UBD-BENCH: BENCHMARKING AND ANALYZING UNIMODAL BRANCH DEGRADATION IN MULTIMODAL PERCEPTION FOR AUTONOMOUS DRIVING

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal perception improves autonomous-driving performance by combining complementary sensor information. However, stronger multimodal predictions do not necessarily imply stronger representations in each modality-specific branch. Such branch-level weaknesses can remain hidden in conventional output-level evaluation. We introduce UBD-Bench, a systematic benchmark for studying unimodal branch degradation (UBD) in multimodal perception. UBD-Bench uses a Paired Frozen-Branch Probe to compare the task-relevant representation quality of the same modality after independent unimodal training and multimodal joint training under matched decoding conditions. Across four autonomous-driving perception tasks and 10 models, we find that Camera branches consistently degrade after multimodal joint training, whereas LiDAR branches exhibit weaker and more model-dependent changes, revealing a pronounced modality asymmetry. Hierarchical probing further shows that Camera-branch degradation extends from the visual backbone through the Camera-to-BEV pathway and is most severe at the pre-fusion Camera BEV representation. Guided by this diagnosis, we introduce unimodal-to-multimodal distillation at the degraded representation. This intervention consistently improves both Camera-branch Probe performance and full-model performance across all four tasks. These results show that multimodal gains can coexist with degraded unimodal representations, highlighting branch-level representation quality as an important dimension for evaluating and improving multimodal perception.

1 INTRODUCTION

Accurate and reliable perception of the surrounding environment is fundamental to safe autonomous driving. To this end, camera–LiDAR fusion has become a prevailing paradigm for autonomous-driving perception, as cameras capture dense appearance and semantic cues while LiDAR provides direct and metrically accurate 3D geometry (Caesar et al., 2020; Yan et al., 2023). Building on this complementarity, multimodal systems have achieved substantial progress across a broad range of autonomous-driving perception tasks, including 3D object detection (Liu et al., 2023), online vectorized map construction (Liao et al., 2022), bird's-eye-view (BEV) map segmentation (Liu et al., 2023), and occupancy prediction (Wang et al., 2023a). Despite substantial differences in model architecture and prediction target, these systems are commonly evaluated using task-level metrics computed from their final predictions, enabling comparisons of overall perception performance.

A natural expectation of multimodal perception is that combining complementary sensor observations should produce more accurate and reliable predictions than relying on either sensor alone. However, Figure 1 shows that this overall advantage does not hold uniformly across scenarios: although the Camera–LiDAR system achieves stronger overall performance, the Camera-only system produces more accurate predictions in certain scenarios.

These examples do not imply that single-sensor perception is generally superior. Instead, they raise a fundamental question: does each modality-specific branch in a jointly trained multimodal model learn a task-relevant representation as effective as that learned by an independently trained unimodal model? Since the two systems differ in both their sensor inputs and processing pipelines, an output-level comparison cannot answer this question or isolate the influence of joint learning on any individual modality. Current evaluations primarily report metrics based on these predictions, providing limited insight into the quality of task-relevant representations learned by individual branches of multimodal models relative to independently trained unimodal models.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

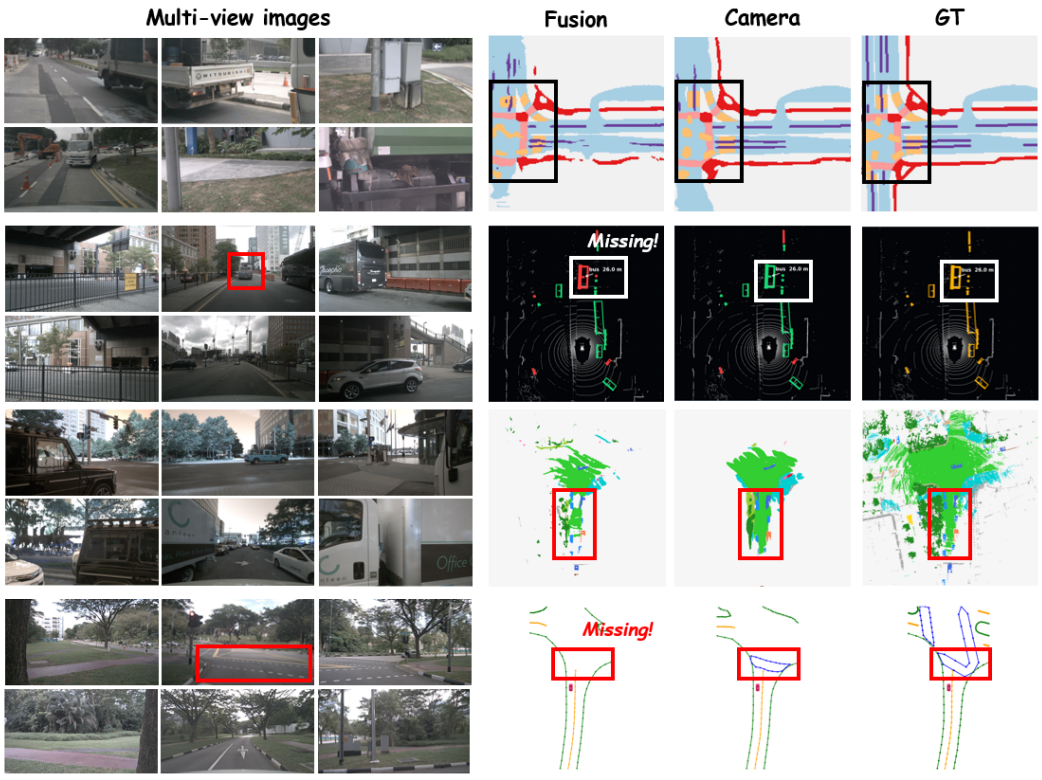


Figure 1: **Prediction Discrepancies Between Camera–LiDAR and Camera-Only Models.** Representative examples from 3D object detection, BEV map segmentation, occupancy prediction, and online vectorized map construction show that Camera–LiDAR models produce less accurate predictions than their Camera-only counterparts in the highlighted regions.

To address this gap, we introduce **UBD-Bench**, a benchmark that uses a **Paired Frozen-Branch Probe** to assess the task-relevant representational quality of the same modality after independent unimodal training and multimodal joint training. We define **unimodal branch degradation (UBD)** as lower probe performance for the jointly trained branch than for its independently trained counterpart. UBD-Bench then examines how widely UBD occurs across tasks and architectures, where it arises along modality-specific representation pathways, and whether targeted interventions mitigate it.Our key contributions are threefold:

- **UBD-Bench establishes the first systematic benchmark for unimodal branch degradation in multimodal autonomous-driving perception.** Built on the Paired Frozen-Branch Probe, UBD-Bench spans four perception tasks and 10 models, demonstrating the prevalence of Camera-branch degradation across tasks and models, as well as a pronounced modality asymmetry between Camera and LiDAR.

- **A hierarchical Probe is developed to localize degradation along the Camera–BEV generation pathway.** Cross-task results show that the degradation is already embedded in the branch representation before the task decoder and involves the entire pathway from image encoding to BEV generation, with the most pronounced gap appearing at the pre-fusion Camera BEV representation.

- **Guided by the hierarchical Probe results, a unimodal-to-multimodal reverse-distillation scheme is designed to intervene in the degraded branch.** Transferring knowledge from an independently trained Camera teacher to the corresponding branch of the multimodal model improves both the task-relevant representation quality measured by the branch Probe and full-model performance, providing intervention-based evidence for the degradation diagnosis and demonstrating that targeted representation enhancement can mitigate unimodal branch degradation.

2 Related Work

Multimodal Perception for Autonomous Driving. Multimodal perception for autonomous driving combines complementary information to perceive the surrounding environment, with camera–LiDAR fusion serving as a prominent example. Recent research has advanced attention-based interaction between heterogeneous sensor features (Yan et al., 2023) and shared BEV fusion for 3D detection and map segmentation (Liu et al., 2023), with related developments extending multimodal perception to online map construction (Zhang et al., 2024; Hao et al., 2024a; Zhou et al., 2024) and occupancy prediction (Ming et al., 2024; Shi et al., 2024; Li et al., 2025). These studies span diverse fusion mechanisms and perception tasks. UBD-Bench draws on this diversity to assess whether unimodal branch degradation recurs across tasks and model architectures.

Evaluation of Multimodal Perception for Autonomous Driving. Evaluation of multimodal perception for autonomous driving examines whether perception models remain accurate and reliable across different tasks, operating conditions, and sensor configurations (Song et al., 2024). Recent benchmarks evaluate natural corruptions in 3D detection (Dong et al., 2023), corruption robustness and domain shifts in BEV perception (Xie et al., 2025), partial or complete sensor failures in detection and map segmentation (Ge et al., 2023), and combined Camera–LiDAR corruptions in online map construction (Hao et al., 2024b). These evaluations reveal how changes in sensing conditions affect final predictions, but do not directly isolate the effect of multimodal joint training on the task-relevant representation quality of each modality-specific branch under uncorrupted inputs. UBD-Bench complements existing evaluations by using a unified frozen-branch probing protocol to compare representations learned through independent unimodal training and multimodal joint training.

