TinyBEV: Compact Temporal Fusion for Multi-View 3D Perception

Hongyu Ke[1] Jack Morris[1] Yongkang Liu[2] Satoshi Kitai [2] Kentaro Oguchi[2]
Yi Ding[1] Haoxin Wang[1]
[1]Georgia State University
[2]InfoTech Labs, Toyota Motor North America R&D
{hke3, jmorris116, yiding, haoxinwang}@gsu.edu
{yongkang.liu, satoshi.kitai, kentaro.oguchi}@toyota.com

Abstract

Multi-view camera-based 3D object detection through unified Bird's Eye View (BEV) representation has become popular for autonomous driving due to its low cost, but efficiently inferring precise spatial and temporal information from cameras alone remains a significant challenge. Transformer-based approaches have shown substantial performance improvements but have the drawback of quadratic memory complexity — making these architectures ill-suited for edge deployment. Recently, State Space Models (SSMs) offer a more favorable balance of computational efficiency and performance in 2D vision, suggesting that they could help here as well. We present TinyBEV, an efficient BEV framework for multi-view 3D perception. For spatial modeling, we replace cross attention with SSMs that fusing BEV and camera images with linear complexity. For temporal modeling, we adopt a lightweight, linear-complexity history-fusion scheme that uses explicit time conditioning and channel-level aggregation instead of cross-frame attention. Both fusion strategies follow small constant scaling with respect to history length and enabling edge-friendly deployment. Experiments on NuScenes datasets demonstrate that TinyBEV is comparable with other state-of-the-art methods across diverse visual perception metrics with advantages in computational efficiency.

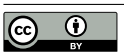
CCS Concepts

• **Computing methodologies → Scene understanding**.

Keywords

Birds' Eye View, Efficient AI

ACM Reference Format:
Hongyu Ke[1] Jack Morris[1] Yongkang Liu[2] Satoshi Kitai [2] Kentaro Oguchi[2] , Yi Ding[1] Haoxin Wang[1], [1]Georgia State University, [2]InfoTech Labs, Toyota Motor North America R&D, {hke3, jmorris116, yiding, haoxinwang}@gsu.edu, {yongkang.liu, satoshi.kitai, kentaro.oguchi}@toyota.com . 2025. TinyBEV: Compact Temporal Fusion for Multi-View 3D Perception. In *The Tenth ACM/IEEE Symposium on Edge Computing (SEC '25), December 3–6, 2025, Arlington, VA, USA.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3769102.3774633



SEC '25, Arlington, VA, USA

ACM ISBN 979-8-4007-2238-7/2025/12
https://doi.org/10.1145/3769102.3774633

1 Introduction

The Bird's-Eye-View (BEV) representation provides a unified view of the world, which has become the standard in modern autonomous systems [1, 5, 9, 34]. This representation is critical for downstream tasks such as 3D object detection [8, 10, 12, 15–19, 32]. Recently, generating BEV representation only from multi-view cameras has made significant strides, largely attributed to Transformer-based architectures [14, 20, 22, 31, 33] that leverage cross-attention to link BEV queries with image features. Despite their effectiveness, the attention mechanism in such models incurs quadratic computational and memory complexity as the number of image tokens or BEV queries grows. On the other hand, when dealing with real-world scenarios that involve high-resolution inputs from multiple cameras and frames, the quadratic computational complexity becomes a major bottleneck, significantly limiting feasibility on resource constrained automotive platforms [13, 26]. To address this, we replace the cross-attention used for constructing BEV features with state space models (SSMs) [4] to retain long-range spatial context with near-linear scaling, and we use a lightweight convolutional recurrent module for temporal fusion to keep overhead modest.

State space models (SSMs)—especially Mamba, which introduced the selective state space model (S6)—offer linear-time sequence processing, achieving strong results in language [3, 23, 24] and increasingly competitive performance in 2D vision [7, 21, 35] where they approximate attention at much lower cost. These properties make SSMs well suited for the spatial association between camera features and the BEV grid, where sequences aggregated across multiple cameras can be very long. Following recent BEV work [12], we therefore replace cross-attention with SSMs for spatial integration, retaining long-range dependencies while avoiding quadratic growth.

