BROADER TASKS PREFER EGO-FRAME INVARIANT REPRESENTATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Computer vision has traditionally advanced through models optimized for individual tasks, such as object detection and semantic segmentation, while recent work increasingly seeks general representations that can support diverse downstream objectives. This raises the question of whether general representations can be guided by structural principles that hold independently of any particular downstream task. We study this question in multi-camera perception through the choice of ego coordinate frame, whose origin and axes are arbitrary. Changing this frame alters the numerical camera geometry while leaving the physical cameras, their relationships, and their visual observations unchanged. For image-projected representations, we show that consistent transformations of the underlying 3D quantities and camera geometry cancel under projection, motivating *ego-frame invariance* (EFI). We hypothesize that EFI is not universally advantageous for every specialized task, but becomes increasingly beneficial as a representation is required to support a broader range of predictions. To test this hypothesis in a setting where the output itself is not invariant, we study 3D object detection, an ego-frame equivariant (EFE) task whose predicted positions and orientations transform with the ego frame. We randomize the ego frame during training, and progressively broaden the task by increasing the number of target classes. Across nuScenes and Argoverse 2, using three camera-conditioning schemes, we evaluate this effect by held-out test performance obtained from selected checkpoints after training. The evaluation provides evidence that the advantage of ego-frame-invariant relative conditioning tends to grow as the prediction task becomes broader. These results support EFI as an inductive bias whose value grows as representations are required to support broader predictions, rather than as a universal advantage for every task.

1 INTRODUCTION

Recent work in embodied learning increasingly moves beyond narrowly defined perception or control tasks toward generalist agents that must operate across diverse tasks, environments, and embodiments (Ghosh et al., 2024; Kim et al., 2025; Team et al., 2025; NVIDIA et al., 2025). This shift places a different demand on the learned representation: rather than serving a single prediction objective, it must capture information that remains useful across a broad range of semantic, geometric, dynamic, and behavioral requirements. Accordingly, substantial effort has been devoted to learning general-purpose visual representations and predictive world models through task-agnostic pretraining objectives, as exemplified by DINOv2 and V-JEPA (Oquab et al., 2024; Assran et al., 2025).

If such a representation is intended to describe the observed world rather than a particular task, however, what dependencies should it retain? The physical content being observed is not altered by arbitrary conventions used to describe it, such as the choice of coordinate frame, basis, or ordering, even though particular numerical quantities may transform under these choices. A general representation should therefore reflect the structure of the observed world without acquiring additional dependence on conventions that are external to it. This principle underlies symmetry-aware representation learning in geometry (Cohen & Welling, 2016; Kondor & Trivedi, 2018; Bronstein et al., 2021; Cohen et al., 2019), as well as architectures that remove artificial dependencies such as the ordering of set-structured inputs (Zaheer et al., 2017; Lee et al., 2019).

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.

In multi-camera perception, one such artificial dependency is the choice of ego coordinate frame. Existing 3D perception models commonly express calibrated camera geometry in a shared ego-centric coordinate system (Philion & Fidler, 2020; Wang et al., 2022; Liu et al., 2022; Li et al., 2022; 2023; Liu et al., 2023). Yet the location and orientation of this coordinate frame are arbitrary: changing the ego origin or axes changes only the numerical description of the camera system, not the physical cameras or their visual observations. Ideally, a representation should transform consistently under such a change of coordinates. In a general learned feature space, however, specifying this transformation law is difficult: even under a simple rotation, different physical quantities may transform as scalars, vectors, or higher-order tensors, while neural activations may entangle these components.

For image-projected representations, this difficulty admits a much simpler resolution. Changing the ego frame does not move the physical cameras or alter the images they observe; it only changes the coordinates used to describe the cameras and the surrounding scene. Consequently, a representation associated with a fixed physical location on the image plane can remain unchanged under such a coordinate redefinition, even though the corresponding quantities in three-dimensional coordinates may transform nontrivially. We formalize this observation for general tensor-valued feature fields in Sec 3. Ego-frame symmetry can therefore be imposed on the image-projected representation without specifying an internal transformation law for the learned features. We refer to this property as *ego-frame invariance* (EFI). This separation between a geometric object and its coordinate realization is closely related to the motivation underlying gauge-equivariant representation design (Cohen et al., 2019).

Prior works have explored relative camera conditioning for novel-view synthesis (Safin et al., 2023; Miyato et al., 2024; Kong et al., 2024; Li et al., 2025), with PRoPE further evaluated on stereo depth estimation and spatial reasoning (Li et al., 2025). In particular, *Cameras as Relative Positional Encoding* studies camera-conditioning mechanisms whose dependence on a global reference frame cancels algebraically, yielding a model that is invariant to a common redefinition of the camera coordinate system (Li et al., 2025). In the tasks considered there, both the visual input and the desired output are themselves unchanged by such a redefinition, so this invariance is imposed at the level of the overall input–output mapping. Our question is different: we ask whether ego-frame invariance is desirable specifically as a property of an intermediate representation even when the downstream output is not invariant. We do not expect EFI to be universally advantageous for a narrow, specialized task, since an ego-frame dependent representation may be equally or more effective on a narrow task by exploiting task-specific regularities. This leads to our central hypothesis: beyond ego-frame-invariant tasks, EFI should be more beneficial for broader downstream tasks that demand a more general representation.

We study this hypothesis using 3D object detection, which provides a natural ego-frame equivariant task. Object positions and orientations are expressed in ego coordinates and therefore transform systematically when the ego frame is changed. To examine the connection between EFI and representation generality, we vary the semantic breadth of the detection task through nested sets of target classes. Our hypothesis is not that EFI must outperform a specialized representation on every detection problem, but rather that its advantage should become more pronounced as the representation is required to support a broader range of predictions. Our primary comparison is between Plücker Raymap, an ego-frame-dependent absolute camera encoding, and PRoPE, an ego-frame-invariant relative conditioning scheme. We additionally evaluate GTA+CamRay as a second ego-frame-invariant relative conditioning method to test whether the observed trend is specific to PRoPE. These conditioning schemes are described in section 4, and we measure how their performance relative to the absolute baseline changes as the number of target classes increases.