Knowledge Distillation for Autonomous-Driving Perception. Knowledge distillation in autonomous-driving perception typically transfers task-relevant knowledge from a stronger teacher model to a student model to improve the student's perception performance (Yan et al., 2026). Recent studies have transferred geometric knowledge from LiDAR teachers to Camera detectors (Wang et al., 2023b; Li et al., 2023), supported knowledge transfer among different sensor configurations through aligned BEV representations (Zhou et al., 2023; Zhao et al., 2024), distilled multimodal knowledge into lightweight Camera map models (Hao et al., 2024a), and used future observations to supervise models operating on current-frame inputs (Li et al., 2026). Their improvements show that prediction-, feature-, and relation-level constraints can preserve useful geometric and semantic structure in representations that would otherwise remain comparatively weak. Building on this finding, UBD-Bench transfers knowledge from independently trained unimodal teachers to the corresponding branches during multimodal joint training, and evaluates the intervention through branch Probe performance and fusion performance.

3 Analysis

3.1 Evaluation Settings and Unified Protocol

Paired Frozen-Branch Probe. A typical Camera–LiDAR perception model comprises a Camera–BEV generation pathway, a LiDAR–BEV generation pathway, a cross-modal fusion module, and a task decoder (Liu et al., 2023; Wang et al., 2023a). The final task metric reflects the combined effects of the fused representation and the task decoder and therefore does not directly measure the task-relevant representation quality of either modality. To enable branch-level evaluation, UBD-Bench introduces the **Paired Frozen-Branch Probe**. The Probe compares representations of the same modality learned through independent unimodal training and multimodal joint training under matched decoding conditions.

For a given probing position, the network from the raw sensor input to that position is defined as the representation extractor g. A paired Probe instantiates two such extractors. One extractor loads its weights from a fully trained unimodal model, whereas the other loads the corresponding branch weights from a fully trained multimodal model. Throughout the Probe, both extractors remain frozen and are each connected to a separate task decoder, with the two decoders having the same architecture and identical initial weights. Only the decoders are trained, using identical data splits, optimization settings, and training budgets. The resulting Probe models are evaluated under the same procedure. Consequently, the source of the frozen representation extractor is the only variable within each paired comparison. Figure 2 illustrates the weight loading and probing procedure for the Camera and LiDAR branches.

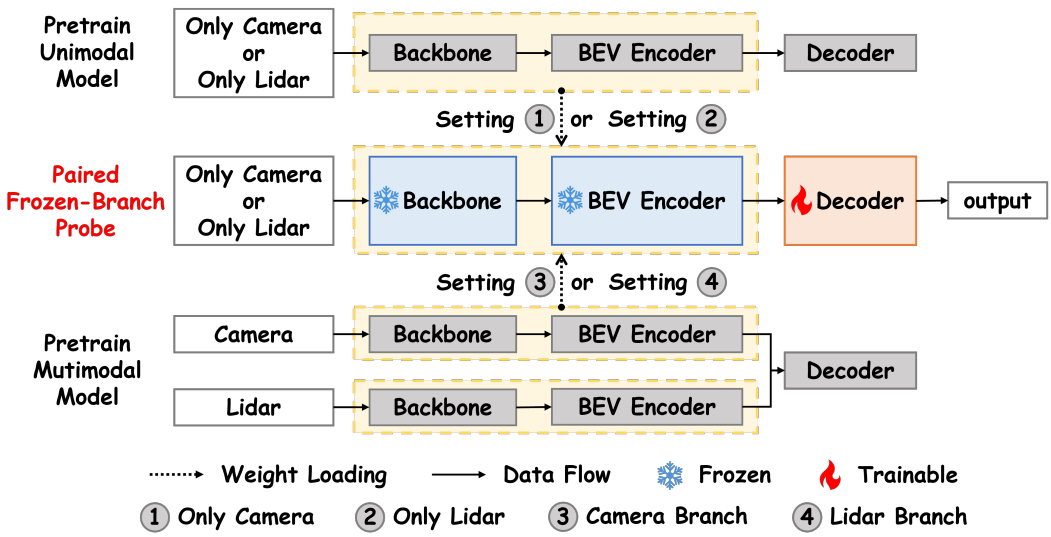


Figure 2: **Paired Frozen-Branch Probe for Camera and LiDAR representations.**

Probe Performance. Let q denote the task decoder attached to the frozen representation extractor g. With g fixed, q_g^* denotes the decoder obtained by training q under the Probe setting defined above. The Probe performance of g is defined as

$$P(g) = S\left(q_g^* \circ g\right), \tag{1}$$

where $S(\cdot)$ is the task-specific metric on a fixed validation set, and $\circ$ denotes function composition. Under a fixed decoder architecture and training setting, $P(g)$ measures how effectively the representation from g supports the target task, and serves as our operational measure of task-relevant representation quality.

Probe Configurations. To identify both the training source and the representation being evaluated, each Probe configuration is denoted by $A2B$. The prefix A specifies how the source model was trained: C, L, and F denote Camera-only, LiDAR-only, and Camera–LiDAR joint training, respectively. The suffix B specifies the representation type, with the same symbols denoting Camera, LiDAR, and fused Camera–LiDAR representations. Let g_B^A denote the type-B representation extractor from a model trained under setting A, and let $P(g_B^A)$ denote its Probe performance. For example, C2C evaluates the Camera representation from an independently trained Camera-only model and provides the reference score $P(g_C^C)$, whereas F2C evaluates the Camera representation from the corresponding branch of a jointly trained multimodal model, yielding $P(g_C^F)$.

Paired Comparison and Degradation Measure. C2C and F2C form a paired evaluation of Camera representations, while L2L and F2L form the corresponding evaluation of LiDAR representations. For modality $v \in \{C, L\}$, we define the branch performance difference as

$$\Delta_v = P\left(g_v^F\right) - P\left(g_v^v\right), \qquad v \in \{C, L\}. \tag{2}$$

For the Camera and LiDAR branches, the corresponding differences are

$$\Delta_C = P\left(g_C^F\right) - P\left(g_C^C\right), \qquad \Delta_L = P\left(g_L^F\right) - P\left(g_L^L\right). \tag{3}$$

A negative Δ_v indicates that the modality-specific branch in the multimodal model achieves lower Probe performance than its independently trained unimodal counterpart. We define this outcome as unimodal branch degradation.

3.2 Unimodal Branch Degradation across Tasks and Models

Tasks and Models. UBD-Bench evaluates unimodal branch degradation across four autonomous-driving perception tasks: 3D object detection, online vectorized map construction, BEV map segmentation, and occupancy prediction. These tasks differ in their prediction targets, spatial representations, and task-decoder architectures. Within each task, the selected models further cover

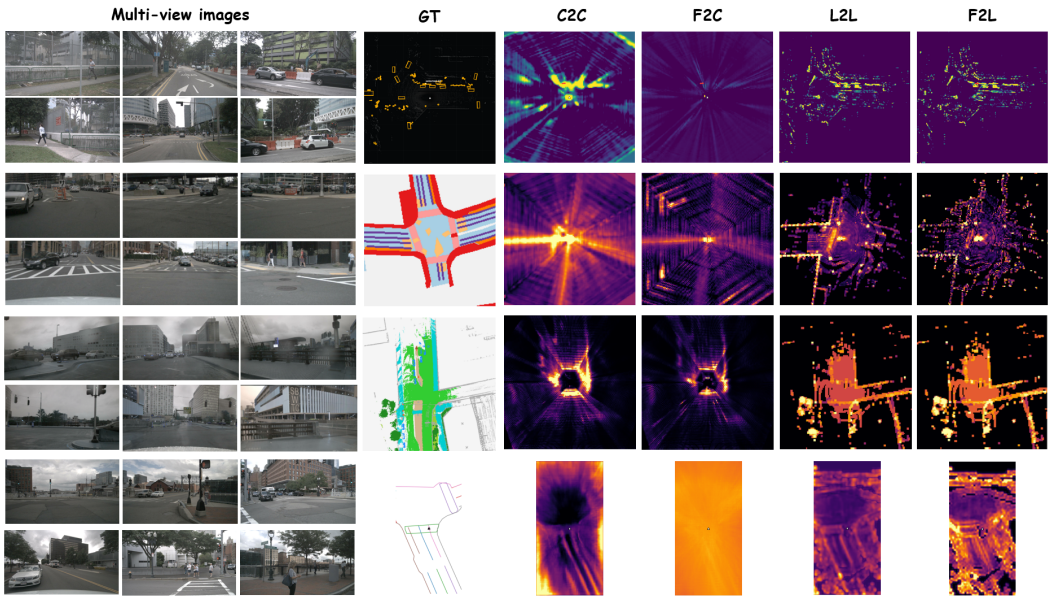


Figure 3: **Qualitative comparison of modality-specific pre-fusion BEV representations across four perception tasks.** The independently trained Camera and LiDAR representations are compared with the corresponding branches learned through multimodal joint training.

different Camera–LiDAR fusion architectures. Specifically, the 3D object detection evaluation includes BEVFusion (Liu et al., 2023) and CMT (Yan et al., 2023). Online vectorized map construction includes MapTR (Liao et al., 2022), HIMap (Zhou et al., 2024), MapTRv2 (Liao et al., 2025), and MapQR (Liu et al., 2024). BEV map segmentation includes BEVFusion (Liu et al., 2023), MapPrior (Zhu et al., 2023), and DifFUSER (Le et al., 2024). Occupancy prediction includes CONet (Wang et al., 2023a). This evaluation scope enables UBD-Bench to assess whether unimodal branch degradation is prevalent across perception tasks and model architectures.

