Learnability-Driven Submodular Optimization for Active Roadside 3D Detection

Supplementary Material

A. Supplementary Material

Due to space limitations in the main manuscript, we provide additional theoretical proofs, detailed dataset specifications, and extensive experimental analyses in this supplementary material. The content is organized as follows:

- **Theoretical Analysis (Sec. A.1):** We provide the formal mathematical proofs regarding the monotonicity and submodularity of our proposed objective functions, guaranteeing the theoretical efficiency of the greedy optimization used in LH3D.
- **Dataset Details (Sec. A.5):** We provide detailed specifications for the primary evaluation dataset, **DAIR-V2X-I**, and the generalization dataset, **Rope3D**. We clarify the evaluation protocol, which employs the standard $AP_{3D|R40}$ **metric** across KITTI-style difficulty levels.
- **Failure Case Analysis (Sec. A.6):** We analyze typical failure modes, highlighting issues with long-range vehicles and occluded pedestrians/cyclists (fragmentation and misclassification). This analysis underscores the inherent limitations of monocular 3D estimation under extreme distance and visual ambiguity.
- **Validation of Hierarchical Stages (Sec. A.7):** We present in-depth discussions and empirical evidence (including visualization and metric analysis) to demonstrate the necessity and effectiveness of each individual stage in our three-stage learnability framework.
- **Ablation Studies on Annotation Budgets (Sec. A.8):** We report extended performance comparisons across a wider range of annotation budgets (from low-budget to high-budget regimes) to verify the robustness of LH3D.
- **Ablation Studies on Stages (Sec. A.9):** This section presents an ablation study of the stage design in LH3D. It examines how the ordering of the three stages, as well as the removal of individual components, affects overall performance.
- **Impact of (τ) and Joint Optimization (Sec. A.10):** We demonstrate the superiority of our staged strategy over joint optimization and determine that $\tau = 1$ optimally balances depth confidence with convergence stability.
- **Ablation at 50% Budget (Sec. A.11):** We show that LH3D surpasses the fully-supervised Oracle using merely half the annotation budget, as intelligent selection of learnable samples proves more effective than training on the full labeled set.
- **Generalization Experiments (Sec. A.12):** We extend our evaluation to the Rope3D dataset and test across different detector architectures (BEVSpread and BEVDet) to demonstrate the generalization ability of our method beyond a specific setup.
- **Computational Complexity (Sec. A.13):** We analyze the time complexity of our selection algorithm, showing that the computational overhead is negligible compared to the training costs.
- **Extended Analysis: Human Study (Sec. A.14):** We detail the controlled human study that isolates the impact of inherent ambiguity, empirically proving that ambiguous samples provide weaker supervision than learnable ones even with perfect ground truth.

A.1. Theoretical Analysis

In this section, we provide the formal proof that the objective functions proposed in our LH3D framework—specifically Φ_A (Depth Confidence), Φ_B (Semantic Balance), and Φ_C (Geometric Variation)—are monotone submodular. This property guarantees that the greedy optimization strategy employed in our multi-stage pipeline achieves a $(1 - 1/e)$-approximation of the optimal solution [3].

A.2. Definitions

Let $\mathcal{U}$ be the finite ground set of unlabeled images. A set function $F : 2^{\mathcal{U}} \to \mathbb{R}$ maps a subset $S \subseteq \mathcal{U}$ to a real value.

Definition 1 (Monotonicity). A set function F is monotone if for all subsets $A \subseteq B \subseteq \mathcal{U}$, it holds that $F(A) \leq F(B)$.

Definition 2 (Submodularity). A set function F is submodular if it satisfies the property of diminishing returns. Formally, for all $A \subseteq B \subseteq \mathcal{U}$ and any element $x \in \mathcal{U} \setminus B$:

$$F(A \cup \{x\}) - F(A) \geq F(B \cup \{x\}) - F(B). \quad (1)$$

A.3. Submodularity of Concave-Over-Modular Functions

Our learnability objectives are formulated using the *concave-over-modular* template defined in Eq. (**??**) of the main paper. We now prove that functions of this form are monotone submodular.

Theorem 1. Let $w_i \geq 0$ be a non-negative weight associated with each element $i \in \mathcal{U}$. Let $g(S) = \sum_{i \in S} w_i$ be a modular function, and let $\phi : \mathbb{R}_{\geq 0} \to \mathbb{R}$ be a non-decreasing, concave function. Then, the composite function $F(S) = \phi(g(S))$ is monotone submodular.

Proof. **Monotonicity:** Since $w_i \geq 0$, if $A \subseteq B$, then $g(A) \leq g(B)$. Because ϕ is non-decreasing, it follows that $\phi(g(A)) \leq \phi(g(B))$. Thus, $F(S)$ is monotone.