Temporal structure is equally critical in driving scenes, and it is essential to consider computational efficiency alongside construction accuracy in mobile edge [11, 25, 27, 28]. Naive extensions that appear across long frame stacks or concatenate many frames along channels quickly degrade efficiency. Motivated by [6], we favor lightweight convolutional recurrence in BEV space over cross-frame attention. This enables linear scaling with history length, which is essential for deployment on resource-constrained automotive platforms.

2 Related Work

Temporal fusion. Temporal fusion is central to camera-only BEV perception. Attention-based temporal fusion [18] lets current BEV queries attend to an ego-aligned history. This approach enables

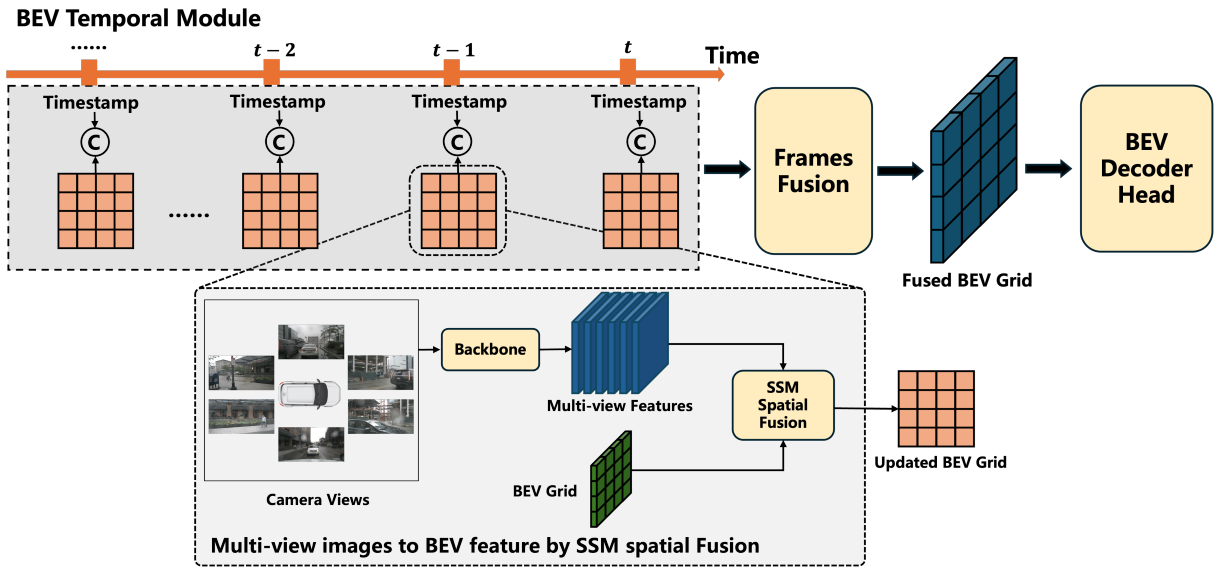


Figure 1: ***Bottom:*** **Multi-view images are encoded and aggregated into a BEV grid using an SSM-based** ***spatial fusion*** **module that replaces Transformer cross-attention.** ***Top:*** **A** ***parallel, attention-free temporal*** **module augments each frame's BEV with a scalar timestamp τ_t, applies a shared 1×1 encoder per frame, concatenates the encoded frames along channels, and performs per–BEV-query pointwise fusion to obtain a temporally consistent BEV grid.**

long-range reasoning, but incurs the quadratic cost in the number of BEV queries and large attention maps. Other work such as [30] warps historical BEV to the current frame, concatenates channels, and fuses with residual convolutions. This avoids attention maps and scales roughly linearly in queries and history, but couples neighboring queries through spatial kernels and typically omits explicit real-time spacing. VideoBEV [6] advocates simple BEV space mixers over cross-frame attention to improve deployability. In contrast, our module is parallel and attention-free. We additionally put a scalar timestamp per frame to expose both order and real time gaps, improving robustness to variable frame rates and multi-camera asynchrony.