Branch-Level Probe Settings. This section employs the Branch-Level Probe, with the probing position fixed at the pre-fusion BEV representation of each modality branch. All representation extraction modules from the raw sensor input to this position remain frozen. The paired comparisons for each model follow the unified protocol in Section 3.1. Decoder architectures and training budgets are determined for the corresponding task and matched within each paired comparison. Evaluation uses the metrics adopted in the official implementations of the respective baselines. Specifically, 3D object detection is evaluated using the official nuScenes mAP and NDS (Caesar et al., 2020), online vectorized map construction using mAP (Liao et al., 2022), BEV map segmentation using mean IoU over six map categories (Liu et al., 2023), and occupancy prediction using IoU and fine-grained semantic mIoU (Wang et al., 2023a).

In addition to the branch-level Probe results, we report the end-to-end performance of the original Camera-only, LiDAR-only, and multimodal models, denoted by Full C, Full L, and Full F, respectively. These results characterize task performance under the models' original sensor configurations and inference pipelines. The fused-representation Probe result, F2F, is also reported to compare performance obtained with a frozen fused representation and a newly trained decoder against the original multimodal model's end-to-end performance.

Branch-Level Probe Results. As shown in Table 1, **Camera-branch degradation occurs consistently across tasks and models.** All reported Camera-branch comparisons yield $\Delta_C < 0$, indicating that Probe performance under F2C is lower than under C2C. For example, CMT's 3D detection mAP decreases from 36.43 under C2C to 10.50 under F2C, a drop of 25.93 percentage points. MapTR's vectorized mapping mAP decreases by 10.67 percentage points. Despite differences in prediction targets and evaluation metrics, the jointly trained Camera branches consistently achieve lower Probe performance than their independently trained counterparts. This pattern indicates that Camera-branch degradation is not confined to a particular task or model architecture within the evaluated setting.

Table 1: **Cross-task and cross-model evaluation of unimodal branch degradation.**

Task	Method	Metric	End-to-end			Camera Probe		LiDAR Probe		Fusion Probe
			Full C	Full L	Full F	C2C	F2C	L2L	F2L	F2F
3D Det.	BEVFusion	mAP↑	35.55	64.68	68.55	33.47	11.58 (**-21.89**)	47.69	47.70 (+0.01)	62.56
		NDS↑	41.21	69.22	71.41	36.49	19.02 (**-17.47**)	53.01	53.47 (+0.46)	61.31
	CMT	mAP↑	40.59	63.21	70.30	36.43	10.50 (**-25.93**)	61.64	61.88 (+0.24)	53.24
		NDS↑	46.04	68.27	72.94	41.94	25.88 (**-16.06**)	67.23	67.78 (+0.55)	60.06
Vec. Map	MapTR	mAP↑	48.57	54.26	63.14	51.09	40.42 (**-10.67**)	53.62	55.88 (+2.26)	62.92
	HIMap	mAP↑	65.02	65.32	73.57	64.41	57.17 (**-7.24**)	65.08	64.37 (**-0.71**)	72.32
	MapTRv2	mAP↑	61.15	59.49	72.28	60.93	53.65 (**-7.28**)	58.53	61.86 (+3.33)	70.87
	MapQR	mAP↑	66.15	61.85	72.56	66.56	59.62 (**-6.94**)	62.28	62.90 (+0.62)	73.72
BEV Seg.	BEVFusion	mIoU↑	57.09	48.58	62.94	56.44	51.20 (**-5.24**)	47.45	45.05 (**-2.40**)	61.48
	MapPrior	mIoU↑	56.76	52.41	63.15	56.56	51.34 (**-5.22**)	47.69	45.67 (**-2.02**)	61.73
	DifFUSER	mIoU↑	53.74	46.28	63.51	55.26	45.11 (**-10.15**)	45.97	46.91 (+0.94)	64.83
Occ.	CONet	IoU↑	21.50	30.00	26.50	20.80	20.10 (**-0.70**)	30.50	30.50 (+0.00)	28.90
		$\mathrm{mIoU}_{\mathrm{fine}}$ ↑	13.60	15.90	20.50	13.10	12.00 (**-1.10**)	16.10	16.20 (+0.10)	20.60

- Values in parentheses in the F2C and F2L columns denote Δ_C and Δ_L, respectively.
- Negative differences indicate branch degradation and are highlighted in bold.

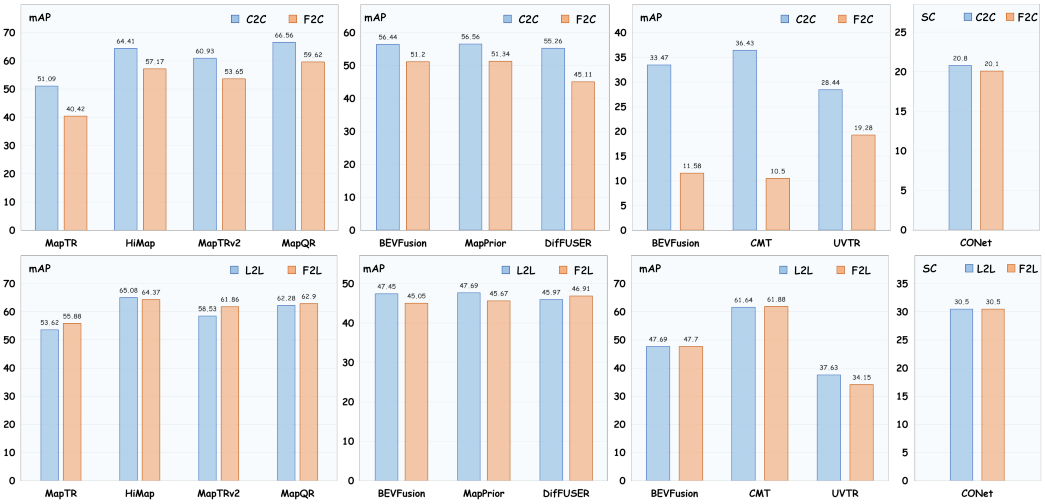


Figure 4: **Cross-task comparison of Camera- and LiDAR-branch Probe performance.** The top row shows consistent Camera-branch degradation across tasks and models, whereas the bottom row shows more model-dependent changes in the LiDAR branch. Blue and orange bars denote independently trained unimodal branches and the corresponding multimodal-model branches, respectively.

The LiDAR branch exhibits greater model dependence, revealing an asymmetric pattern of degradation across modalities. For CMT, Camera-branch mAP decreases by 25.93 percentage points, whereas LiDAR-branch mAP remains nearly unchanged, increasing from 61.64 to 61.88. Other models exhibit both improvements and declines in LiDAR-branch performance. For example, mAP of MapTRv2 increases by 3.33 percentage points, while BEVFusion's BEV map segmentation mIoU decreases by 2.40 percentage points. These results indicate that the changes in branch representations associated with multimodal joint training differ across modalities. Camera branches consistently exhibit degradation within the current evaluation scope, whereas LiDAR-branch outcomes vary across models and tasks.

Qualitative Analysis of Modality-Specific Representations. Figure 3 qualitatively compares the pre-fusion BEV representations learned through independent unimodal training with those learned by the corresponding branches through multimodal joint training. Across the representative models from all four tasks, the Fusion-Camera representations exhibit noticeable changes in spatial organization relative to their Camera-only counterparts. The differences are particularly pronounced in 3D object detection and online vectorized map construction, where the Fusion-Camera activations become less spatially distinctive. In comparison, the change is more subtle for occupancy prediction,

consistent with its smaller Camera-branch performance difference. The Fusion-LiDAR representations, however, largely preserve the spatial patterns observed in their LiDAR-only counterparts and do not exhibit a consistent loss of structure across tasks. These qualitative observations support the Probe results: Camera-branch degradation occurs consistently across tasks, although its magnitude varies, whereas the effect of multimodal joint training on the LiDAR branch depends more strongly on the task and model.

3.3 Hierarchical Probe and Degradation Localization

The branch-level evaluation in Section 3.2 shows that the pre-fusion Camera BEV representations learned through multimodal joint training exhibit degradation. However, only evaluating the branch output cannot determine whether this gap is already present in earlier intermediate representations. To address this question, the Hierarchical Probe progressively moves the probing position upstream along the Camera–BEV generation pathway and retrains all downstream modules to assess representation gaps at different processing stages.