Submodularity: Let $A \subseteq B \subseteq \mathcal{U}$ and $x \in \mathcal{U} \setminus B$. Let $\Delta = w_x \geq 0$ be the weight of the new element. We

define the values of the modular function as $v_A = g(A)$ and $v_B = g(B)$. Since $A \subseteq B$ and weights are non-negative, we have $v_A \leq v_B$. The marginal gain of adding x to A is:

$$\Delta F(x \mid A) = \phi(v_A + \Delta) - \phi(v_A). \tag{2}$$

Similarly, the marginal gain for B is:

$$\Delta F(x \mid B) = \phi(v_B + \Delta) - \phi(v_B). \tag{3}$$

Since ϕ is a concave function, its gradients (or discrete increments) are non-increasing. Therefore, for $v_A \leq v_B$ and any increment $\Delta \geq 0$, the inequality

$$\phi(v_A + \Delta) - \phi(v_A) \geq \phi(v_B + \Delta) - \phi(v_B) \tag{4}$$

holds. This satisfies the definition of submodularity.

A.4. Application to LH3D Objectives

We apply Theorem 1 to the three stages of our framework.

Closure under Summation. We first note that a non-negative linear combination of submodular functions is also submodular. That is, if $F_1, \ldots, F_k$ are submodular, then $F(S) = \sum_k F_k(S)$ is submodular.

- **Stage 1: Depth-Confident Sample Selection (Eq. ??):** $\Phi_A(S) = \sum_{d=1}^{D} \log(\epsilon + Z_d(S))$. Here, $Z_d(S) = \sum_{i \in S} r_i m_{i,d}$ is a modular sum with non-negative weights $r_i m_{i,d}$. The function $\phi(z) = \log(\epsilon + z)$ is concave and non-decreasing for $z \geq 0$ (given $\epsilon > 0$). Thus, each term is submodular, and their sum Φ_A is submodular.
- **Stage 2: Rare-Common Class Balancing (Eq. ??):** $\Phi_B(S) = \sum_{c \in \mathcal{C}} \log(\epsilon + N_c(S))$. Similarly, $N_c(S) = \sum_{i \in S} \alpha_i p_i(c)$ is a modular coverage term. By the same logic, Φ_B is a sum of concave-over-modular functions and is therefore submodular.
- **Stage 3: Geometric Variant Selection (Eq. ??):** $\Phi_C(S) = \sum_{c \in \mathcal{C}} \log(\epsilon + U_c(S))$. With $U_c(S) = \sum_{i \in S} s_{i,c}$ being modular (sum of novelty scores), Φ_C is also submodular.

Conclusion: All three components of our objective function satisfy monotonicity and submodularity. Consequently, the greedy algorithm used in LH3D is theoretically guaranteed to find a solution within $(1 - 1/e)$ of the optimum at each stage.

A.5. Datasets

DAIR-V2X [11] is a large-scale benchmark for vehicle–infrastructure cooperative autonomous driving, offering a rich multi-modal 3D object detection resource. Following prior work [7, 9], we focus on the DAIR-V2X-I subset, which comprises approximately 10k images captured from infrastructure-mounted cameras to study roadside perception. The subset includes 493k 3D bounding box annotations spanning distances from 0 to 200 meter. We adopt the standard data split of 50%, 20%, and 30% for training, validation, and testing, respectively. As the official test annotations are not yet released, all evaluations are conducted on the validation set.

Rope3D [10] is another benchmark for roadside 3D object detection. It comprises 50 k images and over 1.5 M 3D object annotations captured under diverse conditions, including varying lighting (day, night, dusk) and weather (rainy, sunny, cloudy) across 26 distinct intersections, with object distances ranging from 0 m to 200 m. Following the split strategy introduced in Rope3D, we use 70% of the images for training and 30% for testing.

For validation metrics, we leverage $AP_{3D|R40}$ metric to evaluate 3D bounding boxes. The results are reported in three difficulty levels—Easy, Moderate, and Hard—based on box characteristics, following the KITTI [2] evaluation protocol.

A.6. Failure Cases

Despite LH2D's strong performance over baseline methods, our approach still encounters failure cases in challenging roadside scenarios, particularly for distant vehicles and for pedestrians or cyclists that are occluded.

Fig. 1 highlights two primary failure modes: ***distance*** and ***occlusion***. First, long-range objects often lack sufficient visual detail for reliable 3D estimation, leading to missed detections of small or distant vehicles (*top examples*). Cyclists also pose a challenge, as they are easily misclassified in crowded environments.

Second, the model struggles with severe occlusion (*bottom examples*). When vehicles heavily overlap, LH3D frequently fails to distinguish the object in the rear. This issue extends to vulnerable road users; for instance, the visualization shows a cyclist largely screened by a vehicle, resulting in a missed detection due to the lack of visible features.

