A. Limitations and Social Impact

Limitations. As a planning-oriented active learning approach, we have achieved significant effects within the planning metrics. However, the 30% of data we selected still falls far short in training the model's perception and prediction capabilities compared to using 100%. Experiments in Appendix E demonstrate that model perception and prediction gradually strengthen with an increase in data volume. This is typical in similar fields, such as active learning for segmentation, and our method has not overcome this bottleneck. Nonetheless, within the E2E-AD framework, we have effectively identified valuable samples, reduced annotation costs, and avoided overfitting.

Social Impact. Our method will have a positive societal impact by enabling the selection of the most valuable data for annotation from large volumes of autonomous driving data. This will reduce resource waste, improve model performance, and enhance driving safety. However, it is inevitable that this will reduce the demand for annotators, potentially leading to fewer employment opportunities.

B. Related Work

B.1. End-to-End Autonomous Driving

The concept of end-to-end autonomous driving has roots dating back to the 1980s [52]. In the era of deep learning, early efforts focused on the straightforward mapping [49]. Subsequently, [36, 71] explored the application of reinforcement learning to develop an end-to-end driving policy. Some state-of-the-art student models [16, 62] are developed based on them while PlanT [54] suggested employing a Transformer for the teacher model. LBC [7] and DriveAdapter [27] involved initially training a teacher model with privileged inputs. In later works, multiple sensors are used. Transfuser [9, 53] employed a Transformer for camera and LiDAR fusion. LAV [6] adopted PointPainting [61]. Interfuser [57] injected safety-enhanced rules during the decision-making process. ThinkTwice [29] introduced a DETR-like scalable decoder paradigm for the student model. ReasonNet proposed specific modules for student models to better exploit temporal and global information. In [22], they suggested formulating the output of the student as classification problems to avoid averaging. ST-P3 [17] unified the detection, prediction, and planning tasks into the form of BEV segmentation. UniAD [19] adopted Transformer to connect different tasks. Further, VAD [30] reduced some potential redundant modules in UniAD while demonstrating better performance.

B.2. Active Learning

Active learning algorithms exploit the limited annotation budget by selecting the most informative samples for labeling. They select data samples based on the criterion of either uncertainty or diversity. Uncertainty-based algorithms prefer those difficult samples most confusing for the models. The difficulty of each data sample may be measured by prediction entropy [31, 44], prediction inconsistency [14], loss estimation [69] or its potential influence for model training [12, 40]. Alternatively, other methods pay attention to the diversity of the selected subset. Some early works [56, 59] mainly consider the representation diversity in the global image level, while following papers [1, 39] dig into the regional information to deal with fine-grained detection or segmentation tasks. Furthermore, some recent works [65, 66, 68] utilize the strong representation ability of models pretrained on large datasets to measure the image diversity of the target dataset more accurately. Recently, CRB [46] has pioneered active learning to LiDAR-based 3D object detection and KECOR [45] greedily select informative point clouds by maximizing the kernel coding rate in AD.

However, most prior works focus on the traditional tasks like classification, detection [15], or segmentation [20, 55, 64], but the recently prominent planning-oriented end-to-end AD setting is hardly explored. Instead of just simple prediction probability, The task model outputs the future ego-vehicle trajectory. Besides, this task requires to reason from the interaction [24] between ego-vehicle and surroundings, which cannot be reflected from superficial visual patterns. To this end, we fill in this gap by devising novel uncertainty and diversity metrics for active learning of end-to-end AD.

C. More Discussion about ActiveAD

We believe that contributing to the community extends beyond proposing novel neural networks. **Identifying key issues and conducting preliminary explorations are equally vital. True innovation emerges from uncovering and understanding challenges, setting the stage for meaningful progress.** In this work, (1) We take the initial step to point out and analyze the data problem for E2E-AD. (2) Based on the characteristics of AD tasks, we design specific metrics to select samples which could optimize the planning performance by active learning, which fits planning-oriented spirits of E2E-AD. (3) The strong performance and the comprehensive ablation studies verify our claims.

What's more, we notice that recent events in the E2E-AD community further validate the major claims of our work:

1. In CVPR 2024 CARLA challenge (June 2024) [1] , the winner solutions of both sensor and map tracks mention that they filter those less valuable frames during training. As a result, one winner state that by reducing the dataset size by 49%, with slightly improved performance. Their heuristic effectively removes redundant frames without losing information [73].

2. Tesla, one of the world's leading autonomous driving technology company, claims that only about 1/10,000 of

[1] https://opendrivelab.com/challenge2024/#carla

distance driven is useful for training, by their CEO Elon Musk in May, 2024[2].

We could observe that practioners in both academia and industry have both discovered the importance of data filtering. As the first work to study the data issue and active learning for E2E-AD, we believe the discoveies and insights of this work are worth sharing in the community.