State Space Models (SSMs). SSMs have emerged as a compelling alternative to Transformers, offering linear time complexity for sequence modeling. The Structured State Space Sequence (S4) model was a foundational work, demonstrating that SSMs could effectively capture long-range dependencies. A major advancement came with Mamba, which introduced the selective state space model (S6). Adapting these inherently 1D models to 2D vision tasks presents challenges. Approaches like Vision Mamba (Vim) [35] and VMamba [21] treat an image as a sequence of flattened patches. To preserve spatial relationships, they often employ a multi-scan strategy, processing the data multiple times with different flattening orders (e.g., row-wise and column-wise). The application of SSMs to BEV perception is a nascent but growing field. Recent works like MamBEV [12], which replaces the standard cross-attention in Transformer-based models with an SSM-based cross-attention module, retaining the separate query-and-pull fusion paradigm but with linear complexity. This supports us to implement the more efficient encoder/decoder for BEV.

3 Method

3.1 Temporal Fusion

Temporal fusion aims to aggregate a sequence of BEV features into a temporally consistent representation while respecting edge constraints and multi-camera asynchrony. We adopt a *parallel, attention-free* design that is highly parallelizable on modern hardware and whose dominant compute/memory are decoupled from the history length. Classical designs either (i) apply cross-frame attention, which suffers from high computation complexity, or (ii) maintain a recurrent hidden state, which is causal but less parallel. We take a different route: a *parallel, attention-free* mixer that (a) encodes each frame independently, (b) concatenates features along the channel axis under a fixed budget, and (c) fuses per *BEV query* using only pointwise operations. This yields high hardware parallelism and keeps the dominant compute decoupled from the number of history frames.

Let $\{X_t\}_{t=0}^{T}$ be BEV feature maps produced by a shared BEV encoder for the current frame t=0 and T historical frames:

$$X_t \in \mathbb{R}^{B\times C_{\text{bev}}\times H\times W}, \qquad Q = HW \text{ BEV queries per frame.} \tag{1}$$

To encode order and real time gaps, each frame receives a scalar timestamp $\tau_t = \rho\, t$ with $\rho > 0$, broadcast and concatenated as one extra channel:

$$\tilde{X}_t = \mathrm{Cat}\big[X_t,\ \tau_t \mathbf{1}\big] \in \mathbb{R}^{B\times(C_{\text{bev}}+1)\times H\times W}. \tag{2}$$

A shared pointwise block processes each frame independently:

$$Z_t = \phi\Big(\mathrm{BN}\big(W_1 * \tilde{X}_t\big)\Big), \tag{3}$$

where $W_1 \in \mathbb{R}^{C_{\text{bev}} \times (C_{\text{bev}}+1) \times 1 \times 1}$, and $\phi = \text{ReLU}$. Then we concatenate all frames along the channel:

$$Z = \text{Cat}[Z_0, \ldots, Z_T] \in \mathbb{R}^{B \times C_{\text{cat}} \times H \times W}. \tag{4}$$

Temporal mixing at each *BEV grid* is then done by another pointwise block:

$$U = \phi\Big(\text{BN}\big(W_2 * Z\big)\Big), \tag{5}$$

where $W_2 \in \mathbb{R}^{C_{\text{cat}} \times C_{\text{cat}} \times 1 \times 1}$. Then follow a per-query projection and LayerNorm:

$$\begin{aligned} U' &= \text{Reshape}(U) \in \mathbb{R}^{B \times Q \times C_{\text{cat}}}, \\ Y_{b,q} &= \text{LN}\big(U'_{b,q} W_P\big), \end{aligned} \tag{6}$$

where $W_P \in \mathbb{R}^{C_{\text{cat}} \times D}$, $\quad Y \in \mathbb{R}^{B \times Q \times D}$. Eqs. (2)–(6) constitute a single-pass, parallel, attention-free fusion; no recurrent state is maintained.