A.7. Validation of Hierarchical Stages

A.7.1. Stage 1: Depth-Confident Sample Selection

Stage 1 aims to filter out inherently ambiguous scenes by selecting images where the depth estimator exhibits high confidence and balanced depth coverage. At the beginning of active learning, however, the detector is trained on only a very small labeled subset, so its depth predictions remain reliable only on relatively simple, low-ambiguity scenes. As a result, Stage 1 naturally gravitates toward such scenes in the early rounds. These early-selected images typically contain fewer objects, involve minimal occlusion, and show a more even spread of near- and mid-range targets. In practical terms, the selected scenes also tend to have a lower density of vehicles, pedestrians, and cyclists, which prevents the annotation process from spending its limited early budget on congested or difficult scenes that the model is not yet strong enough to learn from.

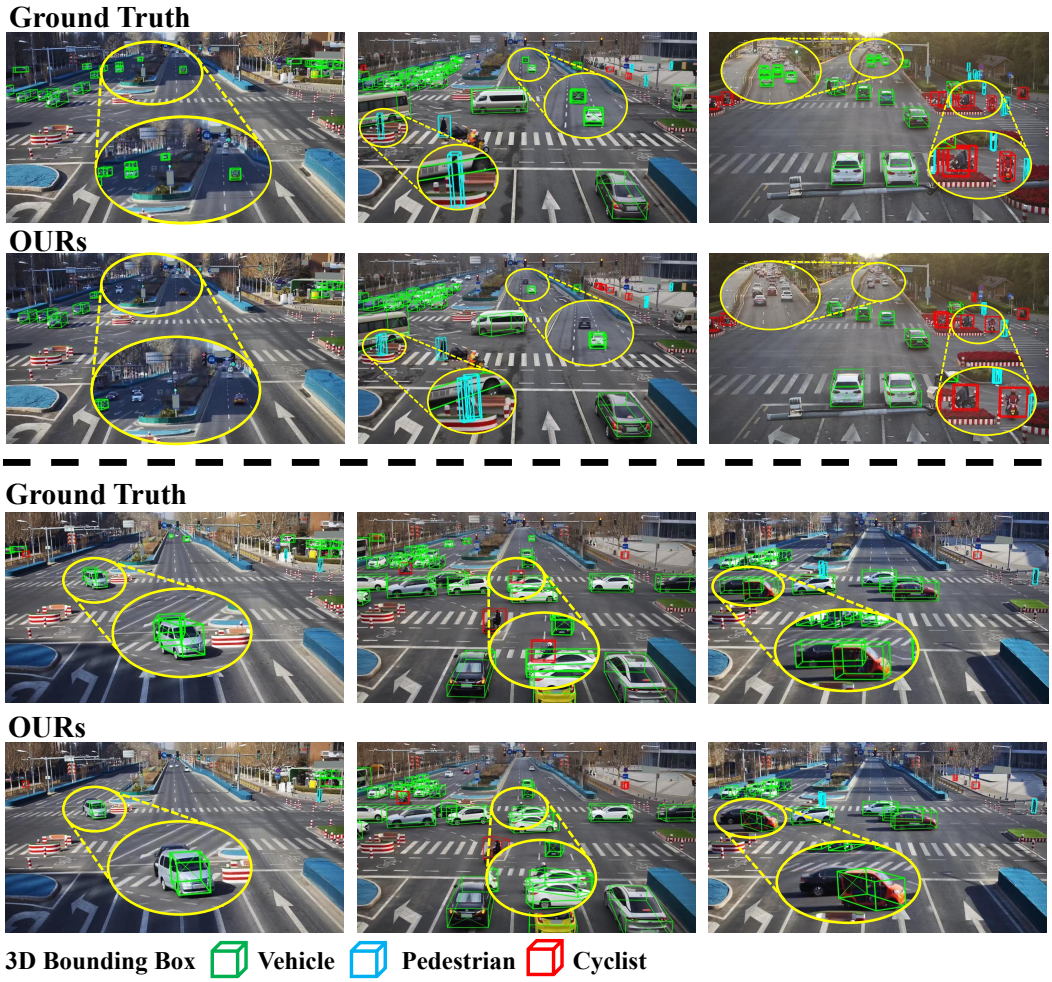

Figure 1. **Failure cases of LH3D on the DAIR-V2X-I validation set: Ground-truth annotations vs. LH3D predictions.** The top row mainly illustrates long-range perception, where distant vehicles provide limited visual cues, leading to missed detections or unstable 3D localization. The bottom row shows failures caused by occlusion, where overlapping objects hinder geometric reasoning and result in incomplete predictions.