Hierarchical Probe Settings. The Camera–BEV generation pathway comprises a visual backbone, an image neck, and a BEV encoder, where the BEV encoder performs view transformation from image features to BEV representations. Five probing configurations, P0–P4, are defined along this pathway. P0 follows the branch-level setting in Section 3.2, freezing the entire Camera–BEV generation pathway and retraining only the task decoder. P1 moves the probing position to the input of the BEV encoder, keeping the visual backbone and image neck frozen while retraining the BEV encoder and task decoder. P2 moves the probing position further upstream to the visual-backbone output, freezing only the backbone and retraining the image neck, BEV encoder, and task decoder. P3 and P4 place the probing position at the inputs to the final one and two backbone stages, respectively. All preceding backbone modules remain frozen, while the corresponding backbone stages and all subsequent modules are retrained.

At each probing position, C2C and F2C follow the paired protocol in Section 3.1, with all downstream modules trained from scratch. One representative model is selected for each task: MapTR for online vectorized map construction, BEVFusion for BEV map segmentation and 3D object detection, and CONet for occupancy prediction. Each task retains the training budget and evaluation metrics used in Section 3.2.

Table 2: **Hierarchical probing along the Camera–BEV generation pathway.**

Vec. Map / MapTR

Metric	Pos.	C2C	F2C (Δ_C)
mAP↑	P0	51.09	40.42 (**-10.67**)
	P1	51.58	43.80 (**-7.78**)
	P2	49.84	45.73 (**-4.11**)
	P3	45.07	42.49 (**-2.58**)
	P4	26.09	26.49 (+0.40)

BEV Seg. / BEVFusion

Metric	Pos.	C2C	F2C (Δ_C)
mIoU↑	P0	56.44	51.20 (**-5.24**)
	P1	56.46	52.04 (**-4.42**)
	P2	55.78	53.79 (**-1.99**)
	P3	55.22	53.08 (**-2.14**)
	P4	52.38	51.02 (**-1.36**)

3D Det. / BEVFusion

Metric	Pos.	C2C	F2C (Δ_C)
mAP↑	P0	33.47	11.58 (**-21.89**)
	P1	34.43	15.11 (**-19.32**)
	P2	33.39	23.84 (**-9.55**)
	P3	33.12	26.67 (**-6.45**)
	P4	29.22	27.51 (**-1.71**)
NDS↑	P0	36.49	19.02 (**-17.47**)
	P1	37.88	22.04 (**-15.84**)
	P2	37.62	29.37 (**-8.25**)
	P3	37.95	32.29 (**-5.66**)
	P4	33.64	32.63 (**-1.01**)

Occ. / CONet

Metric	Pos.	C2C	F2C (Δ_C)
IoU↑	P0	20.80	20.10 (**-0.70**)
	P1	21.50	21.30 (**-0.20**)
	P2	21.30	20.10 (**-1.20**)
	P3	20.90	20.50 (**-0.40**)
	P4	20.10	19.50 (**-0.60**)
$\mathrm{mIoU}_{\mathrm{fine}}$ ↑	P0	13.10	12.00 (**-1.10**)
	P1	13.60	13.00 (**-0.60**)
	P2	13.40	12.90 (**-0.50**)
	P3	13.50	13.20 (**-0.30**)
	P4	11.90	11.70 (**-0.20**)

- Values in parentheses in the F2C column denote Δ_C.
- Negative differences indicate Camera-branch degradation and are highlighted in bold.

Hierarchical Probe Results. Camera-branch degradation extends throughout the Camera–BEV generation pathway. Table 2 shows $\Delta_C < 0$ at P0–P2 for all four tasks. For vectorized mapping, 3D detection, and BEV segmentation, successively retraining the BEV encoder and image neck reduces the degradation while C2C remains relatively stable. In 3D detection, Δ_C changes from -21.89 at P0 to -19.32 at P1 and -9.55 at P2, whereas C2C remains between 33.39 and 34.43 mAP. These paired changes indicate that both the BEV encoder and image neck contribute to the observed degradation. Occupancy prediction shows a non-monotonic trend, with the IoU difference decreasing from -0.20 at P1 to -1.20 at P2 while remaining degraded. Because all downstream modules have been retrained at P2, the remaining negative Δ_C demonstrates degradation in the frozen visual backbone.

The deep backbone also exhibits degradation, whereas P4 reveals a limitation of the Probe. At P3, Δ_C remains negative for every task. For 3D detection, C2C changes only from 33.39 to 33.12 mAP, while Δ_C changes from -9.55 to -6.45 after the deep backbone stage is included in retraining. Vectorized mapping and occupancy prediction show the same reduction in degradation, while BEV segmentation remains degraded by 2.14 points, indicating that degradation also occurs in the deep backbone. At P4, however, C2C decreases across all tasks; for MapTR, C2C/F2C falls from 45.07/42.49 at P3 to 26.09/26.49 mAP. This joint decline indicates that splitting the backbone at this depth prevents it from retaining its original feature-extraction capability, making the P4 Probe unreliable. P4 is therefore excluded from the overall hierarchical comparison. Overall, the visual backbone, image neck, and BEV encoder all exhibit degradation, which is most severe at the pre-fusion Camera BEV representation.

3.4 Distillation Intervention and Degradation Mitigation

Unimodal-to-Multimodal Distillation. The hierarchical probing results in Section 3.3 show that Camera-branch degradation occurs in the visual backbone, image neck, and BEV encoder and is most pronounced overall at the pre-fusion Camera BEV representation. Motivated by this finding, we introduce targeted unimodal-to-multimodal distillation at this position to examine whether improving the task-relevant representation quality of the Camera branch can mitigate the degradation and improve the task performance of the multimodal model.

Specifically, an independently trained Camera-only model serves as the teacher, while the Camera branch of the multimodal model serves as the student. The teacher and student process the same Camera input, and feature distillation is applied to their pre-fusion Camera BEV representations. Since knowledge is transferred from an independently trained unimodal model to the corresponding branch of a multimodal model, we refer to this procedure as **unimodal-to-multimodal distillation**.

Given an input image x, the frozen teacher and the student Camera branch produce pre-fusion Camera BEV features $F_{\mathrm{cam},T}(x)$ and $F_{\mathrm{cam},S}(x)$, respectively, where the subscripts T and S denote the teacher and student. The distillation loss is defined as

$$\mathcal{L}_{\mathrm{KD}} = \mathrm{MSE}\left(F_{\mathrm{cam},S}(x), F_{\mathrm{cam},T}(x)\right). \tag{4}$$

The loss is computed over all spatial positions and channels of the Camera BEV feature map. The multimodal model is optimized using

$$\mathcal{L} = \mathcal{L}_{\mathrm{task}} + \lambda_t \mathcal{L}_{\mathrm{KD}}, \tag{5}$$

where $\mathcal{L}_{\mathrm{task}}$ is the original training objective for task t, and λ_t controls the strength of the distillation constraint. The teacher remains frozen throughout training and is removed afterward. Therefore, the proposed intervention does not alter the inference architecture or introduce additional computational overhead during inference.

Experimental Settings. The distillation experiments use the representative models selected for the four tasks in Section 3.3. BEVFusion is used for 3D object detection and BEV map segmentation, MapTR for online vectorized map construction, and CONet for occupancy prediction. For each task, the multimodal baseline trained without distillation and its distilled counterpart use the same model architecture, parameter initialization, data processing, optimization settings, training budget, and evaluation procedure. The distillation loss is the only additional training constraint.

After training, the full multimodal model is first evaluated using its standard inference pipeline. Its Camera branch is then evaluated using the unified Probe protocol defined in Section 3.1. These two evaluations measure the effects of distillation on the task-relevant representation quality of the Camera branch and the final task performance of the multimodal model, respectively.

Table 3: **Effects of unimodal-to-multimodal reverse distillation.**

Task	Method	Metric	Full multimodal model		Camera-branch Probe	
			Baseline	+KD (Δ)	F2C	F2C + KD (Δ)
3D Det.	BEVFusion	mAP↑	68.27	**69.25** (+0.98)	12.06	**32.48** (+20.42)
		NDS↑	71.29	**71.88** (+0.59)	19.06	**35.12** (+16.06)
Vec. Map	MapTR	mAP↑	61.97	**64.49** (+2.52)	40.42	**43.54** (+3.12)
BEV Seg.	BEVFusion	mIoU↑	59.83	**60.45** (+0.62)	49.30	**55.46** (+6.16)
Occ.	CONet	IoU↑	25.30	**29.30** (+4.00)	20.10	**21.50** (+1.40)
		$\text{mIoU}_{\text{fine}}$ ↑	20.00	**20.80** (+0.80)	12.00	**12.90** (+0.90)

- Parentheses denote absolute changes from matched baselines and bold indicates improvements.

