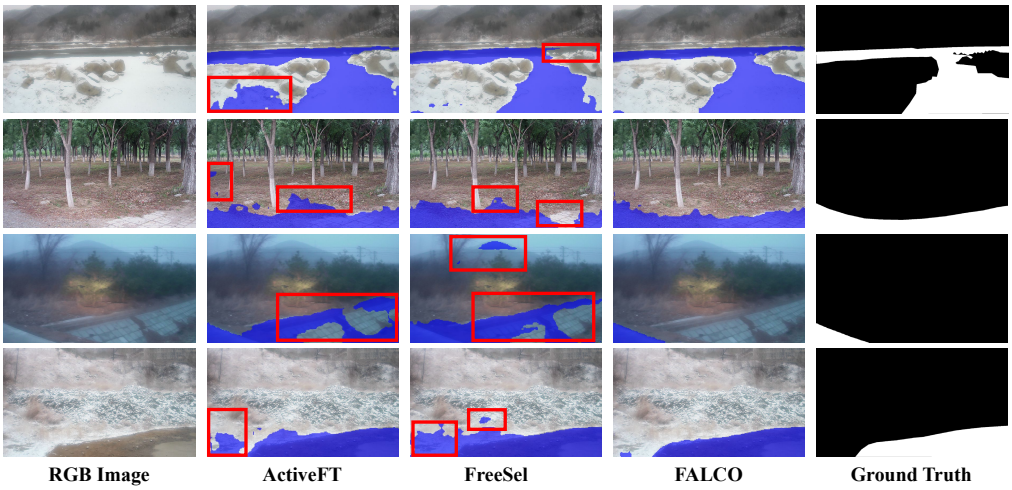

Fig. 5: Qualitative results of ActiveFT, FreeSel and our method on ORAD-3D dataset, with the ratio 1%. The red boxes are the areas where finetuned models using subsets selected by other methods predict incorrectly but ours predicts correctly.

obstacle, and none in edge clarity) compared with the full dataset. As shown in Fig. 7, our method yields a wider IoU range (higher IQR) and captures more long-tail cases that are often overlooked by other methods.

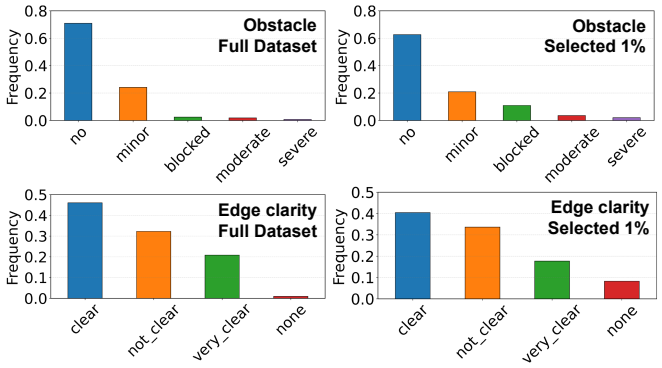

Fig. 6: Semantic distribution of ORAD-3D vs. 1% subset selected by FALCO on dimensions of obstacle and edge clarity. "None" in edge clarity indicates no obvious visible road area.

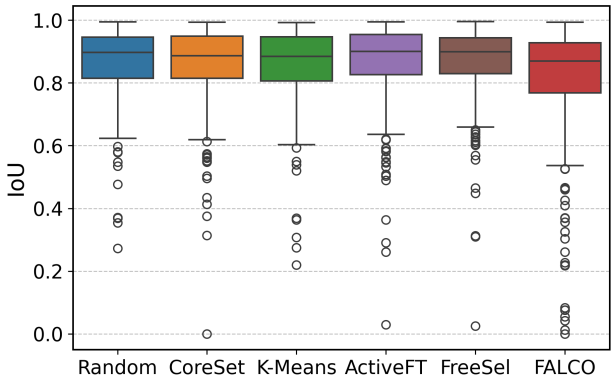

Fig. 7: IoU distribution of 1% ORAD-3D subsets selected by different methods.