As the active learning process progresses, the detector becomes increasingly capable of producing confident depth estimates on more complex layouts. Stage 1 correspondingly begins to admit scenes with richer object arrangements, heavier occlusion, and greater geometric variability. At the same time, the balanced-depth-coverage criterion avoids repeatedly sampling near-range scenes: once these bins are sufficiently covered, the objective encourages selecting images that contribute to underrepresented mid- and far-range regions.

Empirically, this behavior is clearly reflected on the DAIR-V2X-I [11] dataset: LH3D consistently selects scenes with systematically closer and more learnable object configurations. The average distance from annotated objects to the camera is **6.84 m** under LH3D, whereas uncertainty-based and diversity-based baselines select scenes whose average distance consistently exceeds **7.5 m**. This demonstrates that LH3D not only favors depth-confident, easy-to-learn scenes in early rounds but also preserves annotation budget by postponing dense or ambiguous scenes until the model becomes sufficiently strong to extract reliable supervision from them.

A.7.2. Stage 2: Rare-Common Class Balancing

The core function of Stage 2 is to ensure that the selected annotation set maintains high class diversity across multiple active learning rounds, especially under the severe imbalance in roadside 3D datasets (e.g., vehicles vastly outnumber pedestrians and cyclists).

To validate the necessity of Stage 2, we perform an ablation study comparing the full LH3D pipeline against LH3D w/o Stage 2. As shown in Fig. **??**, we track the global class-diversity entropy over 8 active learning rounds (higher en-

Figure 2. **Training samples from DAIR-V2X-I selected during LH3D Stage 1 through depth-confident sample selection.** The top row displays the original images, and the bottom row shows the corresponding 3D bounding box annotations. The samples selected during Stage 1 are characterized by high visual clarity and minimal ambiguity. The selection strategy prioritizes scenes where vehicles, pedestrians, and cyclists appear without occlusion and are positioned at moderate distances. Furthermore, these samples exhibit low scene complexity, avoiding overcrowded traffic environments.

tropy indicates better class balance).

Initially, removing Stage 2 (blue curve) results in slightly higher entropy than the full LH3D pipeline (red curve), peaking around Round 3 (0.862). This surge occurs because, without explicit balancing constraints, the baseline aggressively selects available rare classes (Pedestrians and Cyclists) from the unlabeled pool, leading to a temporary increase in diversity.

However, this high diversity is unsustainable. Since naturally rare classes are quickly depleted in early rounds, the LH3D w/o Stage 2 variant is forced to select mostly common classes (Vehicles) in later rounds, causing entropy to drop significantly (down to ≈ 0.835 by Round 8).

In contrast, the full LH3D pipeline (with Stage 2) enforces a controlled, stable selection across classes. Although its diversity gain is more gradual at the beginning, it maintains high and stable class diversity throughout the process, stabilizing at an entropy of ≈ 0.845 in later rounds. This consistent balancing prevents the selected set from becoming overly biased toward Vehicles and ultimately leads to better final detection performance: incorporating Stage 2 improves $AP_{3D|R40}$ by more than 4 points for Cyclists and 2 points for Pedestrians compared to LH3D w/o Stage 2.

A.7.3. Stage 3: Geometric Variant Selection

Stage 3 is designed to enhance per-class geometric *variation* while still respecting the detector's learned geometric priors, i.e., to select scenes that are novel but not extreme outliers in BEV layout space.

To test this design choice, we construct an ablated variant that inverts the first step of Stage 3: instead of favoring scenes whose BEV layouts are moderately consistent with the labeled set, it explicitly prioritizes scenes whose geometry is as *dissimilar* as possible from previously labeled scenes. In other words, we remove the geometric consistency constraint and aggressively push selection toward maximal geometric novelty.

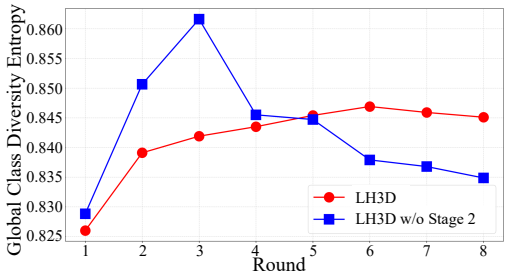

Figure 3. Global class-diversity entropy over AL rounds: comparison between LH3D and LH3D without Stage 2.

On DAIR-V2X-I (Hard setting) with a BEVHeight backbone and the same total annotation budget, this "maximally dissimilar" variant yields substantially worse performance: the final $AP_{3D|R40}$ averaged over Car, Pedestrian, and Cyclist is lower by about 4 percent compared to the full LH3D with Stage 3. Qualitatively, the ablated variant tends to oversample rare, highly anomalous layouts in early rounds, which slows down training and introduces instability, as the model struggles to extract reliable supervision from overly difficult scenes. In contrast, the original Stage 3, which encourages *controlled* geometric deviation, maintains stable training dynamics and consistently achieves higher final detection accuracy.