D. Experiments Details

D.1. Experiments Setup

End-to-end Autonomous Driving Models. ST-P3 [17] is an interpretable end-to-end vision-based network for autonomous driving that achieves better spatial-temporal feature learning. UniAD [18] leverages information from multiple preceding tasks to enhance goal-oriented planning and demonstrates outstanding performance in all aspects, including perception, prediction, and planning. VAD [30] introduces a vectorized paradigm as a substitute for the dense rasterized scene representation used in previous studies. This approach facilitates a more focused analysis of instance-level structural information, leading to excellent end-to-end planning performance. Moreover, it achieves substantial reductions in computational requirements, decreases the reliance on training devices, and accelerates training speed. Consequently, we adopt the lightweight version, VAD-Tiny, as the starting point for our experiments.

Active Learning Baselines. As mentioned in the related works, end-to-end autonomous driving is a novel and underexplored task for active learning. It is difficult to directly transfer existing active learning approaches, which are usually based on predictive probability analysis, to this task. Therefore, we choose three classic methods that are more transferable and relevant as baselines: Coreset, a feature selection-based approach; VAAL, a task-agnostic method; KECOR [45]: a 3D Object Detection active learning method; ActiveFT, which utilizes pre-trained features. 1) Coreset [56] formulates the data selection process as a k-Center problem on the learned embeddings of both labeled and unlabeled data. We utilize the features prior to the trajectory planning head [30] as the embeddings. 2) VAAL [59] employs the adversarial learning paradigm, utilizing a variational autoencoder (VAE)[33] to extract image features from the nuscenes dataset, along with a discriminator network that distinguishes between labeled and unlabeled images. The VAE aims to deceive the discriminator by making it classify all samples as labeled data, while the discriminator strives to accurately identify the unlabeled samples in the data pool. Based on this approach, the selected unlabeled samples are then annotated. 3) KECOR [45] identifies the most informative point clouds to acquire labels for 3D annotations through the lens of information theory. Samples selected based on this criterion are used for our end-to-end training. 4) ActiveFT [65] uses pretrained features to optimize the distance between the distributions of labeled and unlabeled sets. In state-of-the-art autonomous driving methods, BEV features [35] are the commonly used representation. We adopt ActiveFT to use BEV features for data selection, and its strength lies in the ability to select all data under the budget at once, without the need for iterative selection.

Annotation Budget. In the scenario of active learning, the annotation budget is typically predetermined. Considering the complexity of end-to-end autonomous driving models and the diversity of tasks (including the final planning task as well as auxiliary perception and prediction tasks), we have set the annotation budget as 30%. Meanwhile, We further report the performance of ActiveAD with the budget from 10% to 50% of the data in Tab. 8. We observe that the planning performance is saturated around 30 % and thus we choose 30% as the stop threshold in the main paper.

D.2. Metrics Explanation

In this paper, we utilize the evaluation metrics from VAD [30], which is consistent with ST-P3 [17]. Therefore, the results from these two papers can be directly applied. Recently, inconsistencies in the UniAD metrics [19] have been identified within the community [38, 48]. We reference the content in [48] to provide more details about the evaluation metrics. The output trajectory τ is formatted as 6 waypoints in a 3-second horizon, i.e., $\tau = [(x_1, y_1), (x_2, y_2), ..., (x_6, y_6)]$. Then, the L2 loss is computed as:

$$l_2 = \sqrt{(\tau - \hat{\tau})^2} = \left[\sqrt{(x_i - \hat{x}_i)^2 + (y_i - \hat{y}_i)^2}\right]_{i=1}^{6}, \quad (7)$$

where $l_2 \in \mathbb{R}^{6\times1}$ and $\hat{\tau}$ denotes ground truth trajectory. Then, the average L2 loss $\bar{l}_2 \in \mathbb{R}^{6\times1}$ can be computed by averaging l_2 for each sample in the test set.

UniAD [19] uses the value in the exact timestep as the L2 loss at the k-th second $(k = 1, 2, 3)$:

$$L_{2,k}^{\text{UniAD}} = \bar{l}_2[2k]. \quad (8)$$

ST-P3 [17] and VAD [30] use the the average error from 0 to k second as L2 loss at the k-th second:

$$L_{2,k}^{\text{VAD}} = \frac{\sum_{t=1}^{2k} \bar{l}_2[t]}{2k}. \quad (9)$$

Given the collision times $\mathcal{C} \in \mathbb{N}^{6\times1}$ at each timestep. Similarly, UniAD reports the collision $\mathcal{C}_k^{\text{uniad}}$ at the k-th second $(k = 1, 2, 3)$ as $\mathcal{C}[2k]$, while VAD reports $\mathcal{C}_k^{\text{VAD}}$ as the average from 0 to k second.