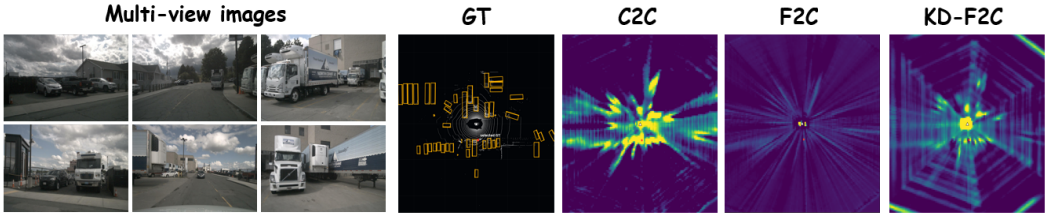


Figure 5: **Qualitative comparison of pre-fusion Camera BEV features for 3D object detection.** From left to right: ground-truth 3D bounding boxes, the C2C reference feature, the F2C feature without distillation, and the F2C feature after unimodal-to-multimodal distillation. The same feature channel and color scale are used for all activation maps.

Distillation Results. Unimodal-to-multimodal distillation improves the task-relevant representation quality of the Camera branch. As shown in Table 3, the distilled Camera branch achieves higher Probe performance on every reported metric. For example, its F2C mAP for 3D object detection improves by 20.42 percentage points. The consistent improvements across tasks indicate that representation supervision from an independently trained Camera-only model can effectively mitigate Camera-branch degradation under multimodal joint training.

Improving the Camera-branch representation further benefits the task performance of the full multimodal model. All reported end-to-end metrics improve after distillation, including a 4.00-percentage-point improvement in IoU for occupancy prediction. The concurrent improvements in branch-level Probe performance and end-to-end task performance indicate that the diagnosed degradation can be mitigated through targeted representation supervision and that improving the affected Camera representation benefits final multimodal prediction.

Qualitative Visualization. Figure 5 provides a representative visualization of the pre-fusion Camera BEV features for 3D object detection. Without distillation, the F2C feature exhibits substantially weaker activations and less distinct spatial structure than the C2C reference. After unimodal-to-multimodal distillation, structured responses become more apparent in the forward region and around the selected object, making the feature qualitatively closer to the C2C reference. This example visually complements the consistent improvements in Camera-branch Probe performance reported in Table 3.

4 Conclusion

This work introduces UBD-Bench for systematically evaluating unimodal branch degradation in multimodal autonomous-driving perception. Across four perception tasks and 10 models, we observe prevalent Camera-branch degradation and a clear asymmetry between Camera and LiDAR. Hierarchical probing further shows that the degradation spans the Camera–BEV pathway and is most pronounced at the pre-fusion Camera BEV representation. Distillation at this position consistently improves both Camera-branch Probe performance and end-to-end performance across all four tasks, providing intervention-based evidence for the diagnosis and demonstrating that the degradation can be mitigated. Overall, UBD-Bench provides a unified framework for measuring and improving modality-specific representations under multimodal joint training.

AI USE STATEMENT

Generative AI tools were used to assist with manuscript drafting and revision. The authors reviewed the AI-assisted content, checked factual claims against the underlying experimental results and cited sources, and made the final decisions on the manuscript. The authors take responsibility for the accuracy and integrity of all content in this work.

REFERENCES

Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *2020 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 11618–11628. IEEE, 2020.

Yinpeng Dong, Caixin Kang, Jinlai Zhang, Zijian Zhu, Yikai Wang, Xiao Yang, Hang Su, Xingxing Wei, and Jun Zhu. Benchmarking robustness of 3d object detection to common corruptions in autonomous driving. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1022–1032. IEEE, 2023.

Chongjian Ge, Junsong Chen, Enze Xie, Zhongdao Wang, Lanqing Hong, Huchuan Lu, Zhenguo Li, and Ping Luo. Metabev: Solving sensor failures for 3d detection and map segmentation. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8687–8697. IEEE, 2023.

Xiaoshuai Hao, Ruikai Li, Hui Zhang, Dingzhe Li, Rong Yin, Sangil Jung, Seung-In Park, ByungIn Yoo, Haimei Zhao, and Jing Zhang. Mapdistill: Boosting efficient camera-based hd map construction via camera-lidar fusion model distillation. In *European Conference on Computer Vision*, pp. 166–183. Springer, 2024a.

Xiaoshuai Hao, Mengchuan Wei, Yifan Yang, Haimei Zhao, Hui Zhang, Yi Zhou, Qiang Wang, Weiming Li, Lingdong Kong, and Jing Zhang. Is your hd map constructor reliable under sensor corruptions? *Advances in Neural Information Processing Systems*, 37:22441–22482, 2024b.

Duy-Tho Le, Hengcan Shi, Jianfei Cai, and Hamid Rezatofighi. Diffusion model for robust multi-sensor fusion in 3d object detection and bev segmentation. In *European conference on computer vision*, pp. 232–249. Springer, 2024.

Heng Li, Yuenan Hou, Xiaohan Xing, Yuexin Ma, Xiao Sun, and Yanyong Zhang. Occmamba: Semantic occupancy prediction with state space models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11949–11959. IEEE, 2025.

Jianing Li, Ming Lu, Jiaming Liu, Yandong Guo, Yuan Du, Li Du, and Shanghang Zhang. Bev-lgkd: A unified lidar-guided knowledge distillation framework for multi-view bev 3d object detection. *IEEE Transactions on Intelligent Vehicles*, 9(1):2489–2498, 2023.

Ruikai Li, Xinrun Li, Mengwei Xie, Hao Shan, Shoumeng Qiu, Xinyuan Chang, Yizhe Fan, Feng Xiong, Han Jiang, Yilong Ren, et al. Amap: Distilling future priors for ahead-aware online hd map construction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24906–24917, 2026.

Bencheng Liao, Shaoyu Chen, Xinggang Wang, Tianheng Cheng, Qian Zhang, Wenyu Liu, and Chang Huang. Maptr: Structured modeling and learning for online vectorized hd map construction. *arXiv preprint arXiv:2208.14437*, 2022.

Bencheng Liao, Shaoyu Chen, Yunchi Zhang, Bo Jiang, Qian Zhang, Wenyu Liu, Chang Huang, and Xinggang Wang. Maptrv2: An end-to-end framework for online vectorized hd map construction. *International Journal of Computer Vision*, 133(3):1352–1374, 2025.

Zhijian Liu, Haotian Tang, Alexander Amini, Xinyu Yang, Huizi Mao, Daniela L Rus, and Song Han. Bevfusion: Multi-task multi-sensor fusion with unified bird's-eye view representation. In *2023 IEEE international conference on robotics and automation (ICRA)*, pp. 2774–2781. ieee, 2023.

Zihao Liu, Xiaoyu Zhang, Guangwei Liu, Ji Zhao, and Ningyi Xu. Leveraging enhanced queries of point sets for vectorized map construction. In *European Conference on Computer Vision*, pp. 461–477. Springer, 2024.

Zhenxing Ming, Julie Stephany Berrio, Mao Shan, and Stewart Worrall. Occfusion: Multi-sensor fusion framework for 3d semantic occupancy prediction. *IEEE Transactions on Intelligent Vehicles*, 2024.

Yining Shi, Kun Jiang, Ke Wang, Jiusi Li, Yunlong Wang, Mengmeng Yang, and Diange Yang. Streamingflow: Streaming occupancy forecasting with asynchronous multi-modal data streams via neural ordinary differential equation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14833–14842. IEEE, 2024.

Ziying Song, Lin Liu, Feiyang Jia, Yadan Luo, Caiyan Jia, Guoxin Zhang, Lei Yang, and Li Wang. Robustness-aware 3d object detection in autonomous driving: A review and outlook. *IEEE Transactions on Intelligent Transportation Systems*, 25(11):15407–15436, 2024.

Xiaofeng Wang, Zheng Zhu, Wenbo Xu, Yunpeng Zhang, Yi Wei, Xu Chi, Yun Ye, Dalong Du, Jiwen Lu, and Xingang Wang. Openoccupancy: A large scale benchmark for surrounding semantic occupancy perception. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 17804–17813. IEEE, 2023a.

Zeyu Wang, Dingwen Li, Chenxu Luo, Cihang Xie, and Xiaodong Yang. Distillbev: Boosting multi-camera 3d object detection with cross-modal knowledge distillation. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8603–8612. IEEE, 2023b.

Shaoyuan Xie, Lingdong Kong, Wenwei Zhang, Jiawei Ren, Liang Pan, Kai Chen, and Ziwei Liu. Benchmarking and improving bird's eye view perception robustness in autonomous driving. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(5):3878–3894, 2025.

Junjie Yan, Yingfei Liu, Jianjian Sun, Fan Jia, Shuailin Li, Tiancai Wang, and Xiangyu Zhang. Cross modal transformer: Towards fast and robust 3d object detection. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 18222–18232. IEEE, 2023.

Weiqing Yan, Changhong Bu, Yuanyang Zhang, Guanghui Yue, Wei Zhou, Xiatian Zhu, and Jungong Han. 3d object detection and knowledge distillation in autonomous driving: A survey. *Information Fusion*, pp. 104706, 2026.

Zhixin Zhang, Yiyuan Zhang, Xiaohan Ding, Fusheng Jin, and Xiangyu Yue. Online vectorized hd map construction using geometry. In *European Conference on Computer Vision*, pp. 73–90. Springer, 2024.