A.8. Ablation Studies on Annotation Budgets

We conduct extensive ablation studies to evaluate the effectiveness of LH3D under varying annotation budgets on the DAIR-V2X-I dataset. The total available training pool con-

tains $246,500$ objects (50% of the total 493k annotations). The annotation budget is defined by the cumulative number of objects annotated. The budgets presented in Table 1 reflect the following proportions of the total training pool:

- 8,000 objects: $\approx 3.24\%$ of the training pool.
- 16,000 objects: $\approx 6.49\%$ of the training pool.
- 24,000 objects: $\approx 9.74\%$ of the training pool.
- 32,000 objects: $\approx 12.98\%$ of the training pool.

Based on these findings, we chose the 32,000 object budget ($\approx$ 13% of the training pool) as the primary comparative budget in the main text. At this level, our method, LH3D, achieves 86.06%, 67.32%, and 78.67% of full-performance for vehicles, pedestrians, and cyclists respectively, significantly outperforming baselines and confirming that learnability, not uncertainty, matters for roadside 3D perception.

A.9. Ablation Studies on Stages

LH3D follows a stage ordering of depth confidence (DC), semantic balance (SB), and geometric variation (GV). **Stage ordering.** Table 2 shows that DC–SB–GV design outperforms all 3! possible permutations. Notably, shifting DC from the first stage or placing GV at the beginning consistently leads to worse results, validating the effectiveness of our prioritized design.

Stage removal. As shown in Table 3, retaining only two stages consistently leads to inferior performance compared with the full three-stage pipeline, demonstrating that all three components are necessary.

A.10. Ablation on τ and Joint Optimization

Our staged strategy respects intrinsic dependencies: depth confidence (Stage 1) is a prerequisite for monocular BEV, followed by semantic balancing (Stage 2) which prevents geometric variation (Stage 3) from overfitting frequent layouts. We validate this design against two alternatives: joint optimization and varying the entropy weight decay τ. **Entropy weight decay**. We set $\tau = 1$ as the default and evaluate $\tau \in \{0.3, 0.6, 1.2\}$ (Tab. 4). Values that are too small (e.g., 0.3) overly suppress entropy and slow convergence, whereas values too large (e.g., 1.2) dilute the depth-confidence signal, leading to noisy initial rankings. $\tau = 1$ achieves the optimal balance between maintaining sharp depth confidence and ensuring stable convergence. **Joint optimization.** We also compare against a joint training baseline that simultaneously optimizes all three objectives (Tab. 4). While theoretically appealing, joint optimization performs inferiorly because it treats depth estimation, semantic balancing, and geometric diversity as equally important—whereas our stages explicitly model their prerequisite relationships. Moreover, joint training exhibits slower convergence, making it less practical for large-scale roadside datasets.

A.11. Ablations at 50% Budget

Table 5 further demonstrates that LH3D, using the BEVHeight backbone, surpasses the fully-supervised Oracle with merely $\sim$50% of the labeled data (120K out of 246.5K images). This seemingly counter-intuitive result highlights a critical insight: not all annotated samples are beneficial for training. By actively filtering out ambiguous or non-learnable roadside scenes that introduce noisy gradients, LH3D effectively trains on a higher-quality subset.

Consequently, LH3D achieves 46.95% AP, outperforming the Oracle's 41.46% by 5.49 percentage points (a relative gain of 13.2%). The improvements are particularly pronounced for safety-critical minority classes: pedestrian detection increases by 5.02% (from 21.11% to 26.13%) and cyclist detection by 9.40% (from 42.09% to 51.49%), while vehicle detection also sees a modest gain of 2.04%.

A.12. Generalization Experiments

Table 6 demonstrates the effectiveness of our proposed LH3D method on the Rope3D dataset. Across both *BEVSpread* and *BEVDet* backbones, LH3D consistently surpasses state-of-the-art active learning baselines. In particular, under the *BEVSpread* configuration, our method outperforms the PPAL baseline by more than 2.6 points in the *Easy Vehicle* category. For the *Cyclist* category on the *Hard* difficulty setting (using the BEVSpread backbone), LH3D achieves an AP_{3D} of 17.65. This represents a substantial improvement of +3.61 points over the nearest competitor, PPAL (14.04), highlighting our model's effectiveness in mitigating the ambiguities often associated with vulnerable road users.

We observe that the performance on the Rope3D dataset is lower compared to DAIR-V2X-I. This discrepancy can be attributed to the higher complexity of the Rope3D scenarios and the limited scale of the validation set ($1,688$ images), which poses a greater challenge for the model under the current active learning constraints. In future work, we will increase the annotation budget to select a larger number of informative samples for training, thereby further improving the model's generalization capability.