Our contributions are twofold. 1) We formulate the hypothesis that ego-frame invariance is more beneficial for broader ego-frame-equivariant tasks, motivated by the invariance of image projections to arbitrary ego-frame choices. 2) We provide empirical evidence supporting this hypothesis across multiple camera-conditioning schemes and datasets.

2 Related Works

Absolute and relative camera conditioning. Multiview models commonly condition image tokens on absolute camera rays, represented by ray origins and directions or by Plücker coordi-

nates (Sajjadi et al., 2022; Suhail et al., 2022; Li et al., 2025). Because these quantities are expressed in a selected global or ego frame, a common change of that frame changes the geometric encoding despite leaving the images and physical camera configuration unchanged. Relative camera-conditioning methods instead use pairwise camera relations, thereby removing dependence on a globally selected reference frame. RePAST introduces relative-pose attention for scene representation transformers (Safin et al., 2023), while GTA and EscherNet develop related geometry-aware attention mechanisms for multi-view transformers and view synthesis (Miyato et al., 2024; Kong et al., 2024).

The work most closely related to ours is *Cameras as Relative Positional Encoding* (Li et al., 2025), which systematically compares token-level raymaps with attention-level relative conditioning and introduces Projective Positional Encoding (PRoPE). Unlike earlier relative-pose formulations, PRoPE incorporates both camera intrinsics and extrinsics through pairwise projective transformations. It is evaluated on novel-view synthesis, stereo depth estimation, spatial cognition, and generalization to unseen camera intrinsics and numbers of input views. In these settings, the desired prediction remains unchanged under a common redefinition of the external coordinate frame, so invariance is imposed at the level of the overall input–output mapping. Our setting instead separates the symmetry of the intermediate representation from that of the output: the encoder can be ego-frame invariant while ego-centric 3D detection must remain ego-frame equivariant. We further study whether the benefit of relative conditioning changes as the downstream task is broadened.

Symmetry in intermediate representations. Geometric deep learning incorporates known transformation structure into neural architectures rather than requiring it to be learned entirely from data (Cohen & Welling, 2016; Kondor & Trivedi, 2018; Bronstein et al., 2021). This principle has motivated rotation- and translation-equivariant models for images, point clouds, and geometric fields (Fuchs et al., 2020; Satorras et al., 2021). Gauge-equivariant networks further emphasize the distinction between a geometric object and its numerical realization under an arbitrary choice of local reference frame (Cohen et al., 2019).

Most symmetry-aware architectures prescribe how representations or outputs transform under a known group action. Our setting instead exploits a case in which the desired transformation of the image-projected representation reduces to invariance, even though the downstream prediction remains equivariant. We impose this invariance only on the camera-conditioned encoder and retain ego-frame-dependent geometry in the detection head. Rather than treating symmetry as universally beneficial, we study whether this stage-wise separation becomes more useful as the downstream task broadens and under ego-frame realizations outside the training distribution.

3 Ego-frame Invariance of Image-projected Representations

This section provides a theoretical motivation for our hypothesis. Our discussion begins from the premise that a general representation should faithfully encode properties of the physical world that can be regarded as existing independently of the particular coordinate conventions used to describe them. Under this premise, changing an arbitrary ego coordinate frame should alter only the numerical description of the scene, rather than the underlying physical content represented.

More concretely, suppose that the physical content of a scene can be faithfully represented by an ideal semantic-geometric feature field over space. This field need not be explicitly reconstructed by a neural network; it serves as a mathematical description of the semantic and geometric information contained in the scene.

Let $\boldsymbol{x} \in \mathbb{R}^3$ denote the coordinates of a physical point in an ego frame. Consider a passive change of ego coordinates

$$\boldsymbol{x}' = \boldsymbol{R}\,\boldsymbol{x} + \boldsymbol{t}, \qquad g = (\boldsymbol{R}, \boldsymbol{t}) \in SE(3)\,, \tag{1}$$

with homogeneous transformation $\bar{\boldsymbol{x}}' = \boldsymbol{T}_g \bar{\boldsymbol{x}}$. For concreteness, let $F^{\mu i_1 \cdots i_n}(\boldsymbol{x})$ denote the components of the field, where μ indexes scalar semantic channels and $i_1, \ldots, i_n$ are spatial tensor indices. Its coordinate components transform as

$$F'^{\mu i_1 \cdots i_n}(\boldsymbol{x}') = R^{i_1}{}_{j_1} \cdots R^{i_n}{}_{j_n} F^{\mu j_1 \cdots j_n}(\boldsymbol{x})\,. \tag{2}$$

Thus, the three-dimensional field is equivariant to the change of ego coordinates.

Let $\boldsymbol{E}$ map homogeneous ego coordinates to a fixed physical camera, and let $\boldsymbol{K}$ be its intrinsic matrix. The image coordinate of $\boldsymbol{x}$ is

$$\boldsymbol{y} = \pi_{\boldsymbol{E}}(\boldsymbol{x}) = s(\boldsymbol{K}\boldsymbol{E}\bar{\boldsymbol{x}})\,, \tag{3}$$

where s denotes perspective division. Because the camera is physically unchanged, its extrinsic matrix in the new ego frame is

$$\boldsymbol{E}' = \boldsymbol{E}\boldsymbol{T}_g^{-1}. \tag{4}$$

Consequently,

$$\begin{aligned}\pi_{\boldsymbol{E}'}(\boldsymbol{x}') &= s(\boldsymbol{K}\boldsymbol{E}'\bar{\boldsymbol{x}}') \\ &= s\big(\boldsymbol{K}\boldsymbol{E}\boldsymbol{T}_g^{-1}\boldsymbol{T}_g\bar{\boldsymbol{x}}\big) \\ &= \pi_{\boldsymbol{E}}(\boldsymbol{x}).\end{aligned} \tag{5}$$