Lingjun Zhao, Jingyu Song, and Katherine A Skinner. Crkd: Enhanced camera-radar object detection with cross-modality knowledge distillation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15470–15480. IEEE, 2024.

Shengchao Zhou, Weizhou Liu, Chen Hu, Shuchang Zhou, and Chao Ma. Unidistill: A universal cross-modality knowledge distillation framework for 3d object detection in bird's-eye view. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5116–5125. IEEE, 2023.

Yi Zhou, Hui Zhang, Jiaqian Yu, Yifan Yang, Sangil Jung, Seung-In Park, and ByungIn Yoo. Himap: Hybrid representation learning for end-to-end vectorized hd map construction. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15396–15406. IEEE, 2024.

Xiyue Zhu, Vlas Zyrianov, Zhijian Liu, and Shenlong Wang. Mapprior: Bird's-eye view map layout estimation with generative models. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8194–8205. IEEE, 2023.

Appendix

A Algorithmic Specifications

This section provides pseudocode for the Paired Frozen-Branch Probe, hierarchical Probe, and unimodal-to-multimodal reverse-distillation procedures used in this work.

A.1 Paired Frozen-Branch Probe

Algorithm A.1 specifies the construction and evaluation of each paired Probe.

Algorithm A.1: Paired Frozen-Branch Probe

Input: unimodal checkpoint W_v^v; multimodal checkpoint W_v^F; target modality $v \in \{C, L\}$; probing position p; downstream architecture q; shared downstream initialization θ_0; matched training configuration $\mathcal{C}$; validation set $\mathcal{D}_{\text{val}}$.

1. Extract $g_v^v \leftarrow \text{ExtractBranch}(W_v^v, v, p)$ and $g_v^F \leftarrow \text{ExtractBranch}(W_v^F, v, p)$.
2. Load the upstream parameters from their respective checkpoints.
3. Freeze g_v^v and g_v^F and keep both extractors in evaluation mode.
4. Instantiate two downstream networks, $q_v^v \leftarrow q(\theta_0)$ and $q_v^F \leftarrow q(\theta_0)$, with identical initial parameters.
5. Train q_v^v and q_v^F independently while keeping their corresponding extractors frozen and using the same configuration $\mathcal{C}$.
6. Evaluate $P(g_v^v) = S((q_v^v)^* \circ g_v^v)$ and $P(g_v^F) = S((q_v^F)^* \circ g_v^F)$ on $\mathcal{D}_{\text{val}}$.
7. Compute $\Delta_v = P(g_v^F) - P(g_v^v)$.

Output: $P(g_v^v)$, $P(g_v^F)$, and Δ_v.

A.2 Hierarchical Probe

Algorithm A.2 specifies the evaluation procedure at successive probing positions along the Camera–BEV generation pathway.

Algorithm A.2: Hierarchical Probe

Input: Camera-only checkpoint W_C^C; multimodal Camera-branch checkpoint W_C^F; ordered probing positions $\mathcal{P} = (P0, P1, \ldots, PK)$; module chain $\mathcal{M}$; matched position-specific initializations $\{\theta_p\}_{p\in\mathcal{P}}$; training configuration $\mathcal{C}$; validation set $\mathcal{D}_{\text{val}}$.

1. For each position $p \in \mathcal{P}$, partition $\mathcal{M}$ into the upstream extractor g^p ending at p and the downstream network q^p following p.
2. Extract $g_C^{C,p}$ from W_C^C and $g_C^{F,p}$ from W_C^F; load their checkpoint parameters and keep them frozen in evaluation mode.
3. Instantiate $(q_C^{C,p}, q_C^{F,p})$ from the same position-specific initialization θ_p.
4. Train the two downstream networks independently under the same configuration $\mathcal{C}$ while their upstream extractors remain frozen.
5. Evaluate $P(g_C^{C,p})$ and $P(g_C^{F,p})$ on $\mathcal{D}_{\text{val}}$.
6. Record $\Delta_C^p = P(g_C^{F,p}) - P(g_C^{C,p})$, together with the trainable downstream scope at position p.

Output: $\{P(g_C^{C,p}), P(g_C^{F,p}), \Delta_C^p\}_{p\in\mathcal{P}}$.

A.3 Unimodal-to-Multimodal Reverse Distillation

Algorithm A.3 specifies the training procedure for transferring the pre-fusion Camera representation from a frozen unimodal teacher to the Camera branch of a multimodal student.

Algorithm A.3: Unimodal-to-Multimodal Reverse Distillation

Input: pretrained Camera-only teacher T; multimodal student M; training samples (x_C, x_L, y); task loss $\mathcal{L}_{\text{task}}$; distillation weight λ_t; training configuration $\mathcal{C}$.

1. Freeze all parameters of T and keep the teacher in evaluation mode.
2. For each minibatch (x_C, x_L, y), compute the teacher's pre-fusion Camera BEV feature $F^T_{\text{cam}} = T_{\text{cam}}(x_C)$ without gradient tracking.
3. Run the multimodal student on (x_C, x_L) to obtain its pre-fusion Camera BEV feature F^S_{cam} and task prediction $\hat{y}$.
4. Compute $\mathcal{L}_{\text{KD}} = \text{MSE}(F^S_{\text{cam}}, F^T_{\text{cam}})$.
5. Compute $\mathcal{L} = \mathcal{L}_{\text{task}}(\hat{y}, y) + \lambda_t \mathcal{L}_{\text{KD}}$.
6. Update the trainable parameters of M under $\mathcal{C}$; do not update T.
7. After training, evaluate the distilled full multimodal model and its Camera branch using the standard full-model and F2C protocols.

Output: distilled multimodal model M^{KD} and its full-model and F2C performance.

B Additional Controlled and Hierarchical Probe Results

B.1 Controlled Evaluation across Visual Backbones

The cross-task results may be affected by the design and capacity of the visual backbone. We therefore conduct a controlled study to determine whether Camera-branch degradation persists when the visual backbone changes while the remaining representation and evaluation pipeline is held fixed.

For both MapTR and HIMap, we evaluate three convolutional backbones (ResNet-18, ResNet-50, and ResNet-101) and two hierarchical Transformer backbones (Swin-T and Swin-S). Within each method, the Camera-to-BEV module, task decoder, data split, map range, training schedule, training budget, and Probe protocol remain unchanged. For every backbone, we train a Camera-only source model and a multimodal source model and then construct the corresponding C2C and F2C Probes.

Table 4: Camera Probe performance under controlled changes of the visual backbone. All values are mAP points.

Method	Backbone	C2C	F2C	Δ_C
MapTR	ResNet-18	44.83	34.90	**-9.93**
	ResNet-50	51.09	40.42	**-10.67**
	ResNet-101	51.94	42.11	**-9.83**
	Swin-T	52.05	44.41	**-7.64**
	Swin-S	55.26	46.45	**-8.81**
HIMap	ResNet-18	60.60	53.23	**-7.37**
	ResNet-50	64.41	57.17	**-7.24**
	ResNet-101	64.54	56.00	**-8.54**
	Swin-T	65.98	58.68	**-7.30**
	Swin-S	67.88	60.83	**-7.05**

All ten comparisons exhibit Camera-branch degradation. Averaged over the five backbones, the degradation is 9.38 mAP for MapTR and 7.50 mAP for HIMap, remaining substantial across convolutional and Transformer backbones. Stronger backbones improve both C2C and F2C: from ResNet-18 to Swin-S, MapTR increases from 44.83 to 55.26 in C2C and from 34.90 to 46.45 in F2C, while HIMap increases from 60.60 to 67.88 and from 53.23 to 60.83, respectively. Thus, greater visual capacity benefits both training sources.

However, degradation does not decrease monotonically with backbone capacity. It ranges from 7.64 to 10.67 mAP for MapTR and from 7.05 to 8.54 mAP for HIMap. Even with the best-performing Swin-S, degradation remains 8.81 mAP for MapTR and 7.05 mAP for HIMap; the ResNet variants likewise change absolute Probe performance without consistently reducing degradation. Therefore, adopting a stronger backbone improves the model's performance level but does not eliminate Camera-branch degradation. Taken together, the controlled backbone ablation demonstrates that Camera-branch degradation persists across different backbone families, model capacities, and model architectures.

B.2 Additional Hierarchical Camera Probe on DifFUSER

Because of the space limit, the main paper reports the complete BEV-segmentation hierarchy only for BEVFusion. We therefore provide the full five-position DifFUSER results in the appendix to examine whether the stage-wise pattern extends to a substantially different fusion architecture. Unlike the conventional BEV-level fusion used by BEVFusion, DifFUSER refines multimodal features through a Gated Self-conditioned Modulated latent diffusion module and arranges its fusion blocks in a hierarchical cMini-BiFPN architecture (Le et al., 2024). This difference makes DifFUSER a complementary test of whether the hierarchical Camera-branch trend depends on one particular fusion design.

Table 5: Hierarchical Camera Probe results on DifFUSER. All values are percentage points.