BEV query view. Index a *BEV query* by $q \in \{1, \ldots, Q\}$ and denote the per-frame BEV grid feature by $x_{t,q} \in \mathbb{R}^{C_{\text{bev}}}$. With (2) and (3),

$$\begin{aligned} \tilde{x}_{t,q} &= [x_{t,q};\, \tau_t] \in \mathbb{R}^{C_{\text{bev}}+1}, \\ z_{t,q} &= \phi(W_1 \tilde{x}_{t,q}) \in \mathbb{R}^{C_{\text{bev}}}, \\ z_q &= [z_{0,q}; \ldots; z_{T,q}] \in \mathbb{R}^{C_{\text{cat}}}. \end{aligned}$$

Then (5) and (6) reduce to a per-query feed-forward mixer:

$$y_q = \text{LN}\Big(\phi(W_2\, z_q)\, W_P\Big). \tag{7}$$

Thus, temporal fusion preserves *per-query independence*: the block mixes *time* but not *space*.

Why time conditioning? The timestamp channel is crucial beyond a mere positional tag.

(i) Order & spacing identifiability. Without τ_t, frames differ only by their block index inside z_q, which encodes order but not *physical spacing*. Assume $x_{t,q}$ evolves smoothly from a latent process sampled at interval Δt; a first-order expansion gives

$$x_{t+\delta,q} \;\approx\; x_{t,q} + \frac{\partial x_q}{\partial t}\,\delta. \tag{8}$$

Changing FPS or encountering dropped/late frames alters $\delta = \Delta t$; a mapping that depends only on block indices cannot re-scale its response by Δt, yielding miscalibrated velocity/temporal geometry. Providing $\tau_t = \rho t$ exposes real time gaps, enabling the network to learn functions that are (approximately) equivariant to time rescalings $t \mapsto \alpha t$ because τ changes accordingly.

(ii) Breaking frame-exchange symmetry. Eq. (3) shares W_1 across t. Injecting τ_t induces a *time-gated affine* encoder per BEV query,

$$z_{t,q} = \phi\big(A\, x_{t,q} + b(\tau_t)\big), \tag{9}$$

where $b(\tau_t)$ (realized via the extra channel and biases) adapts the response by temporal position/spacing and stabilizes generalization under variable history length T and non-uniform sampling.

(iii) Constructive cue. A prototypical temporal cue is a finite difference normalized by elapsed time, $\big(x_{1,q} - x_{0,q}\big)/(\tau_1 - \tau_0)$. With τ present, the mixer in (5)–(6) can *approximate* such normalization by (a) difference-like mixing across time-concatenated channels in W_2 and (b) gain modulation as a (piecewise-linear) function of τ; without τ, the denominator is absent and the model overfits to a single training interval.

Properties and discussion. *Parallel & attention-free* are independent over frames and fuse the entire window in one pass via pointwise mixing; no quadratic token-pair interactions occur. *Per-query locality:* temporal mixing is applied independently to each BEV query, decoupling temporal fusion from spatial resolution. *Fixed-budget behavior:* the dominant computation depends on C_{total}, aligning with edge deployment. *Robustness to irregular sampling:* τ_t captures real time gaps, improving behavior under variable FPS, asynchrony, and late/dropped frames. *Practical notes:* ρ may be set to the camera sweep period (s/frame) or learned; variable T at test time can be handled by zero-padding to a preset maximum or by weight sharing with pooling before.

3.2 Spatial Fusion

Given multi-view image features, we seek to integrate them into the BEV representation by *replacing cross-attention with state space models* so that each *BEV query* can aggregate long-range, multi-view signals with near-linear scaling. This follows recent evidence that Mamba provides content-adaptive sequence mixing with linear-time scans while remaining competitive with attention in accuracy for vision tasks.