The same physical point therefore remains at the same image coordinate. For a physical point $\boldsymbol{x}$ projected to $\boldsymbol{y} = \pi_{\boldsymbol{E}}(\boldsymbol{x})$, define the corresponding image-projected semantic-geometric field as

$$L^{\mu\alpha_1\ldots\alpha_n}(\boldsymbol{y}) = J_{\pi}^{\alpha_1}{}_{i_1}(\boldsymbol{x})\cdots J_{\pi}^{\alpha_n}{}_{i_n}(\boldsymbol{x})F^{\mu i_1\ldots i_n}(\boldsymbol{x}), \tag{6}$$

where $\alpha_1, \ldots, \alpha_n$ denote image-coordinate indices. Under the ego-frame change, the projection Jacobian satisfies

$$\boldsymbol{J}_{\pi'}(\boldsymbol{x}') = \boldsymbol{J}_{\pi}(\boldsymbol{x})\boldsymbol{R}^{-1}. \tag{7}$$

Each factor of $\boldsymbol{R}^{-1}$ from the projection Jacobian therefore cancels the corresponding factor of $\boldsymbol{R}$ in Equation equation 2. Since $\boldsymbol{y}' = \boldsymbol{y}$ from Equation equation 5, the projected field satisfies

$$\boxed{L'^{\mu\alpha_1\ldots\alpha_n}(\boldsymbol{y}) = L^{\mu\alpha_1\ldots\alpha_n}(\boldsymbol{y})}. \tag{8}$$

Although derived for an ideal image-projected field rather than an arbitrary network activation, this result suggests that a learned representation of physical observations should not depend on an arbitrary ego-frame convention. We regard this invariance as a structural property of general representations: a specialized representation may exploit a fixed ego frame, but such dependence should become less useful as task breadth grows. This motivates our hypothesis that EFI becomes increasingly beneficial for broader tasks, even when their outputs are EFE.

4 Encoder Architecture and Camera Conditioning

To study the effect of EFI on the intermediate representation, we use a common encoder–head architecture across all conditioning methods. The encoder produces image-aligned multi-camera features, which are passed to a BEVDepth-based detection head (Li et al., 2023). The detection head necessarily receives ego-frame camera geometry to construct an ego-centric BEV representation and predict ego-frame 3D boxes. In contrast, we vary whether the encoder representation itself depends on the choice of ego frame.

To avoid introducing architecture-specific geometric structure into the encoder, we use a CNN only to tokenize each image, followed by standard self-attention across the resulting multi-camera tokens. Camera geometry is supplied through the positional conditioning of these tokens. We compare three conditioning schemes studied in *Cameras as Relative Positional Encoding* (Li et al., 2025): Plücker Raymap, GTA+CamRay, and PRoPE. The encoder and detection head contain approximately 20M and 10M parameters, respectively; for a given dataset and class configuration, the detection head is identical across conditioning methods, while Plücker Raymap and GTA+CamRay include an additional token-embedding module, adding approximately 0.7M encoder parameters relative to PRoPE.

Plücker Raymap. For each image token, let $\boldsymbol{o}_i$ and $\boldsymbol{d}_i^{u,v}$ denote the camera origin and the corresponding unit ray direction expressed in the ego frame. Its Plücker ray is

$$\boldsymbol{p}_i^{u,v} = [\boldsymbol{o}_i \times \boldsymbol{d}_i^{u,v}, \boldsymbol{d}_i^{u,v}]\,. \tag{9}$$

We embed this ray and add it to the corresponding image token before self-attention. Because both the ray origin and direction are expressed in the selected ego frame, this conditioning changes under a common change of ego coordinates and therefore provides an ego-frame-dependent, absolute camera encoding.

Table 1: Comparison of the camera-conditioning methods used in this work. Plücker raymaps encode camera geometry in the selected ego frame, whereas GTA+CamRay and PRoPE use ego-frame-invariant conditioning.

Method	Extrinsic relation	Intrinsic encoding	Injection	EFI
Plücker Raymap	Absolute	Yes	Token-level	No
PRoPE	Relative	Yes	Attention-level	Yes
GTA+CamRay	Relative	Yes (CamRay)	Attention + token	Yes

GTA+CamRay. Geometric Transform Attention (GTA) (Miyato et al., 2024) incorporates camera geometry directly into self-attention through relative $SE(3)$ transformations between camera pairs. Since a common change of ego frame cancels in the pairwise transformation, GTA does not depend on the particular realization of the ego frame. Standard GTA, however, encodes camera extrinsics but not intrinsics. We therefore follow Li et al. (2025) and supplement GTA with CamRay, a pixel-aligned ray direction expressed in the local camera frame,

$$\boldsymbol{d}_{i,\mathrm{cam}}^{u,v} \propto \boldsymbol{K}_i^{-1}[u, v, 1]^\top. \tag{10}$$

CamRay supplies token-level intrinsic information without introducing dependence on the ego frame.

PRoPE. Projective Positional Encoding (PRoPE) (Li et al., 2025) extends relative attention from camera poses to complete camera frustums. Let $\widetilde{\boldsymbol{P}}_i$ denote the homogeneous projection matrix of camera i, containing both its intrinsics and extrinsics. PRoPE conditions attention on the pairwise projective transformation

$$\widetilde{\boldsymbol{P}}_i \widetilde{\boldsymbol{P}}_j^{-1}, \tag{11}$$

which describes the relative projective geometry between two views. A common change of ego frame appears in both projection matrices and cancels in this product, making the conditioning ego-frame invariant. Unlike GTA, PRoPE therefore incorporates both intrinsic and extrinsic geometry directly at the attention level without requiring an additional token-level camera encoding.