Position	C2C	F2C	Δ_C
P0	55.26	45.11	**-10.15**
P1	54.97	46.88	**-8.09**
P2	55.51	50.63	**-4.88**
P3	54.78	51.79	**-2.99**
P4	51.53	51.16	-0.37

At P0, where the complete Camera-to-BEV pathway remains frozen, DifFUSER exhibits its most severe Camera-branch degradation of 10.15 points. As the Camera-to-BEV transformation and image neck are successively included in retraining, C2C remains stable between 54.97 and 55.51, while the degradation magnitude represented by Δ_C decreases from 10.15 points at P0 to 8.09 points at P1 and 4.88 points at P2. This pattern shows that both stages exhibit degradation. The remaining 4.88-point degradation at P2 is embedded in the frozen visual backbone. When the deep backbone stage is additionally retrained at P3, the degradation decreases to 2.99 points, showing that the deep backbone also exhibits degradation.

At P4, C2C decreases from 54.78 to 51.53, while F2C changes from 51.79 to 51.16. This may result from the Probe splitting the backbone into frozen and retrained components, which prevents the backbone from retaining its original feature-extraction capability under the current training setting. This indicates that the Probe is no longer reliable for stage-wise comparison at this depth. We therefore exclude P4 from the overall hierarchical comparison. Based on P0–P3, DifFUSER and BEVFusion exhibit the same stage-wise pattern despite their different fusion mechanisms. These results show that Camera-branch degradation involves the complete Camera–BEV pathway and is most severe at the pre-fusion Camera BEV representation.

B.3 Hierarchical LiDAR Probe for BEV Segmentation

The cross-task evaluation shows that the LiDAR branch undergoes degradation in some models, although this behavior is less consistent and generally less severe than Camera-branch degradation. This modality asymmetry raises a further question: when LiDAR-branch degradation occurs, does it follow the same stage-wise pattern observed along the Camera pathway? To examine this question, we move the probing position progressively into the LiDAR pathway of BEVFusion for BEV map segmentation.

P0 retrains the BEV decoder and map head while keeping the complete LiDAR representation pathway frozen. P1 additionally includes the output transformation layer `conv_out`. P2–P4 then successively include the final three LiDAR encoder layers, allowing us to examine how the measured degradation changes as increasingly deeper parts of the LiDAR backbone are retrained.

Table 6: Hierarchical LiDAR Probe results on BEVFusion map segmentation.

Position	L2L	F2L	Δ_L
P0	47.45	45.05	**-2.40**
P1	47.21	45.34	**-1.87**
P2	46.96	46.21	-0.75
P3	46.59	45.81	-0.78
P4	47.05	46.03	-1.02

At P0, the LiDAR branch exhibits 2.40 points of degradation, confirming LiDAR-branch degradation, although it is less severe than Camera-branch degradation. Including `conv_out` in retraining at P1 increases F2L from 45.05 to 45.34 while L2L changes only slightly from 47.45 to 47.21, reducing the degradation to 1.87 points. At P2, retraining the final LiDAR encoder layer increases F2L to 46.21 while L2L remains stable at 46.96, reducing the degradation to 0.75 point. Similar to the Camera hierarchy, the P0–P2 results show degradation in both the output transformation and the LiDAR backbone.

At P3 and P4, the degradation changes to 0.78 and 1.02 points rather than continuing a stable reduction. These results support degradation in the LiDAR backbone as a whole but cannot localize it to a particular internal layer. Overall, the LiDAR hierarchy resembles the Camera hierarchy at P1 and P2 but differs in both degradation severity and consistency within the deeper backbone. This supports the modality asymmetry observed in the main benchmark while showing that the LiDAR branch is also affected by multimodal joint training.

B.4 Cross-Model Category-Level Consistency

The aggregate BEV-segmentation results show consistent Camera-branch degradation but a more variable LiDAR-branch response. An aggregate mIoU difference, however, cannot determine whether degradation is shared across semantic categories or dominated by a small number of classes. We therefore examine the class-wise branch differences at the pre-fusion BEV position. BEV map segmentation is suitable for this analysis because its six semantic classes provide a sufficiently detailed yet compact decomposition of the aggregate metric. Table 7 compares three architectures with different fusion designs: BEVFusion, MapPrior, and DifFUSER.

Table 7: Class-wise branch differences at the pre-fusion BEV representation.

Method	Branch	mIoU	Drivable	Ped. crossing	Walkway	Stop line	Carpark	Divider
BEVFusion	Camera	-5.24	-3.64	-6.15	-6.20	-6.30	-4.02	-5.12
	LiDAR	-2.40	-2.53	-1.32	-2.03	-1.52	-5.25	-1.76
MapPrior	Camera	-5.22	-3.65	-6.06	-5.75	-6.19	-4.67	-5.05
	LiDAR	-2.02	-2.77	-2.25	-2.18	-2.86	+0.28	-2.37
DifFUSER	Camera	-10.15	-7.41	-12.93	-10.73	-13.06	-8.81	-7.96
	LiDAR	+0.94	+0.76	-0.04	+0.88	+0.23	+3.99	-0.14

Camera rows report F2C−C2C, and LiDAR rows report F2L−L2L. Negative values indicate lower Probe performance for the branch extracted from the multimodal model. All values are percentage points.

All 18 Camera class comparisons are negative. Camera-branch degradation ranges from 3.64 to 6.30 points for BEVFusion, from 3.65 to 6.19 points for MapPrior, and from 7.41 to 13.06 points for DifFUSER. Across all three architectures, degradation affects both large-area and smaller structural classes. Although its severity varies across models and categories, its direction remains unchanged in every comparison. The negative aggregate mIoU differences are therefore not caused by category averaging or one particularly difficult class.

The LiDAR results follow a different pattern. BEVFusion exhibits degradation in all six classes, whereas MapPrior exhibits degradation in five classes and a 0.28-point improvement for car parking. DifFUSER improves in four classes, while pedestrian crossing and divider remain lower by 0.04 and 0.14 point. For car parking, BEVFusion degrades by 5.25 points, whereas MapPrior and DifFUSER improve by 0.28 and 3.99 points. LiDAR-branch behavior therefore depends on both architecture and semantic category. Together, Camera-branch degradation is consistent across models and semantic targets, whereas LiDAR lacks the same uniform pattern, reinforcing the modality asymmetry. The next subsection examines whether this distinction persists across hierarchical probing positions.

B.5 Stage-Wise Category-Level Consistency

The preceding subsection shows that degradation at the pre-fusion BEV representation is consistent across Camera categories but more variable for LiDAR. We next test whether this distinction persists as the Probe moves through successive representation stages. The pre-fusion comparison alone

cannot show whether different categories change together when individual modules are included in retraining. Table 8 therefore reports the six class-wise IoU scores from the matched BEVFusion Camera and LiDAR hierarchies at P0–P4.

Table 8: Class-wise BEVFusion hierarchical Probe results.

Branch	Pos.	mIoU	Drivable	Ped. crossing	Walkway	Stop line	Carpark	Divider
Camera	P0	56.44/51.20	81.11/77.47	53.55/47.40	57.77/51.57	47.21/40.91	52.60/48.58	46.38/41.26
	P1	56.46/52.04	81.06/77.90	53.62/48.51	57.42/52.32	46.99/42.02	53.22/49.13	46.43/42.37
	P2	55.78/53.79	80.65/78.87	52.73/50.90	57.06/54.72	47.18/44.39	51.07/50.32	45.98/43.52
	P3	55.22/53.08	80.49/78.94	52.82/50.48	56.37/54.67	46.42/44.15	49.83/46.33	45.39/43.88
	P4	52.38/51.02	78.70/78.73	50.81/49.73	54.93/53.67	42.41/41.32	44.07/40.23	43.39/42.46
LiDAR	P0	47.45/45.05	74.91/72.38	46.28/44.96	55.91/53.88	32.05/30.53	35.35/30.10	40.20/38.44
	P1	47.21/45.34	74.58/72.41	46.44/45.46	55.14/54.70	31.40/30.48	35.63/30.60	40.07/38.38
	P2	46.96/46.21	74.60/73.90	47.10/45.29	55.84/54.61	32.42/30.83	31.63/33.75	40.16/38.88
	P3	46.59/45.81	74.59/74.44	46.35/45.80	55.77/54.84	32.19/31.41	30.64/29.12	39.99/39.21
	P4	47.05/46.03	74.56/74.23	47.54/46.08	55.85/55.48	32.59/31.72	31.99/29.22	39.78/39.43

Camera entries report C2C/F2C, and LiDAR entries report L2L/F2L. The first value is obtained from the independently trained branch and the second from the corresponding branch of the multimodal model. All values are percentage points.

Consistent with the preceding cross-model results, F2C is lower than C2C for all six Camera categories from P0 to P3. Every category exhibits less degradation at P2 than at P0 after retraining the Camera-to-BEV transformation and image neck. At P3, all six categories still exhibit degradation, ranging from 1.51 points for dividers to 3.50 points for car parking, supporting degradation in the deep visual backbone. At P4, C2C and F2C decrease together in five categories, indicating that the backbone split prevents it from retaining its original feature-extraction capability. The Probe is therefore unreliable at this depth, and P4 is excluded from the overall hierarchical comparison.