Besides the variations in calculation methodologies, there is a distinction in the generation of ground truth occupancy maps between the two metrics. UniAD exclusively accounts

[2]https://x.com/elonmusk/status/1787768103449010597

Table 6. **Planning Performance with VAD-Base.** ActiveAD (w/o incremental) refers to the selection of all data solely based on diversity selection. ActiveAD (w/ incremental) indicates performing incremental selection based on an initial set.

Base Model	Percent	Selection Method	Average L2 (m) ↓				Average Collision (%) ↓			
			1s	2s	3s	Avg.	1s	2s	3s	Avg.
ST-P3 [17]	100%	-	1.33	2.11	2.90	2.11	0.23	0.62	1.27	0.71
UniAD† [19]	100%	-	0.42	0.64	0.91	0.67	-	-	-	-
VAD-Base* [30]	100%	-	0.39	0.66	1.01	0.69	0.08	0.16	0.37	0.20
VAD-Tiny* [30]	100%	-	0.38	0.68	1.04	0.70	0.15	0.22	0.39	0.25
VAD-Base	10%	Random	0.49	0.81	1.20	0.83	0.38	0.57	0.91	0.62
	10%	ActiveAD(w/o incremental)	**0.48**	**0.76**	**1.14**	**0.79**	**0.24**	**0.43**	**0.68**	**0.45**
VAD-Base	20%	Random	0.47	0.78	1.15	0.80	0.32	0.47	0.75	0.51
	20%	ActiveAD(w/o incremental)	0.44	0.75	1.10	0.76	0.25	**0.34**	**0.61**	0.40
	20%	ActiveAD(w/ incremental)	**0.42**	**0.70**	**1.08**	**0.73**	**0.16**	0.35	0.64	**0.38**
VAD-Base	30%	Random	0.44	0.74	1.08	0.75	0.16	0.34	0.54	0.35
	30%	ActiveAD(w/o incremental)	0.42	0.71	1.05	0.73	0.14	0.29	0.49	0.31
	30%	ActiveAD(w/ incremental)	**0.40**	**0.67**	**0.93**	**0.67**	**0.09**	**0.21**	**0.35**	**0.22**

for the vehicle category in creating ground truth occupancy maps, whereas ST-P3 and VAD incorporates both vehicle and pedestrian categories. This discrepancy results in different collision rates for the same planned trajectories when evaluated by these metrics, although it has no effect on the L2 error measurement. As a result, the collision rate in UniAD may be higher than reported, and this has been confirmed in [38] where VAD demonstrates superior performance in terms of collision rates. Consequently, we use a '-' in Tab. 1 instead of displaying specific values.

Taking into account the advantages of VAD in terms of model lightweighting (for instance, the ability to train using a 3090 GPU) as well as its leading position in comprehensive performance, we explore active learning based on the VAD model in this paper. This exploration is conducted from the perspective of data, aiming to provide insightful analysis.

D.3. Experiment Results for VAD-Base

Tab. 6 presents the experimental results of our method based on the VAD-Base model. Compared to the baseline of random selection, our method—whether it be the one-time sample selection based on Ego-Diversity or the complete method that performs Incremental Selection starting from an initial dataset—has shown significant advantages. Consistent with the conclusions in the main paper, using 30% of the data, our approach achieves performance on par with using the entire dataset, validating the effectiveness and universality of our method.

D.4. Experiments Results for Different Annotation Increment.

In active learning, the sample selection ratio in each round plays a crucial role in determining performance. In our main experiments, we chose 10% as the number of samples selected per round. Here, we set the interval to 5%, resulting in training data proportions of 5%, 10%, 15%, ..., up to 30%. Since other active learning methods do not show a significant advantage over random selection, we present a comparison of our method ActiveAD, with Random in the Tab. 7. Our findings reveal that, across different initialization ratios and selection intervals, ActiveAD consistently demonstrates robust performance advantages, underscoring its versatility across various labeling scenarios.