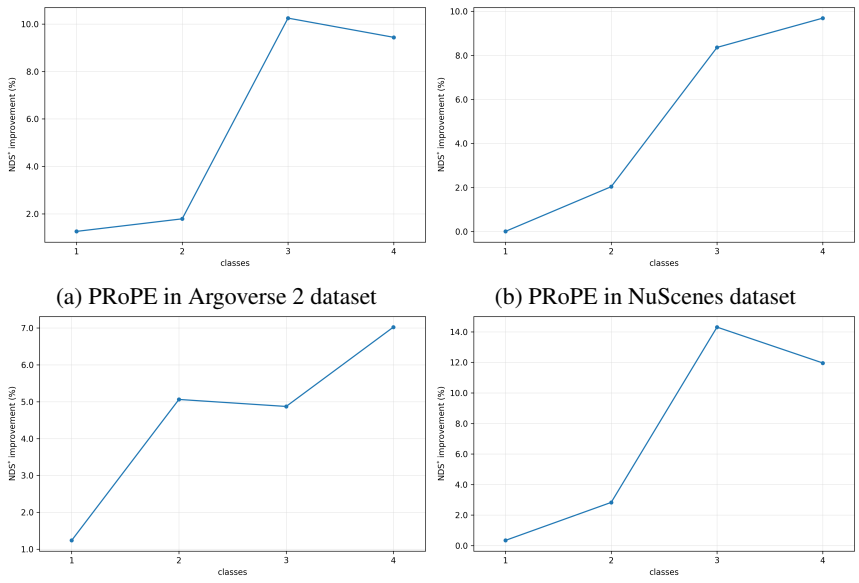
5 Experiment

Task design. Although 3D object detection is intrinsically EFE, standard datasets express annotations in a fixed ego coordinate frame. We explicitly vary this frame during training and held-out test evaluation. For each sample, we apply a random yaw over the full 360° range and a planar translation with displacement in $[0, 0.5]\,\mathrm{m}$. Camera images remain unchanged, while camera geometry and all ego-frame-dependent targets are re-expressed in the sampled coordinate frame.

To keep the target set unchanged under this transformation, objects are selected within $55\,\mathrm{m}$ before randomization and retained afterward without re-filtering. The detection range is set to $60\,\mathrm{m}$ to provide a guard region, and the translation magnitude is kept small to avoid crop-induced directional bias.

We additionally apply random left–right reflection to suppress fixed handedness and scene-layout conventions in the datasets. Unlike ego-frame randomization, reflection is treated as a transformation of the physical scene rather than a passive coordinate change; images, camera geometry, and detection targets are mirrored consistently. Reflection also increases the overall difficulty of the task by removing an additional dataset-specific shortcut. We include it to avoid an overly easy regime in which different camera-conditioning schemes may perform similarly, making differences in the learned representation difficult to observe.

Datasets and semantic breadth. Experiments are conducted independently on nuScenes (Caesar et al., 2020) and the Argoverse 2 Sensor Dataset (Wilson et al., 2021). For both datasets, we construct nested detection tasks with 4, 3, 2, and 1 semantic superclasses. Starting from the 4-class configuration, we progressively remove the superclass with the lowest class-wise AP to obtain narrower tasks. This avoids a confound in which reducing the number of classes could inadvertently



(a) PRoPE in Argoverse 2 dataset (b) PRoPE in NuScenes dataset

(c) GTA+CamRay in Argoverse 2 dataset (d) GTA+CamRay in NuScenes dataset

Figure 1: Relative NDS^* improvement over Plücker Raymap as task breadth increases for PRoPE / GTA+CamRay in Argoverse 2 dataset (a)/(c) and Nuscenes dataset (b)/(d).

increase task difficulty by removing easier categories, making class count a cleaner control of semantic breadth. Each dataset is mapped independently from its native taxonomy to these superclasses; complete mappings are provided in Appendix A.

To reduce the effect of class imbalance, training samples are selected according to the class distribution of the 4-class configuration. Specifically, frames are greedily selected by prioritizing the currently least represented sampling class according to cumulative box counts until the target number of training frames is reached. For both Argoverse 2 and nuScenes, sampling is performed once based on the 4-class setting, and the exact same set of sampled training examples is reused for the 3-, 2-, and 1-class tasks. This keeps the underlying training samples fixed across different levels of task breadth.

Because the official test annotations are unavailable for our modified evaluation protocol, each official validation split is divided by episode-index parity into disjoint validation and held-out test subsets. Model selection is performed only on the validation subset.

Evaluation. We report mean Average Precision (mAP) and the modified nuScenes Detection Score NDS^* introduced by CoIn3D (Kuang et al., 2026). It combines mAP with mean Average Translation Error (mATE), mean Average Scale Error (mASE), and mean Average Orientation Error (mAOE):

$$\mathrm{NDS}^* = \frac{1}{6}\left[3\,\mathrm{mAP} + \sum_{e\in\mathcal{E}}(1-\min(1,e))\right], \qquad \mathcal{E} = \{\mathrm{mATE}, \mathrm{mASE}, \mathrm{mAOE}\}. \tag{12}$$

Compared with the original nuScenes NDS (Caesar et al., 2020), NDS^* excludes velocity and attribute errors, allowing the same aggregate measure of detection and geometric box quality to be used for both datasets. This is useful in our setting because ego-frame equivariance directly affects frame-dependent translation and orientation.

All models are trained for 50 epochs, and the checkpoint with the highest NDS^* on the validation split is selected for evaluation. Ego-frame randomization and reflection are disabled during both validation and held-out test evaluation, so all checkpoints and final results are measured under the same canonical ego-frame configuration.

Results. Figure 1 presents the main results. On both nuScenes and Argoverse 2, the relative advantage of ego-frame-invariant conditioning is small for the narrowest tasks and becomes substantially larger for broader tasks, although the trend is not strictly monotonic across all configurations. On nuScenes, the relative NDS^* improvement of PRoPE over Plücker Raymap is 0.01%, 2.04%, 8.36%, and 9.69% for one through four classes, respectively. Argoverse 2 shows a similar separation, with improvements of 1.26%, 1.79%, 10.25%, and 9.44%. The same overall pattern is observed with GTA+CamRay: on nuScenes, the corresponding gains are 0.34%, 2.83%, 14.31%, and 11.96%, while on Argoverse 2 they are 1.24%, 5.06%, 4.87%, and 7.02%.

These results do not indicate that the benefit of ego-frame invariance must increase monotonically with every additional class. Rather, they show a consistent shift from little or modest advantage on narrow tasks to substantially larger gains on broader tasks. With model capacity and training budget held fixed, the effect of adding a single class can still depend on dataset-specific difficulty and on the particular strengths and weaknesses of each camera-conditioning scheme. Such configuration-specific effects can dominate individual transitions, even when the broader trend remains consistent. We therefore interpret the results in terms of the overall separation between narrow and broad tasks rather than requiring strict step-by-step monotonicity. Across both datasets and both relative conditioning schemes, this separation supports the hypothesis that ego-frame-invariant representations become more beneficial as the downstream prediction task broadens.

Tables 4, 5 in Appendix. B report the corresponding absolute NDS^*, mAP, mAOE, mATE, and mASE values for each configuration, together with class-wise AP, AOE, ATE, and ASE for every included class.