The LiDAR hierarchy is less uniform. F2L is lower than L2L for every category at P0 and P1. At P2, five classes remain lower, while car parking is 2.12 points higher; degradation decreases for the drivable area, walkway, car-parking area, and divider, whereas pedestrian crossing and stop line do not follow the same pattern. At P3 and P4, all six categories are lower again, but degradation does not show a stable progression within the deeper backbone. These results support degradation in the LiDAR backbone as a whole without localizing it to a particular internal layer. Thus, the cross-model consistency extends across the Camera hierarchy, whereas LiDAR-branch degradation remains milder and more dependent on category and probing position.

C Additional Analysis of Reverse Distillation

C.1 Spatial Supervision Ablation on MapTR

We first examine which spatial regions of the Camera BEV representation should be supervised during reverse distillation. All variants use a frozen Camera-only MapTR teacher and a Camera–LiDAR MapTR student trained from the same base initialization. The distillation loss is applied to the student's pre-fusion Camera BEV representation. R1 applies mean squared error (MSE) to all BEV cells, R2 applies MSE only within foreground regions defined by the ground-truth vector map, and R3 retains full-BEV supervision while assigning a weight of four to foreground cells and a weight of one to background cells.

Table 9: Spatial supervision ablation for MapTR reverse distillation.

Run	Spatial supervision	Final mAP	Best mAP	Δ mAP	Divider	Ped. crossing	Boundary
R0	No distillation	61.97	61.97	–	61.53	54.90	69.47
R1	Full-BEV MSE	**64.49**	**64.49**	**+2.52**	62.59	**60.40**	**70.49**
R2	Foreground-mask MSE	63.21	63.21	+1.24	61.84	57.92	69.86
R3	Full-BEV MSE, foreground 4×	64.02	64.04	+2.05	**62.90**	59.24	69.92

Δ mAP is measured relative to the non-distilled R0 baseline. The category columns report AP at the final checkpoint.

All three distillation variants outperform the non-distilled baseline, showing that the independently trained Camera representation provides useful supervision for multimodal training. Full-BEV MSE

achieves the highest final mAP of 64.49, improving R0 by 2.52 points. Restricting supervision to the vector-map foreground reduces the improvement to 1.24 points, while increasing the foreground weight yields 2.05 points. The result indicates that useful distillation signals are present beyond annotated foreground cells, including the surrounding spatial context represented in the Camera BEV feature.

The class-wise results show improvements in all three map categories under R1. The largest increase occurs for pedestrian crossings, from 54.90 to 60.40 AP, while divider and boundary AP increase by 1.06 and 1.02 points. R3 produces the highest divider AP but remains below R1 in overall mAP. We therefore use full-BEV MSE as the default MapTR distillation objective because it provides the strongest aggregate result without requiring a hand-designed spatial mask or foreground weighting rule.

C.2 Modality and Loss Ablations on BEVFusion

We next evaluate whether reverse distillation should target the Camera or LiDAR representation and whether more structured feature constraints improve over direct BEV matching. The student is a fully trainable BEVFusion model, and each teacher remains frozen. Distillation is performed at the modality-specific BEV representation immediately before fusion; the final segmentation output is not distilled. In addition to raw Camera BEV MSE, we evaluate normalized MSE, an energy-masked cosine objective, a local relation constraint, and a multilevel Camera-feature constraint.

Table 10: Modality and loss ablations for BEVFusion reverse distillation.

Run	Distilled representation	Objective	Best mIoU	Final mIoU	Δ best
Vanilla	–	No distillation	59.83	59.56	–
Camera MSE	Camera BEV	Raw MSE	**60.45**	**60.17**	**+0.62**
LiDAR MSE	LiDAR BEV	Raw MSE	59.60	59.19	-0.23
B1	Camera BEV	Normalized MSE	59.97	59.58	+0.14
B2	Camera BEV	Top-30% energy-masked cosine	59.92	59.51	+0.09
B3	Camera BEV	B2 + multiscale local relation	59.92	59.65	+0.09
B4	Camera BEV and neck	B3 + multilevel constraint	59.59	59.35	-0.24

Δ best is the change in best validation mIoU relative to the non-distilled Vanilla model. The local relation objective uses dilation rates 1, 2, and 4.

Camera BEV MSE increases the best mIoU from 59.83 to 60.45, whereas LiDAR BEV MSE decreases it to 59.60. This contrast agrees with the Probe diagnosis: Camera-branch degradation is consistent across the evaluated models, while the LiDAR response is more model dependent. The intervention result therefore supports applying reverse distillation specifically to the Camera branch.

The direct Camera BEV MSE also outperforms all four structured alternatives. Normalization, energy masking, and relation matching yield only 0.09–0.14 point improvements, and adding the multilevel constraint decreases the best mIoU by 0.24 point. These results show that increasing the complexity of the distillation objective does not provide a corresponding benefit in this setting. Direct MSE at the pre-fusion Camera BEV representation is consequently used as the default BEVFusion objective.

C.3 Cross-Architecture Generalization to DifFUSER

To test whether the task-level benefit of Camera BEV distillation depends on the BEVFusion architecture, we apply the same raw MSE constraint to DifFUSER. The Camera-only teacher is frozen, and the loss is computed at the pre-fusion Camera BEV representation of the fully trainable multimodal student. The training and evaluation protocol within each architecture is otherwise kept unchanged.

Table 11: Task-level effect of Camera BEV distillation across BEV segmentation architectures.

Architecture	Vanilla Fusion	Camera BEV-MSE	Δ mIoU
BEVFusion	59.83	60.45	+0.62
DifFUSER	63.51	64.13	+0.62

Table 12: Class-wise DifFUSER results with Camera BEV distillation.

Setting	mIoU	Drivable	Ped. crossing	Walkway	Stop line	Carpark	Divider
Vanilla Fusion	63.51	88.24	64.38	70.21	54.22	46.73	57.29
Camera BEV-MSE	**64.13**	87.79	**64.66**	70.17	**54.82**	**50.22**	57.10

Camera BEV distillation improves both architectures by 0.62 mIoU. On DifFUSER, the largest class-wise increase is observed for car parking, which rises from 46.73 to 50.22 IoU, followed by a 0.60-point increase for stop lines. Pedestrian-crossing IoU also increases, while the remaining categories change only slightly. The aggregate improvement therefore arises from a different distribution of class-level changes than a uniform increase across all categories.

These results show that the task-level benefit of the intervention extends to a second BEV-segmentation architecture. Because a post-distillation F2C Probe has not been evaluated for DifFUSER, this experiment supports improved final segmentation performance but does not by itself establish an improvement in the independently decodable quality of the DifFUSER Camera branch.

C.4 Branch-Level Improvements after Distillation

The preceding ablations evaluate the final multimodal prediction. We further apply the Paired Frozen-Branch Probe after distillation to determine whether the intervention also improves the Camera representation itself. For BEVFusion, we report class-wise F2C results at P0 and P2 before and after Camera BEV distillation. For MapTR, we report the mAP of the original and distilled F2C Probes together with the independently trained C2C reference.

Table 13: Class-wise BEVFusion Camera-branch Probe performance before and after distillation.

Pos.	Probe	mIoU	Drivable	Ped. crossing	Walkway	Stop line	Carpark	Divider
P0	Original F2C	49.30	75.97	46.36	50.90	39.36	42.25	40.97
	Distilled F2C	**55.46**	**80.40**	**53.17**	**57.40**	**46.37**	**49.68**	**45.74**
	Improvement	+6.16	+4.43	+6.81	+6.50	+7.01	+7.43	+4.77
P2	Original F2C	52.75	78.77	49.58	54.24	44.11	45.68	44.10
	Distilled F2C	**55.69**	**80.42**	**53.23**	**57.37**	**47.41**	**49.59**	**46.12**
	Improvement	+2.94	+1.65	+3.65	+3.13	+3.30	+3.91	+2.02

Table 14: MapTR Camera-branch Probe performance after full-BEV distillation.

Representation source	Probe mAP	Change from original F2C
Original Fusion Camera branch	40.42	–
Distilled Fusion Camera branch	**43.54**	**+3.12**
Independently trained Camera branch	51.09	–

At both BEVFusion positions, the distilled F2C Probe improves in all six semantic categories. At P0, mIoU increases by 6.16 points, with class-wise increases ranging from 4.43 points for the drivable area to 7.43 points for car parking. At P2, mIoU increases by 2.94 points, and every class remains higher after distillation. The improvement is therefore distributed across large-area and structural categories rather than being driven by a single semantic class.

MapTR exhibits the same direction at the pre-fusion Camera BEV position: the F2C Probe increases from 40.42 to 43.54 mAP after full-BEV distillation. This branch-level improvement accompanies the full-model increase reported in Table 9. Together, the BEVFusion and MapTR results show that reverse distillation improves the task-relevant information available in the Camera representation as well as the final multimodal task performance.