We visualize the prediction results of the finetuned model by our method and compare it with those of ActiveFT and FreeSel with a 1% selection ratio on the ORAD-3D dataset in Fig. 5. Although all models perform well on common cases, such as clear weather and well-defined roads, their performance degrades in rare and challenging scenarios. In particular, models trained with subsets selected by ActiveFT and FreeSel, often misclassify snow-covered regions, artificial ditches, or obstacles as drivable areas. This is mainly because the subsets chosen by them do not sufficiently cover such long-tailed and difficult conditions, leading to limited finetuning and degraded robustness in these cases.

F. Ablation Study

To investigate the contribution of different components in our proposed method, we conducted a systematic ablation study, as shown in Table IV. Specifically, we compared the full model with two variants: one without criticality score-based weighted sampling and the other without semantic grid based sampling. The results demonstrate that each component contributes positively to performance. Moreover, when both are integrated, the algorithm achieves not only improved average mIoU but also enhanced robustness on challenging samples.

TABLE IV: Ablation study on the effectiveness of different modules in FALCO.

Method	1%		5%	
	mIoU	$\text{mIoU}_{1\%}$	mIoU	$\text{mIoU}_{1\%}$
FALCO w/o Crit.	89.94%	39.83%	93.14%	60.32%
FALCO w/o Sem.	89.99%	47.05%	92.93%	59.50%
FALCO (full)	**90.14%**	**51.45%**	**93.26%**	**61.80%**

Notes. "w/o Crit." = removing the *criticality score*, i.e., not performing probability sampling based on scoring. "w/o Sem." = removing the *semantic vector*, i.e., not performing sampling after grid partitioning in the semantic space.

V. CONCLUSION AND FUTURE WORK

In this paper, we introduced FALCO, a novel active learning framework that leverages foundation models to enable cost-effective freespace detection in unstructured off-road environments. By combining vision foundation models, model uncertainty, and vision–language models, our method provides a reliable measure of sample criticality. Furthermore, the proposed semantic grid based sampling ensures balanced semantic coverage while prioritizing rare but safety-critical scenarios. Extensive experiments on ORAD-3D demonstrate that FALCO substantially improves robustness on difficult cases while maintaining competitive overall performance under comparable labeling budgets.

While our results highlight the promise of integrating foundation models into active learning pipelines, several directions remain open for exploration. (1) Available datasets for off-road autonomous driving are still limited, and task definitions often differ across benchmarks. For example, ORFD and ORAD-3D adopt a binary traversability formulation, whereas another popular off-road RELLIS-3D [35] defines fine-grained multi-class terrain categories. This mismatch prevents direct cross-dataset evaluation, and advancing this field will require larger and more diverse datasets as well as unified semantic abstractions. (2) Although our scoring framework combines vision–language models with other signals to provide a more accurate assessment of sample criticality, the ability of current large models to fully understand off-road scenes and model predictions remains limited. Future research should explore fine-tuning foundation models with domain-specific data to enhance their semantic understanding of complex off-road environments. (3) Our study focuses on image-based approaches, whereas extending active learning to 3D LiDAR data and multimodal perception systems represents an important future direction.

REFERENCES

[1] C. Min *et al.*, "Autonomous driving in unstructured environments: How far have we come?" *ArXiv*, vol. abs/2410.07701, 2024.

[2] O. Sener *et al.*, "Active learning for convolutional neural networks: A core-set approach," *arXiv: Machine Learning*, 2017.

[3] W. Xu *et al.*, "Activedc: Distribution calibration for active finetuning," *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16 996–17 005, 2024.

[4] N. Ravi *et al.*, "Sam 2: Segment anything in images and videos," *arXiv preprint arXiv:2408.00714*, 2024.

[5] J. Zhang *et al.*, "Vision-language models for vision tasks: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 8, pp. 5625–5644, 2024.

[6] C. Min *et al.*, "Advancing off-road autonomous driving: The large-scale orad-3d dataset and comprehensive benchmarks," *ArXiv*, vol. abs/2510.16500, 2025.

[7] A. K. Paigwar *et al.*, "Gndnet: Fast ground plane estimation and point cloud segmentation for autonomous vehicles," *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 2150–2156, 2020.

[8] B. Forkel *et al.*, "Probabilistic terrain estimation for autonomous off-road driving," *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 13 864–13 870, 2021.