We replace cross-attention with a *Spatial Cross Mamba* (XQSSM) [12] module that performs SSM-based cross-attention between BEV queries and multi-view image features. For each BEV query q, we generate a small set of reference points by lifting (x, y) to pillar samples (x, y, z) and projecting them onto each camera; only locations inside the image are kept. We then interleave the corresponding query copies and the sampled image features into a single 1D sequence and process it with Mamba kernel.

To ensure that BEV *query* tokens do not modify the SSM hidden state during the scan, we gate the discretization step size by token type. Let $\text{type}(k) \in \{\text{img}, \text{query}\}$ denote the k-th token's type and define:

$$\Delta_k = \begin{cases} \Delta_{\text{img}} > 0, & \text{type}(k) = \text{img}, \\ 0, & \text{type}(k) = \text{query}. \end{cases}$$

With token-dependent selective parameters $(A_k, B_k, C_k) = \psi_\theta(t_k)$ and zero-order-hold (ZOH) discretization, we have:

$$\bar{A}_k = e^{\Delta_k A_k}, \qquad \bar{B}_k = \big(\bar{A}_k - I\big) A_k^{-1} B_k,$$

the discrete SSM update reads:

$$h_{k+1} = \bar{A}_k\, h_k + \bar{B}_k\, t_k, \qquad y_k = C_k\, h_k.$$

For *query* tokens we set $\Delta_k = 0$, hence $\bar{A}_k = I$ and $\bar{B}_k = 0$, which generate the read-only property:

$$\text{type}(k) = \text{query} \implies h_{k+1} = h_k, \quad y_k = C_k\, h_k.$$

This allows BEV queries read from nearby image tokens at their insertion indices, so that the interaction of two modalities follows near-linear scaling.

This SSM-based cross-attention is hardware friendly and empirically competitive with Transformer counterparts, while improving input-scaling efficiency.

4 Experiment

We follow the methodologies of the previous work of [18, 29] and [30]. We evaluate on one backbones: ResNet50 which are pretrained

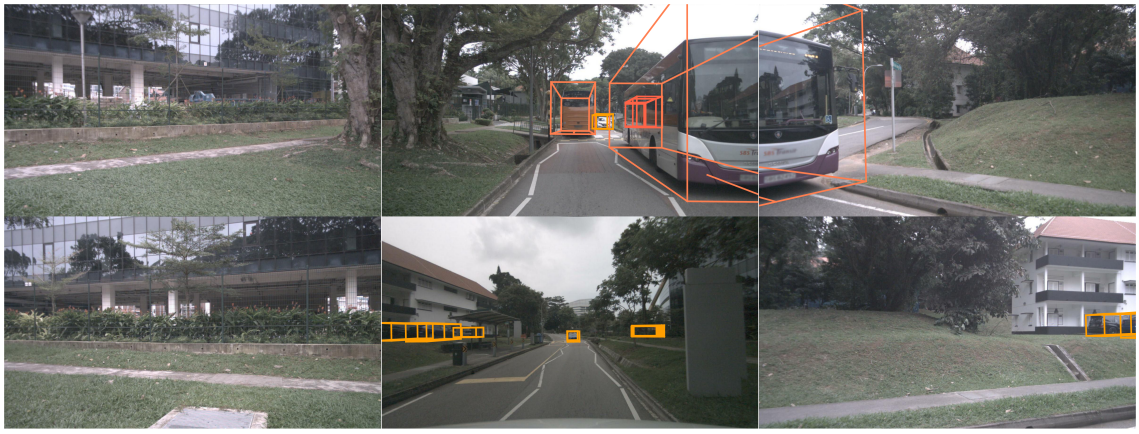


Figure 2: Visualization results of TinyBEV on nuScenes val set.

Figure 3: Visualization results of TinyBEV on nuScenes val set.