6 Conclusion

Prior work has demonstrated the benefits of ego-frame invariant camera conditioning for tasks whose outputs are themselves ego-frame invariant. We showed theoretically that a general image-projected representation should also be independent of an arbitrary choice of ego coordinates. However, not every downstream task requires a general representation. We therefore hypothesized that ego-frame invariance would remain beneficial for ego-frame-dependent prediction when the downstream task is sufficiently broad.

We tested this hypothesis through 3D object detection with progressively increasing numbers of target classes. Across nuScenes and Argoverse2, ego-frame-invariant conditioning shows little or modest advantage on narrow tasks but substantially larger gains on broader ones. Although the improvement is not strictly monotonic for every increase in task breadth, the overall separation between narrow and broad settings is consistent across datasets and conditioning schemes. These results extend the relevance of ego-frame invariance beyond invariant tasks, suggesting that it can also benefit ego-frame dependent prediction when broader representations are required.

Additionally, the results indicate a practical advantage of relative camera conditioning itself. Replacing absolute conditioning with relative conditioning introduces little additional computational or architectural cost; in terms of parameter count, PRoPE is even smaller than the Plücker Raymap baseline. Despite this, relative conditioning improves NDS^* by up to 14.31% in our experiments. Independent of our broader argument about representation generality, these results show that relative conditioning can provide substantial performance gains over absolute conditioning at essentially no additional model cost.

Limitations. Our study approximates task breadth primarily through semantic class expansion and is limited to two datasets and one downstream task. The observed relationship is also not strictly monotonic across all class configurations. More generally, controlling experimental variables across different levels of task breadth is difficult. We keep model capacity, training budget, and hyperparameters fixed across datasets, class configurations, and camera-conditioning schemes in order to provide a common comparison. However, using identical settings does not necessarily imply an equally favorable or equally converged training regime for every configuration. In particular, fixing training to 50 epochs across all levels of task breadth may affect models differently as the number and difficulty of target classes increase. Dataset-specific properties and the particular strengths and weaknesses of each conditioning scheme may likewise influence individual transitions between adjacent breadth settings. Consequently, we interpret the results as evidence of an overall shift between

narrow and broad tasks rather than as a strictly monotonic relationship for every additional class. Future work should examine this relationship under broader notions of task breadth, alternative training budgets and model capacities, additional conditioning mechanisms, architectures, camera rigs, and ego-frame-dependent tasks.

AI USE STATEMENT

In this work, we used generative AI tools to assist with language editing, manuscript organization, literature search, and LaTeX/BibTeX preparation. Generative AI was also used as a discussion aid for refining the presentation of research ideas and experimental motivations. The research hypotheses, experimental design, implementation, experiments, data analysis, and scientific conclusions were developed and verified by the authors. All AI-assisted text, references, and technical statements were reviewed by the authors, and cited works were checked against their original sources. We take responsibility for the final content of this work, including any text or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work studies representation design for multi-camera 3D perception using publicly available autonomous-driving datasets. We do not collect new data or conduct experiments involving human subjects. Our experiments use Argoverse 2 and nuScenes according to their respective terms of use. While the methods studied here may eventually be relevant to safety-critical autonomous systems, this work evaluates representation properties in offline perception benchmarks and does not establish safety or reliability for real-world deployment.

REPRODUCIBILITY STATEMENT

We describe the model architecture, camera-conditioning methods, ego-frame randomization, dataset construction, training protocol, and evaluation procedure in the main paper. The complete implementation, including all configuration files and experimental settings, will be released publicly on GitHub in the form used to produce the reported results. The released code will therefore provide the exact implementation details needed to reproduce the experiments beyond the methodological description given in the paper.

REFERENCES

Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Mojtaba Komeili, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zholus, Sergio Arnaud, Abha Gejji, Ada Martin, Francois Robert Hogan, Daniel Dugas, Piotr Bojanowski, Vasil Khalidov, Patrick Labatut, Francisco Massa, Marc Szafraniec, Kapil Krishnakumar, Yong Li, Xiaodong Ma, Sarath Chandar, Franziska Meier, Yann LeCun, Michael Rabbat, and Nicolas Ballas. V-JEPA 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985*, 2025.

Michael M. Bronstein, Joan Bruna, Taco Cohen, and Petar Veličković. Geometric deep learning: Grids, groups, graphs, geodesics, and gauges. *arXiv preprint arXiv:2104.13478*, 2021.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11621–11631, 2020.

Taco Cohen and Max Welling. Group equivariant convolutional networks. In *Proceedings of the 33rd International Conference on Machine Learning*, pp. 2990–2999, 2016.

Taco Cohen, Maurice Weiler, Berkay Kicanaoglu, and Max Welling. Gauge equivariant convolutional networks and the icosahedral CNN. In *Proceedings of the 36th International Conference on Machine Learning*, pp. 1321–1330, 2019.

Fabian B. Fuchs, Daniel E. Worrall, Volker Fischer, and Max Welling. SE(3)-transformers: 3d roto-translation equivariant attention networks. In *Advances in Neural Information Processing Systems*, volume 33, pp. 1970–1981, 2020.

Dibya Ghosh, Homer Rich Walke, Karl Pertsch, Kevin Black, Oier Mees, Sudeep Dasari, Joey Hejna, Tobias Kreiman, Charles Xu, Jianlan Luo, You Liang Tan, Lawrence Yunliang Chen, Quan Vuong, Ted Xiao, Pannag R. Sanketi, Dorsa Sadigh, Chelsea Finn, and Sergey Levine. Octo: An open-source generalist robot policy. In *Proceedings of Robotics: Science and Systems*, 2024.

Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan P. Foster, Pannag R. Sanketi, Quan Vuong, Thomas Kollar, Benjamin Burchfiel, Russ Tedrake, Dorsa Sadigh, Sergey Levine, Percy Liang, and Chelsea Finn. OpenVLA: An open-source vision-language-action model. In *Proceedings of the 8th Conference on Robot Learning*, volume 270, pp. 2679–2713, 2025.

Risi Kondor and Shubhendu Trivedi. On the generalization of equivariance and convolution in neural networks to the action of compact groups. In *Proceedings of the 35th International Conference on Machine Learning*, pp. 2747–2755, 2018.