[9] C. Chung *et al.*, "Pixel to elevation: Learning to predict elevation maps at long range using images for autonomous offroad navigation," *IEEE Robotics and Automation Letters*, vol. 9, pp. 6170–6177, 2024.

[10] K. Viswanath *et al.*, "Offseg: A semantic segmentation framework for off-road driving," *2021 IEEE 17th International Conference on Automation Science and Engineering (CASE)*, pp. 354–359, 2021.

[11] B. Gao *et al.*, "Fine-grained off-road semantic segmentation and mapping via contrastive learning," *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 5950–5957, 2021.

[12] B. Gao *et al.*, "An active and contrastive learning framework for fine-grained off-road semantic segmentation," *IEEE Transactions on Intelligent Transportation Systems*, vol. 24, no. 1, pp. 564–579, 2023.

[13] J. Seo *et al.*, "Learning off-road terrain traversability with self-supervisions only," *IEEE Robotics and Automation Letters*, vol. 8, pp. 4617–4624, 2023.

[14] S. Jung *et al.*, "V-strong: Visual self-supervised traversability learning for off-road navigation," *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 1766–1773, 2024.

[15] T. Sun *et al.*, "Rod: Rgb-only fast and efficient off-road freespace detection," in *2025 IEEE International Conference on Robotics and Automation (ICRA)*, 2025, pp. 9787–9793.

[16] Y. Liu *et al.*, "Petrv2: A unified framework for 3d perception from multi-camera images," *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3239–3249, 2023.

[17] A. Narr *et al.*, "Stream-based active learning for efficient and adaptive classification of 3d objects," *2016 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 227–233, 2016.

[18] J. Ruckin *et al.*, "Informative path planning for active learning in aerial semantic mapping," *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 11 932–11 939, 2022.

[19] Z. Liu *et al.*, "Influence selection for active learning," *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9254–9263, 2021.

[20] K. Li *et al.*, "Roial: Region of interest active learning for characterizing exoskeleton gait preference landscapes," *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 3212–3218, 2021.

[21] K. Wang *et al.*, "Cost-effective active learning for deep image classification," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 27, pp. 2591–2600, 2017.

[22] A. J. Joshi *et al.*, "Multi-class active learning for image classification," *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2372–2379, 2009.

[23] S. Tong *et al.*, "Support vector machine active learning with applications to text classification," *J. Mach. Learn. Res.*, pp. 45–66, 2000.

[24] S. Hwang *et al.*, "Joint semi-supervised and active learning via 3d consistency for 3d object detection," *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 4819–4825, 2023.

[25] A. LaGrassa *et al.*, "Task-oriented active learning of model preconditions for inaccurate dynamics models," *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 16 445–16 445, 2024.

[26] S. Mandalika *et al.*, "Segxal: Explainable active learning for semantic segmentation in driving scene scenarios," in *International Conference on Pattern Recognition*, 2024.

[27] Y. Xie *et al.*, "Towards free data selection with general-purpose models," *ArXiv*, vol. abs/2309.17342, 2023.

[28] Y. Xie *et al.*, "Active finetuning: Exploiting annotation budget in the pretraining-finetuning paradigm," *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23 715–23 724, 2023.

[29] Q. Xie *et al.*, "Self-training with noisy student improves imagenet classification," *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10 684–10 695, 2020.

[30] A. Kendall *et al.*, "What uncertainties do we need in bayesian deep learning for computer vision?" *ArXiv*, vol. abs/1703.04977, 2017.

[31] F. Pedregosa *et al.*, "Scikit-learn: Machine learning in python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011.

[32] H. Ye *et al.*, "M2f2-net: Multi-modal feature fusion for unstructured off-road freespace detection," *2023 IEEE Intelligent Vehicles Symposium (IV)*, pp. 1–7, 2023.

[33] C. Min *et al.*, "Orfd: A dataset and benchmark for off-road freespace detection," *2022 International Conference on Robotics and Automation (ICRA)*, pp. 2532–2538, 2022.

[34] J. E. Hu *et al.*, "Lora: Low-rank adaptation of large language models," *ArXiv*, vol. abs/2106.09685, 2021.

[35] P. Jiang *et al.*, "Rellis-3d dataset: Data, benchmarks and analysis," *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 1110–1116, 2021.