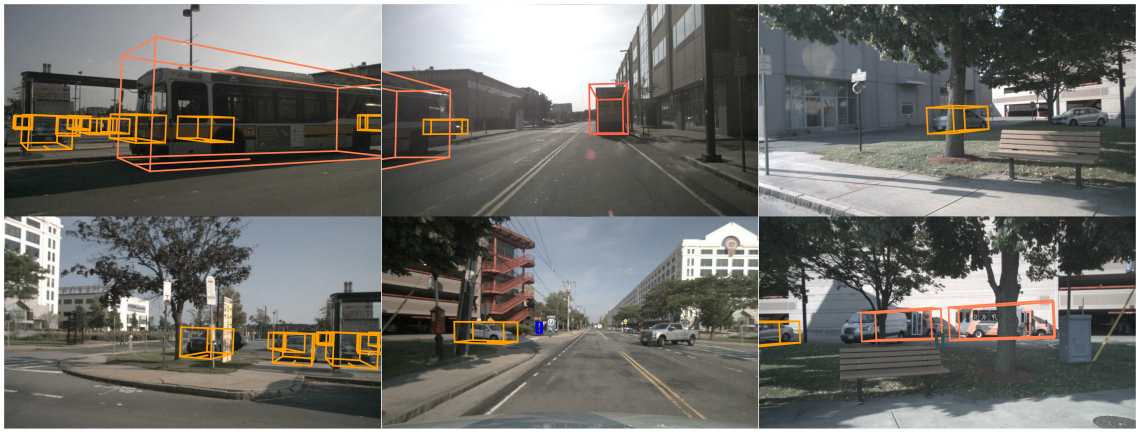


Figure 4: Visualization results of TinyBEV on nuScenes val set. We observe that our model can detect highly occluded objects.

on a depth prediction task and COCO. During training, the first stage of the backbone is frozen, and all other stages are trained at a 10% learning rate to fine-tune their latent representations to the multi-view autonomous driving setting.

4.1 Dataset and Metrics

We evaluate on the nuScenes dataset [2], a large-scale multi-modal autonomous driving benchmark comprising 1,000 urban driving

Method	Backbone	# Frames	NDS↑	mAP↑	mATE↓	mASE↓	mAOE↓	mAVE↓	mAAE↓
BEVFormerV1-T	ResNet50	3	0.354	0.252	0.900	0.294	0.655	0.657	0.216
BEVFormerV2-T*	ResNet50	3	0.397	0.270	0.820	0.301	0.594	0.469	0.195
MamBEV-T	ResNet50	3	0.399	0.266	0.794	0.298	0.575	0.469	0.199
BEVDiffuser	ResNet50	3	0.391	0.283	0.859	0.285	0.558	0.592	0.212
TinyBEV	ResNet50	3	0.396	0.267	0.856	0.300	0.553	0.462	0.203

Table 1: Our main results for 3D detection on nuScenes val set. * indicates models are trained by us.

# Frames	NDS↑	mAP↑	mATE↓	mASE↓	mAOE↓	mAVE↓	mAAE↓	Params↓	MamBEV-T
3	0.396	0.267	0.856	0.300	0.553	0.462	0.203	39M	71M
4	0.405	0.277	0.841	0.302	0.543	0.450	0.205	40M	96M
5	0.413	0.282	0.812	0.295	0.540	0.435	0.199	40M	128M
8	0.412	0.280	0.815	0.291	0.547	0.430	0.197	42M	266M

Table 2: Performance of our model across all metrics NDS of models on nuScenes validation set using different numbers of temporal frames.

scenes recorded in Boston and Singapore. Each scene is approximately 20 seconds in duration. The key frames are annotated at 2 Hz with 3D bounding boxes and attributes for 23 categories, of which 10 constitute the official 3D detection benchmark. Each scene is recorded with a 360° sensor suite—six cameras, five radars, and one LiDAR—together with GPS/IMU and full calibration. The evaluation metrics and framework for computing them are provided as part of the nuScenes devkit. The metrics used for evaluation are 1) the mean average precision (mAP), which evaluates both the localization and classification performance of the predicted results over four different thresholds using center distance on the ground plane, and five true-positive error metrics are reported for matched detections: average translation error (ATE), average scale error (ASE), average orientation error (AOE), average velocity error (AVE), and average attribute error (AAE). The overall nuScenes Detection Score (NDS) aggregates mAP with the five mean TP metrics to produce a single summary of both detection and box-quality performance. In our method and experiments, only the camera frames, sensor calibration data, and GPS data are used in making predictions.