A.13. Computational Complexity

We evaluate the computational efficiency of our proposed approach by comparing the training duration against several baseline methods. Table 7 presents the training time comparison on the DAIR-V2X-I dataset with the *BEVHeight* backbone.

While the RANDOM strategy achieves the lowest training time (3.55 hours) due to its lack of selection overhead, our method, LH3D, maintains competitive efficiency. With a total training time of 4.47 hours, LH3D proves to be more efficient than both PPAL (4.70 hours).

Table 1. $AP_{3D|R40}$ **performance on the DAIR-V2X-I validation set** under different annotation budgets. The backbone detector is *BEVHeight*.

	Vehicle (*IoU=0.5*)			Pedestrian (*IoU=0.25*)			Cyclist (*IoU=0.25*)			Average		
Method (Object Budget)	Easy	Mod.	Hard	Easy	Mod.	Hard	Easy	Mod.	Hard	Easy	Mod.	Hard
RANDOM (8000)	51.95	43.94	43.91	14.01	13.28	13.45	23.42	30.56	31.07	29.79	29.26	29.48
ENTROPY (8000)	51.85	42.56	42.64	**14.96**	**14.13**	**14.27**	22.74	33.60	34.07	29.85	30.10	30.33
PPAL (8000)	50.18	42.40	42.49	13.00	12.19	12.28	**26.73**	**39.33**	**39.47**	**29.97**	**31.31**	**31.41**
LH3D (8000)	**57.47**	**49.85**	**49.93**	10.90	10.72	10.83	21.10	32.43	32.41	29.82	31.00	31.06
RANDOM (16000)	56.46	46.63	46.61	12.06	11.30	11.38	23.03	31.40	31.89	30.52	29.78	29.96
ENTROPY (16000)	59.24	50.59	50.67	14.41	13.63	13.80	24.82	32.30	32.72	32.82	32.17	32.40
PPAL (16000)	60.07	51.23	51.28	13.61	12.81	12.90	**28.99**	**36.61**	**37.03**	34.22	33.55	33.74
LH3D (16000)	**63.03**	**52.93**	**52.41**	**15.81**	**14.84**	**14.94**	27.64	33.64	34.30	**35.49**	**33.80**	**33.88**
RANDOM (24000)	57.36	48.88	48.97	13.50	12.98	12.97	26.66	**37.36**	**37.53**	32.51	33.07	33.16
ENTROPY (24000)	57.44	48.99	49.14	13.58	12.73	12.83	27.68	34.27	34.69	32.90	32.00	32.22
PPAL (24000)	58.31	48.14	48.27	11.98	11.69	11.83	29.01	35.35	35.64	33.10	31.73	31.91
LH3D (24000)	**63.55**	**53.71**	**52.78**	**16.58**	**15.65**	**15.82**	**29.93**	36.95	37.33	**36.69**	**35.44**	**35.31**
RANDOM (32000)	61.90	51.37	51.41	13.63	13.23	13.42	30.04	38.70	39.38	35.19	34.43	34.74
ENTROPY (32000)	63.42	54.42	54.51	17.50	16.57	16.72	31.45	36.86	38.57	37.46	36.67	36.53
PPAL (32000)	60.20	51.38	51.44	**19.09**	**18.47**	**18.07**	**34.41**	39.13	39.71	37.90	36.33	36.41
LH3D (32000)	**65.36**	**56.00**	**56.03**	18.51	17.50	17.67	32.44	**41.49**	**41.79**	**38.77**	**38.33**	**38.50**
ORACLE (246500)	73.05	61.32	61.19	22.10	21.57	21.11	42.85	42.26	42.09	46.00	41.72	41.46

Table 2. Ablation study on stage ordering of our LH3D framework using the BEVHeight backbone (Hard setting). DC = Depth Confidence, SB = Semantic Balance, GV = Geometric Variation.

Order	**Car**	**Pedestrian**	**Cyclist**	**Average**
DC–GV–SB	50.62	16.83	37.10	34.85
SB–GV–DC	51.82	15.14	36.81	34.59
SB–DC–GV	55.90	12.46	35.95	34.77
GV–DC–SB	40.04	13.02	32.67	28.58
GV–SB–DC	47.31	15.16	36.63	33.03
Ours (DC–SB–GV)	**56.03**	**17.67**	**41.79**	**38.50**

Table 3. LH3D stage-removal ablation with BEVHeight (Hard). DC = Depth Confidence, SB = Semantic Balance, GV = Geometric Variation.

Order	**Car**	**Pedestrian**	**Cyclist**	**Average**
SB–GV	51.17	17.34	34.93	34.48
DC–GV	55.58	17.16	35.22	35.99
DC–SB	54.27	16.58	37.63	36.16
Ours (DC–SB–GV)	**56.03**	**17.67**	**41.79**	**38.50**

Table 4. DAIR-V2X-I ablation: impact of Stage-1 entropy weight decay (τ) and joint optimization (BEVHeight, Hard).