Xin Kong, Shikun Liu, Xiaoyang Lyu, Marwan Taher, Xiaojuan Qi, and Andrew J. Davison. EscherNet: A generative model for scalable view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9503–9513, 2024.

Zhaonian Kuang, Rui Ding, Haotian Wang, Xinhu Zheng, Meng Yang, and Gang Hua. CoIn3D: Revisiting configuration-invariant multi-camera 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 40707–40716, 2026.

Juho Lee, Yoonho Lee, Jungtaek Kim, Adam Kosiorek, Seungjin Choi, and Yee Whye Teh. Set transformer: A framework for attention-based permutation-invariant neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, pp. 3744–3753, 2019.

Ruilong Li, Brent Yi, Junchen Liu, Hang Gao, Yi Ma, and Angjoo Kanazawa. Cameras as relative positional encoding. In *Advances in Neural Information Processing Systems*, volume 38, pp. 18020–18045, 2025.

Yinhao Li, Zheng Ge, Guanyi Yu, Jinrong Yang, Zengran Wang, Yukang Shi, Jianjian Sun, and Zeming Li. BEVDepth: Acquisition of reliable depth for multi-view 3d object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 1477–1485, 2023.

Zhiqi Li, Wenhai Wang, Hongyang Li, Enze Xie, Chonghao Sima, Tong Lu, Yu Qiao, and Jifeng Dai. BEVFormer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers. In *European Conference on Computer Vision*, pp. 1–18, 2022.

Yingfei Liu, Tiancai Wang, Xiangyu Zhang, and Jian Sun. PETR: Position embedding transformation for multi-view 3d object detection. In *European Conference on Computer Vision*, pp. 531–548, 2022.

Yingfei Liu, Junjie Yan, Fan Jia, Shuailin Li, Aqi Gao, Tiancai Wang, and Xiangyu Zhang. PETRv2: A unified framework for 3d perception from multi-camera images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3262–3272, 2023.

Takeru Miyato, Bernhard Jaeger, Max Welling, and Andreas Geiger. GTA: A geometry-aware attention mechanism for multi-view transformers. In *International Conference on Learning Representations*, 2024.

NVIDIA, Johan Bjorck, Fernando Castañeda, Nikita Cherniadev, Xingye Da, Runyu Ding, Linxi Fan, Yu Fang, Dieter Fox, Fengyuan Hu, Spencer Huang, Joel Jang, Zhenyu Jiang, Jan Kautz, Kaushil Kundalia, Lawrence Lao, Zhiqi Li, Zongyu Lin, Kevin Lin, Guilin Liu, Edith Llontop, Loic Magne, Ajay Mandlekar, Avnish Narayan, Soroush Nasiriany, Scott Reed, You Liang Tan, Guanzhi Wang, Zu Wang, Jing Wang, Qi Wang, Jiannan Xiang, Yuqi Xie, Yinzhen Xu, Zhenjia Xu, Seonghyeon Ye, Zhiding Yu, Ao Zhang, Hao Zhang, Yizhou Zhao, Ruijie Zheng, and Yuke Zhu. GR00T N1: An open foundation model for generalist humanoid robots. *arXiv preprint arXiv:2503.14734*, 2025.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024.

Jonah Philion and Sanja Fidler. Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d. In *European Conference on Computer Vision*, pp. 194–210, 2020.

Aleksandr Safin, Daniel Duckworth, and Mehdi S. M. Sajjadi. RePAST: Relative pose attention scene representation transformer. *arXiv preprint arXiv:2304.00947*, 2023.

Mehdi S. M. Sajjadi, Henning Meyer, Etienne Pot, Urs Bergmann, Klaus Greff, Noha Radwan, Suhani Vora, Mario Lučić, Daniel Duckworth, Alexey Dosovitskiy, Jakob Uszkoreit, Thomas Funkhouser, and Andrea Tagliasacchi. Scene representation transformer: Geometry-free novel view synthesis through set-latent scene representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6229–6238, 2022.

Víctor Garcia Satorras, Emiel Hoogeboom, and Max Welling. E(n) equivariant graph neural networks. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 9323–9332, 2021.

Mohammed Suhail, Carlos Esteves, Leonid Sigal, and Ameesh Makadia. Light field neural rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8269–8279, 2022.

Gemini Robotics Team, Saminda Abeyruwan, Joshua Ainslie, Jean-Baptiste Alayrac, Montserrat Gonzalez Arenas, Travis Armstrong, Ashwin Balakrishna, Robert Baruch, Maria Bauza, Michiel Blokzijl, Steven Bohez, Konstantinos Bousmalis, Anthony Brohan, Thomas Buschmann, Arunkumar Byravan, Serkan Cabi, Ken Caluwaerts, Federico Casarini, Oscar Chang, Jose Enrique Chen, Xi Chen, Hao-Tien Lewis Chiang, Krzysztof Choromanski, David D'Ambrosio, Sudeep Dasari, Todor Davchev, Coline Devin, Norman Di Palo, Tianli Ding, Adil Dostmohamed, Danny Driess, Yilun Du, Debidatta Dwibedi, Michael Elabd, Claudio Fantacci, Cody Fong, Erik Frey, Chuyuan Fu, Marissa Giustina, Keerthana Gopalakrishnan, Laura Graesser, Leonard Hasenclever, Nicolas Heess, Brandon Hernaez, Alexander Herzog, R. Alex Hofer, Jan Humplik, Atil Iscen, Mithun George Jacob, Deepali Jain, Ryan Julian, Dmitry Kalashnikov, M. Emre Karagozler, Stefani Karp, Chase Kew, Jerad Kirkland, Sean Kirmani, Yuheng Kuang, Thomas Lampe, Antoine Laurens, Isabel Leal, Alex X. Lee, Tsang-Wei Edward Lee, Jacky Liang, Yixin Lin, Sharath Maddineni, Anirudha Majumdar, Assaf Hurwitz Michaely, Robert Moreno, Michael Neunert, Francesco Nori, Carolina Parada, Emilio Parisotto, Peter Pastor, Acorn Pooley, Kanishka Rao, Krista Reymann, Dorsa Sadigh, Stefano Saliceti, Pannag Sanketi, Pierre Sermanet, Dhruv Shah, Mohit Sharma, Kathryn Shea, Charles Shu, Vikas Sindhwani, Sumeet Singh, Radu Soricut, Jost Tobias Springenberg, Rachel Sterneck, Razvan Surdulescu, Jie Tan, Jonathan Tompson, Vincent Vanhoucke, Jake Varley, Grace Vesom, Giulia Vezzani, Oriol Vinyals, Ayzaan Wahid, Stefan Welker, Paul Wohlhart, Fei Xia, Ted Xiao, Annie Xie, Jinyu Xie, Peng Xu, Sichun Xu, Ying Xu, Zhuo Xu, Yuxiang Yang, Rui Yao, Sergey Yaroshenko, Wenhao Yu, Wentao Yuan, Jingwei Zhang, Tingnan Zhang, Allan Zhou, and Yuxiang Zhou. Gemini robotics: Bringing AI into the physical world. *arXiv preprint arXiv:2503.20020*, 2025.