4.2 Implementation Details

We used a learning rate of 8×10^{-4}, with a linear warmup for 10% of the scheduled steps starting from $\frac{8}{3}\times10^{-4}$ Following the warmup, the learning rate follows an epoch based cosine annealing schedule with a minimum learning rate of 8×10^{-7}. We trained with an effective batch size of 32 with no gradient accumulation on 2 Ada 6000 for 24 epochs. An exponential moving average according to the function $w'_t = (1-0.0002)w_t + 0.0002w_t$ is applied to all weights starting from the beignning of training. An AdamW optimizer with a 0.01 weight decay is used, and training employs an automatic mixed precision optimizer wrapper with an initial gradient scaling of 512. A 0.1 multiplier is applied to the learning rate of the backbone weights and the deformable attention sampling offsets [36]. We train the models from scratch using a randomly initialized network for the encoder layers.

4.3 Main Results

We first report our main results, followed by a series of ablation experiments designed to assess the contributions of individual model components. Models in these experiments were trained for 24 epochs and employed a ResNet-50 backbone, pre-trained on the COCO object detection dataset.

Main Results. We present a comparison of our results in Table 1 against state-of-the-art methods at compatible parameter and image input scales. We only use camera features; additionally, we do not make use of any auxiliary loss as in the works of Yang et al. [30]. We aligned the definitions of tiny model with MamBEV-T. We trained only the **tiny** version of our TinyBEV. Our TinyBEV model has 39M parameters. For comparison, MamBEV's tiny model has 71M parameters.

Effectiveness. Table 1 presents a comparison of our proposed models with various state-of-the-art methods. Our TinyBEV has comparable performance under similar conditions for the ResNet50 backbone. This demonstrates the effectiveness of our methods.

Efficiency. In table 2, we test the parameters when the temporal information increases. We have consistently maintained a significantly lower number of parameters than MamBEV-T. From the result, increasing frames helps overall performance until approximately 5 frames. Unsurprisingly, the mAVE, a measurement of velocity prediction error, is also significantly higher when we increase the history frames. The total number of parameters scales linearly with respect to the history frames.

5 Conclusion

We presented **TinyBEV**, a camera-only BEV perception framework that replaces cross-attention with an SSM-based spatial encoder and adopts a parallel, attention-free temporal fusion. Spatially, the SSMs based method aggregates multi-view images for BEV query without taking quadratic computation complexity. Temporally, we have comparable performance with SOTA while having less and stable parameters.

6 Limitations and Future Work

(L1) Diminishing returns with longer histories. As shown in Table 2, increasing temporal information from 3→8 frames makes NDS changes from 0.396→0.412 and even saturates several metrics.

(F1) Beyond spatial SSM: temporal and end-to-end SSMs. An appealing direction is to replace temporal fusion with state space models: e.g., a causal, streaming Mamba that maintains a compact BEV state with learned time gated updates. We leave these directions for future exploration.

Acknowledgments

Research was sponsored by funds from Toyota Motor North America and the Army Research Laboratory and was accomplished under Cooperative Agreement Number W911NF-23-2-0224. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Laboratory or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.

References

[1] Ben Agro, Quinlan Sykora, Sergio Casas, Thomas Gilles, and Raquel Urtasun. 2024. Uno: Unsupervised occupancy fields for perception and forecasting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 14487–14496.

[2] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. 2020. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 11621–11631.

[3] Albert Gu and Tri Dao. 2023. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752* (2023).

[4] Albert Gu, Karan Goel, and Christopher Ré. 2021. Efficiently modeling long sequences with structured state spaces. *arXiv preprint arXiv:2111.00396* (2021).

[5] Shadi Hamdan and Fatma Güney. 2024. CarFormer: Self-Driving with Learned Object-Centric Representations. In *ECCV*.