Method	**Car**	**Pedestrian**	**Cyclist**	**Average**
LH3D ($\tau = 0.3$)	55.21	17.33	36.77	36.44
LH3D ($\tau = 0.6$)	47.80	16.68	41.47	35.32
LH3D ($\tau = 1.2$)	50.73	14.31	38.33	34.46
Joint Optimization	51.12	14.34	34.68	33.38
Ours (LH3D $\tau = 1.0$)	**56.03**	**17.67**	**41.79**	**38.50**

Table 5. LH3D outperforms the fully supervised *Oracle* using ∼50% annotation budget on DAIR-V2X-I (Hard, BEVHeight).

Order	**Car**	**Pedestrian**	**Cyclist**	**Average**
LH3D (120K)	**63.23**	**26.13**	**51.49**	**46.95**
Oracle (246.5K)	61.19	21.11	42.09	41.46

A.14. Extended Analysis: Human Study on Ambiguity

To empirically validate our hypothesis that *inherently ambiguous samples* provide weaker supervision signals than learnable samples—even when accurate ground truth is provided—we conducted a controlled human study. This study isolates the impact of visual ambiguity from other factors like class imbalance or label noise.

A.14.1. Study Setup and Partitioning

We enlisted three expert annotators (well-trained PhD students in the computer vision domain) to manually partition the unlabeled training pool into two distinct subsets: **Learnable** and **Ambiguous**. The classification was based on three primary visual criteria strictly from a monocular perspective:

- **Object Distance:** Scenes dominated by objects at extreme ranges (e.g., > 55m) where objects have lower resolution than closer objects.
- **Occlusion Level:** Scenes where key objects suffer from severe occlusion ($> 70\%$) or are truncated.
- **Scene Clutter:** High-density scenes where object boundaries are visually indistinguishable.

To ensure a fair comparison, the annotators strictly controlled the selection to maintain a consistent class distribution (Car, Pedestrian, Cyclist) between the two subsets, eliminating semantic imbalance as a confounding variable.

A.14.2. Experimental Protocol

We designed an iterative training protocol to mimic the active learning process, but with manual selection:

Table 6. $AP_{3D|R40}$ **results on the Rope3D validation set** with 20% queried boxes. Backbones include *BEVSpread* and *BEVDet*.

Backbone	Method	Vehicle (IoU=0.5)			Pedestrian (IoU=0.25)			Cyclist (IoU=0.25)			Average		
		Easy	Mod.	Hard	Easy	Mod.	Hard	Easy	Mod.	Hard	Easy	Mod.	Hard
BEVSpread	RANDOM	24.49	22.52	22.41	1.90	1.83	1.86	8.85	12.03	12.01	11.75	12.13	12.09
	ENTROPY	24.72	20.63	20.52	0.80	0.81	0.83	7.46	10.76	10.77	10.33	10.73	11.37
	BGADL [6]	24.13	23.89	22.13	1.13	1.05	1.39	13.23	12.00	12.11	12.83	12.31	11.88
	CORESET [5]	24.61	22.62	21.41	1.38	1.18	1.20	9.04	12.20	12.17	11.68	12.00	11.59
	BADGE [1]	24.70	24.13	23.89	1.28	1.10	1.04	12.74	13.99	13.10	12.91	13.07	12.67
	PPAL [8]	28.19	25.47	24.03	2.47	2.56	2.62	11.14	14.03	14.04	13.93	14.02	13.56
	HUA [4]	23.99	22.17	20.87	1.96	1.86	1.89	7.91	11.65	11.63	11.29	11.89	11.46
	LH3D (Ours)	**30.85**	**26.75**	**26.60**	**2.53**	**2.53**	**2.57**	**14.30**	**17.69**	**17.65**	**15.89**	**15.66**	**15.61**
BEVDet	RANDOM	23.50	21.73	21.02	1.57	1.71	1.72	7.58	13.18	13.27	10.88	12.21	11.99
	ENTROPY	25.84	22.64	22.62	1.12	1.16	1.18	9.46	13.85	13.07	12.14	12.55	12.29
	BGADL [6]	23.25	20.17	20.71	1.02	1.01	1.08	9.41	11.66	11.47	11.23	10.95	11.09
	CORESET [5]	26.26	22.61	22.59	1.80	1.68	1.72	10.63	14.13	14.21	12.90	12.81	12.84
	BADGE [1]	24.77	22.70	21.03	1.81	1.61	1.90	**11.72**	12.63	12.28	12.77	12.31	11.74
	PPAL [8]	21.53	20.29	18.96	1.78	1.78	1.80	7.20	10.43	10.43	10.17	10.83	10.40
	HUA [4]	21.39	20.29	20.22	1.26	1.15	1.16	6.81	10.36	10.39	9.82	10.60	10.59
	LH3D (Ours)	**28.19**	**26.09**	**25.97**	**1.78**	**1.84**	**1.90**	11.59	**16.37**	**16.44**	**13.85**	**14.76**	**14.77**

Table 7. **Training Time Comparison on DAIR-V2X-I** using the *BEVHeight* backbone.