Yue Wang, Vitor Campagnolo Guizilini, Tianyuan Zhang, Yilun Wang, Hang Zhao, and Justin Solomon. DETR3D: 3d object detection from multi-view images via 3d-to-2d queries. In *Proceedings of the 5th Conference on Robot Learning*, pp. 180–191, 2022.

Benjamin Wilson, William Qi, Tanmay Agarwal, John Lambert, Jagjeet Singh, Siddhesh Khandelwal, Bowen Pan, Ratnesh Kumar, Andrew Hartnett, Jhony Kaesemodel Pontes, Deva Ramanan, Peter Carr, and James Hays. Argoverse 2: Next generation datasets for self-driving perception and forecasting. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, 2021.

Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabás Póczos, Ruslan Salakhutdinov, and Alexander J. Smola. Deep sets. In *Advances in Neural Information Processing Systems*, volume 30, pp. 3391–3401, 2017.

A SUPERCLASS TAXONOMY AND CLASS MAPPINGS

The native object taxonomies of Argoverse 2 and nuScenes differ substantially. We therefore define dataset-specific mappings from the original categories to a smaller set of superclasses used in our experiments. The mappings are designed to preserve broadly corresponding object types within each dataset rather than to enforce an exact class correspondence across datasets. Detection tasks of different breadth are constructed by progressively adding these superclasses. Classes not assigned to any superclass are ignored during training and evaluation.

A.1 ARGOVERSE 2

Table 2 shows the mapping from Argoverse 2 native categories to the superclasses used in our experiments.

Table 2: Superclass mapping for Argoverse 2.

Superclass	Argoverse 2 categories
Car	`REGULAR_VEHICLE`
LargeVehicle	`LARGE_VEHICLE`, `BUS`, `SCHOOL_BUS` `TRUCK`, `BOX_TRUCK`
Pedestrian	`PEDESTRIAN`, `BICYCLIST` `MOTORCYCLIST`, `WHEELED_RIDER`
Cycle	`BICYCLE`, `MOTORCYCLE`, `WHEELED_DEVICE`, `WHEELCHAIR`, `STROLLER`

The nested Argoverse 2 tasks are defined as follows:

- 1 classes: Car.
- 2 classes: Car, LargeVehicle, .
- 3 classes: Car, LargeVehicle, Pedestrian.
- 4 classes: Car, LargeVehicle, Pedestrian, Cycle.

A.2 NUSCENES

Table 3 shows the corresponding mapping for nuScenes.

Table 3: Superclass mapping for nuScenes.

Superclass	nuScenes categories
Car	`vehicle.car`
Pedestrian	`human.pedestrian.adult`, `human.pedestrian.child`, `human.pedestrian.construction_worker`, `human.pedestrian.police_officer`
Cycle	`vehicle.bicycle`, `vehicle.motorcycle`
LargeVehicle	`vehicle.truck`, `vehicle.bus.bendy`, `vehicle.bus.rigid`, `vehicle.trailer`, `vehicle.construction`

The nested nuScenes tasks are defined as follows:

- 1 classes: Car.
- 2 classes: Car, Pedestrian.
- 3 classes: Car, Pedestrian, Cycle.
- 4 classes: Car, Pedestrian, Cycle, LargeVehicle.

Table 4: 3D object detection on Argoverse2 under ID ego-frame variation. ID uses ego translations of 0–0.5 m. Higher is better for NDS* and AP; lower is better for AOE, ATE, and ASE.