E. Perception and Prediction Performance.

Existing end-to-end training models [18, 30] often utilize visual information as auxiliary tasks to assist core objective planning. The main experiment shown in Tab. 1, demonstrates our advantage in planning metrics, while we are also curious about perception and prediction task performance. Tab. 8 displays the performance after training with different proportions of data. The perception metrics include NDS(nuScenes detection score), mAP(mean Average Precision), mATE(mean Average Translation Error), mASE(mean Average Scale Error), mAOE(mean Average Orientation Error), mAVE(mean Average Velocity Error), mAAE(mean Average Attribute Error) which are sourced from the nuScenes dataset setting [4]. The prediction metrics include minADE (minimum Average Displacement Error), minFDE (minimum Final Displacement Error) and MR (Miss Rate) and

Table 7. **Planning Performance with 5% Annotation Budget Per Selection Round on VAD-Tiny.**

Selection Method	Percent	Average L2 (m) ↓				Average Collision (%) ↓			
		1s	2s	3s	Avg.	1s	2s	3s	Avg.
Random	5%	0.66	1.10	1.60	1.12	0.21	0.52	1.18	0.64
ActiveAD	5%	**0.63**	**1.04**	**1.51**	**1.06**	**0.15**	**0.49**	**1.02**	**0.55**
Random	10%	0.51	0.83	1.23	0.86	0.40	0.62	0.98	0.67
ActiveAD	10%	**0.45**	**0.81**	**1.17**	**0.81**	**0.17**	**0.38**	**0.76**	**0.44**
Random	15%	0.49	0.81	1.21	0.84	0.27	0.54	0.84	0.55
ActiveAD	15%	**0.47**	**0.76**	**1.15**	**0.79**	**0.24**	**0.37**	**0.63**	**0.41**
Random	20%	0.49	0.80	1.17	0.82	0.36	0.49	0.77	0.54
ActiveAD	20%	**0.43**	**0.77**	**1.11**	**0.77**	**0.19**	**0.35**	**0.66**	**0.40**
Random	25%	0.47	0.77	1.13	0.79	0.23	0.37	0.59	0.40
ActiveAD	25%	**0.41**	**0.69**	**1.05**	**0.72**	**0.16**	**0.29**	**0.54**	**0.33**
Random	30%	0.45	0.76	1.12	0.78	0.17	0.30	0.63	0.37
ActiveAD	30%	**0.42**	**0.67**	**1.00**	**0.70**	**0.08**	**0.19**	**0.41**	**0.23**

Table 8. **All tasks' performance under different selection ratio.**

Ratio	Planning		Perception							Prediction			
	Avg. L2 ↓	Avg. Col. ↓	NDS ↑	mAP ↑	mATE ↓	mASE ↓	mAOE ↓	mAVE ↓	mAAE ↓	minADE ↓	minFDE ↓	MR ↓	EPA ↑
10%	0.83	0.43	16.56	9.80	0.95	0.43	0.98	1.31	0.47	1.28	1.89	0.195	0.230
20%	0.76	0.39	21.46	14.77	0.83	0.45	0.84	0.99	0.49	1.10	1.59	0.161	0.373
30%	0.68	**0.21**	25.60	15.85	0.84	0.39	0.78	0.83	0.40	1.01	1.43	0.147	0.402
40%	**0.66**	0.24	27.12	18.20	0.81	0.36	0.83	0.79	0.35	0.96	1.36	0.145	0.414
50%	0.68	0.23	29.29	19.72	0.85	0.34	0.80	0.76	0.31	0.93	1.28	0.142	0.430
100%	0.70	0.25	**36.11**	**26.65**	**0.74**	**0.31**	**0.76**	**0.67**	**0.23**	**0.84**	**1.16**	**0.134**	**0.534**

EPA (End-to-end Prediction Accuracy) [19].

It can be clearly observed that there still exists a significant performance gap in these metrics between utilizing a small amount of data and using complete data. This observation aligns with common sense in active learning tasks [56, 59, 65, 70], where a small sample size can not outperform the entire dataset in traditional image classification and segmentation tasks. We would like to offer some thoughts on this phenomenon.

- In the field of optimization with uncertain coefficients, recent studies [5, 11, 47] have also found that when a task involves both prediction and decision-making, with the prediction output serving as the input for the decision task, there can be a misalignment between the optimization objective of the prediction and the overall decision objective. In other words, better predictions do not necessarily lead to better decisions. For example, in the shortest path problem on a graph with unknown paths mentioned in [11], when the path costs are predicted using a dataset during the prediction phase and directly used for downstream solving, the solution obtained is not optimal.
- The enhanced visual perception capabilities afforded by larger datasets can be expected and align with common sense within the Active Learning community. A plausible explanation is that while driving trajectories might show a long-tail distribution, the repetition across different visual scenarios for vehicles is relatively low. Different environments, road sections, and lighting conditions inevitably lead to varied scenarios, making saturated training valuable. However, in end-to-end AD tasks, where these serve as auxiliary losses, our primary goal is decision-making, specifically trajectory planning. Thus, avoiding data redundancy and long-tail overfitting becomes even more critical.

It also raises the question of how to balance other losses in E2E-AD, considering planning as the ultimate objective, and whether there are better training paradigms. Our active learning approach provides a means to optimize training data while reducing costs. We believe that future work on multitask learning or hard case mining holds promise for enhancing planning performance.