Method	Time (*hours*)
RANDOM	3.55
ENTROPY	4.40
PPAL	4.70
HUA	4.08
LH3D (Ours)	4.47

- **Budget Constraints:** The total annotation budget was fixed at $10,000$ objects.
- **Iterative Selection:** The process spanned 10 rounds. In each round, annotators selected up to 50 images from their respective pools (Learnable vs. Ambiguous) to add to the training set.
- **Training Settings:** The model was trained for 10 epochs per round. To simulate a realistic active learning cycle, the model for round k was initialized with the weights from round $k-1$ (incremental learning).
- **Labeling:** Both groups were trained using the official Ground Truth labels from the dataset.

A.14.3. Results and Discussion

The results, visualized in Fig. **??** of the main paper, reveal a critical finding:

Ambiguity Limits Monocular Learnability. Despite using the exact same detector architecture, optimizer, and reliable ground truth labels, the model trained on the *Ambiguous* split consistently underperformed the model trained on the *Learnable* split. Specifically, the performance gap is most pronounced for Vehicles and Pedestrians. This indicates that ambiguous samples suffer from low signal-to-noise ratios; even with correct labels, the image features (due to blur or occlusion) are insufficient for the network to learn a generalized geometric mapping.

Implication for Active Learning. This experiment confirms that in the roadside monocular setting, *uncertainty* is not equivalent to *informativeness*. High-uncertainty samples in this domain are often inherently ambiguous cases that confuse the model rather than strengthen it. This validates the core motivation of LH3D: prioritizing learnability over mere uncertainty.

References

[1] Jordan T Ash, Chicheng Zhang, Akshay Krishnamurthy, John Langford, and Alekh Agarwal. Deep batch active learning by diverse, uncertain gradient lower bounds. *arXiv preprint arXiv:1906.03671*, 2019. 7

[2] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the KITTI vision benchmark suite. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3354–3361. IEEE, 2012. 2

[3] George L. Nemhauser, Laurence A. Wolsey, and Marshall L. Fisher. An analysis of approximations for maximizing submodular set functions—I. *Mathematical Programming*, 14 (1):265–294, 1978. 1

[4] Younghyun Park, Wonjeong Choi, Soyeong Kim, Dong-Jun Han, and Jaekyun Moon. Active learning for object detection with evidential deep learning and hierarchical uncertainty aggregation. In *The Eleventh International Conference on Learning Representations*, 2023. 7

[5] Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. *arXiv preprint arXiv:1708.00489*, 2017. 7

[6] Toan Tran, Thanh-Toan Do, Ian Reid, and Gustavo Carneiro. Bayesian generative active deep learning. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, pages 6295–6304. PMLR, 2019. 7

[7] Wenjie Wang, Yehao Lu, Guangcong Zheng, Shuigen Zhan, Xiaoqing Ye, Zichang Tan, Jingdong Wang, Gaoang Wang, and Xi Li. BEVSpread: Spread voxel pooling for bird's-eye-view representation in vision-based roadside 3D object detection. In *Proceedings of the IEEE/CVF Conference on*

Computer Vision and Pattern Recognition (CVPR), pages 14718–14727, 2024. 2

[8] Chenhongyi Yang, Lichao Huang, and Elliot J Crowley. Plug and play active learning for object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 17784–17793, 2024. 7

[9] Lei Yang, Kaicheng Yu, Tao Tang, Jun Li, Kun Yuan, Li Wang, Xinyu Zhang, and Peng Chen. BEVHeight: A robust framework for vision-based roadside 3D object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21611–21620, 2023. 2

[10] Xiaoqing Ye, Mao Shu, Hanyu Li, Yifeng Shi, Yingying Li, Guangjie Wang, Xiao Tan, and Errui Ding. Rope3d: The roadside perception dataset for autonomous driving and monocular 3d object detection task. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21341–21350, 2022. 2

[11] Haibao Yu, Yizhen Luo, Mao Shu, Yiyi Huo, Zebang Yang, Yifeng Shi, Zhenglong Guo, Hanyu Li, Xing Hu, Jirui Yuan, and Zaiqing Nie. DAIR-V2X: A large-scale dataset for vehicle-infrastructure cooperative 3D object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21361–21370, 2022. 2, 3