Classes	Target	Metric	PRoPE	GTA+CamRay	Plücker Raymap
4 classes	Overall	NDS ↑	0.232	0.227	0.212
	Overall	mAP ↑	0.153	0.149	0.145
	Overall	mAOE ↓	0.899	0.903	1.000
	Overall	mATE ↓	0.922	0.939	0.922
	Overall	mASE ↓	0.245	0.244	0.240
	Car	AP ↑	0.345	0.347	0.331
	Car	AOE ↓	0.483	0.560	0.736
	Car	ATE ↓	0.761	0.762	0.769
	Car	ASE ↓	0.195	0.188	0.192
	LargeVehicle	AP ↑	0.113	0.117	0.128
	LargeVehicle	AOE ↓	0.455	0.527	0.630
	LargeVehicle	ATE ↓	0.959	0.968	0.903
	LargeVehicle	ASE ↓	0.265	0.264	0.259
	Pedestrian	AP ↑	0.114	0.101	0.095
	Pedestrian	AOE ↓	1.419	1.406	1.400
	Pedestrian	ATE ↓	0.972	1.010	1.002
	Pedestrian	ASE ↓	0.240	0.238	0.237
	Cycle	AP ↑	0.040	0.034	0.027
	Cycle	AOE ↓	1.237	1.120	1.234
	Cycle	ATE ↓	0.996	1.014	1.015
	Cycle	ASE ↓	0.280	0.287	0.272
3 classes	Overall	NDS ↑	0.302	0.287	0.273
	Overall	mAP ↑	0.218	0.207	0.196
	Overall	mAOE ↓	0.750	0.796	0.807
	Overall	mATE ↓	0.869	0.876	0.913
	Overall	mASE ↓	0.226	0.227	0.228
	Car	AP ↑	0.380	0.362	0.356
	Car	AOE ↓	0.308	0.360	0.452
	Car	ATE ↓	0.720	0.742	0.748
	Car	ASE ↓	0.182	0.183	0.191
	LargeVehicle	AP ↑	0.149	0.146	0.128
	LargeVehicle	AOE ↓	0.475	0.539	0.517
	LargeVehicle	ATE ↓	0.928	0.929	0.973
	LargeVehicle	ASE ↓	0.257	0.259	0.253
	Pedestrian	AP ↑	0.126	0.112	0.104
	Pedestrian	AOE ↓	1.468	1.489	1.451
	Pedestrian	ATE ↓	0.960	0.957	1.018
	Pedestrian	ASE ↓	0.239	0.240	0.240
2 classes	Overall	NDS ↑	0.375	0.387	0.369
	Overall	mAP ↑	0.248	0.254	0.249
	Overall	mAOE ↓	0.428	0.377	0.438
	Overall	mATE ↓	0.834	0.844	0.860
	Overall	mASE ↓	0.233	0.218	0.237
	Car	AP ↑	0.368	0.369	0.359
	Car	AOE ↓	0.362	0.355	0.425
	Car	ATE ↓	0.723	0.732	0.743
	Car	ASE ↓	0.186	0.185	0.189
	LargeVehicle	AP ↑	0.129	0.139	0.139
	LargeVehicle	AOE ↓	0.493	0.398	0.452
	LargeVehicle	ATE ↓	0.945	0.957	0.977
	LargeVehicle	ASE ↓	0.279	0.251	0.286
1 class	Overall	NDS ↑	0.535	0.535	0.528
	Overall	mAP ↑	0.417	0.417	0.408
	Overall	mAOE ↓	0.202	0.201	0.199
	Overall	mATE ↓	0.670	0.671	0.685
	Overall	mASE ↓	0.172	0.172	0.171
	Car	AP ↑	0.417	0.417	0.408
	Car	AOE ↓	0.202	0.201	0.199
	Car	ATE ↓	0.670	0.671	0.685
	Car	ASE ↓	0.172	0.172	0.171

B FULL DATA

Here, we present the full data from the experiments described in Sec. 5, in two tables 4, 5.

Table 5: 3D object detection on nuScenes under ID ego-frame variation. ID uses ego translations of 0–0.5 m. Higher is better for NDS* and AP; lower is better for AOE, ATE, and ASE.

Classes	Target	Metric	PRoPE	GTA+CamRay	Plücker Raymap
4 classes	Overall	NDS ↑	0.249	0.254	0.227
	Overall	mAP ↑	0.187	0.184	0.173
	Overall	mAOE ↓	0.936	0.892	0.988
	Overall	mATE ↓	0.864	0.867	0.891
	Overall	mASE ↓	0.269	0.269	0.279
	Car	AP ↑	0.361	0.360	0.346
	Car	AOE ↓	0.485	0.434	0.579
	Car	ATE ↓	0.697	0.703	0.706
	Car	ASE ↓	0.175	0.174	0.180
	LargeVehicle	AP ↑	0.103	0.101	0.108
	LargeVehicle	AOE ↓	0.720	0.634	0.790
	LargeVehicle	ATE ↓	0.997	0.990	0.996
	LargeVehicle	ASE ↓	0.306	0.303	0.316
	Pedestrian	AP ↑	0.166	0.172	0.142
	Pedestrian	AOE ↓	1.485	1.427	1.407
	Pedestrian	ATE ↓	0.896	0.902	0.950
	Pedestrian	ASE ↓	0.298	0.308	0.306
	Cycle	AP ↑	0.117	0.103	0.096
	Cycle	AOE ↓	1.054	1.072	1.178
	Cycle	ATE ↓	0.868	0.873	0.914
	Cycle	ASE ↓	0.295	0.293	0.314
3 classes	Overall	NDS ↑	0.260	0.274	0.240
	Overall	mAP ↑	0.214	0.227	0.189
	Overall	mAOE ↓	0.992	0.969	1.083
	Overall	mATE ↓	0.828	0.813	0.865
	Overall	mASE ↓	0.260	0.252	0.263
	Car	AP ↑	0.348	0.370	0.335
	Car	AOE ↓	0.382	0.389	0.560
	Car	ATE ↓	0.716	0.702	0.721
	Car	ASE ↓	0.177	0.175	0.182
	Pedestrian	AP ↑	0.175	0.184	0.157
	Pedestrian	AOE ↓	1.367	1.471	1.476
	Pedestrian	ATE ↓	0.908	0.889	0.939
	Pedestrian	ASE ↓	0.305	0.302	0.296
	Cycle	AP ↑	0.118	0.126	0.076
	Cycle	AOE ↓	1.228	1.046	1.211
	Cycle	ATE ↓	0.860	0.847	0.934
	Cycle	ASE ↓	0.299	0.278	0.310
2 classes	Overall	NDS ↑	0.341	0.343	0.334
	Overall	mAP ↑	0.300	0.297	0.293
	Overall	mAOE ↓	0.858	0.836	0.867
	Overall	mATE ↓	0.761	0.757	0.770
	Overall	mASE ↓	0.236	0.240	0.241
	Car	AP ↑	0.396	0.398	0.387
	Car	AOE ↓	0.262	0.271	0.287
	Car	ATE ↓	0.652	0.643	0.661
	Car	ASE ↓	0.171	0.169	0.180
	Pedestrian	AP ↑	0.203	0.196	0.200
	Pedestrian	AOE ↓	1.454	1.401	1.446
	Pedestrian	ATE ↓	0.871	0.871	0.878
	Pedestrian	ASE ↓	0.301	0.311	0.303
1 class	Overall	NDS ↑	0.537	0.539	0.537
	Overall	mAP ↑	0.414	0.415	0.408
	Overall	mAOE ↓	0.218	0.210	0.207
	Overall	mATE ↓	0.632	0.633	0.629
	Overall	mASE ↓	0.171	0.171	0.168
	Car	AP ↑	0.414	0.415	0.408
	Car	AOE ↓	0.218	0.210	0.207
	Car	ATE ↓	0.632	0.633	0.629
	Car	ASE ↓	0.171	0.171	0.168