[6] Chunrui Han, Jinrong Yang, Jianjian Sun, Zheng Ge, Runpei Dong, Hongyu Zhou, Weixin Mao, Yuang Peng, and Xiangyu Zhang. 2024. Exploring recurrent long-term temporal fusion for multi-view 3d perception. *IEEE Robotics and Automation Letters* 9, 7 (2024), 6544–6551.

[7] Ali Hatamizadeh and Jan Kautz. 2025. Mambavision: A hybrid mamba-transformer vision backbone. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 25261–25270.

[8] Jinghua Hou, Tong Wang, Xiaoqing Ye, Zhe Liu, Shi Gong, Xiao Tan, Errui Ding, Jingdong Wang, and Xiang Bai. 2024. Open: Object-wise position embedding for multi-view 3d object detection. In *European Conference on Computer Vision*. Springer, 146–162.

[9] Yihan Hu, Jiazhi Yang, Li Chen, Keyu Li, Chonghao Sima, Xizhou Zhu, Siqi Chai, Senyao Du, Tianwei Lin, Wenhai Wang, et al. 2023. Planning-oriented autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 17853–17862.

[10] Haoxuanye Ji, Pengpeng Liang, and Erkang Cheng. 2024. Enhancing 3d object detection with 2d detection-guided query anchors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 21178–21187.

[11] Hongyu Ke, Wanxin Jin, and Haoxin Wang. 2024. Carboncp: Carbon-aware dnn partitioning with conformal prediction for sustainable edge intelligence. *arXiv preprint arXiv:2404.16970* (2024).

[12] Hongyu Ke, Jack Morris, Kentaro Oguchi, Xiaofei Cao, Yongkang Liu, Haoxin Wang, and Yi Ding. 2025. MamBEV: Enabling State Space Models to Learn Birds-Eye-View Representations. In *The Thirteenth International Conference on Learning Representations*.

[13] Hongyu Ke and Haoxin Wang. 2023. Poster: Real-Time Object Substitution for Mobile Diminished Reality with Edge Computing. In *Proceedings of the Eighth ACM/IEEE Symposium on Edge Computing*. 279–281.

[14] Hongyang Li, Hao Zhang, Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, and Lei Zhang. 2024. Taptr: Tracking any point with transformers as detection. In *European Conference on Computer Vision*. Springer, 57–75.

[15] Hongyang Li, Hao Zhang, Zhaoyang Zeng, Shilong Liu, Feng Li, Tianhe Ren, and Lei Zhang. 2023. Dfa3d: 3d deformable attention for 2d-to-3d feature lifting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 6684–6693.

[16] Yinhao Li, Han Bao, Zheng Ge, Jinrong Yang, Jianjian Sun, and Zeming Li. 2023. Bevstereo: Enhancing depth estimation in multi-view 3d object detection with temporal stereo. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37. 1486–1494.

[17] Yinhao Li, Zheng Ge, Guanyi Yu, Jinrong Yang, Zengran Wang, Yukang Shi, Jianjian Sun, and Zeming Li. 2023. Bevdepth: Acquisition of reliable depth for multi-view 3d object detection. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 37. 1477–1485.

[18] Zhiqi Li, Wenhai Wang, Hongyang Li, Enze Xie, Chonghao Sima, Tong Lu, Yu Qiao, and Jifeng Dai. 2022. BEVFormer: Learning Bird's-Eye-View Representation from Multi-Camera Images via Spatiotemporal Transformers. *arXiv preprint arXiv:2203.17270* (2022).

[19] Haisong Liu, Yao Teng, Tao Lu, Haiguang Wang, and Limin Wang. 2023. Sparsebev: High-performance sparse 3d object detection from multi-camera videos. In *Proceedings of the IEEE/CVF international conference on computer vision*. 18580–18590.

[20] Xianpeng Liu, Ce Zheng, Ming Qian, Nan Xue, Chen Chen, Zhebin Zhang, Chen Li, and Tianfu Wu. 2024. Multi-view attentive contextualization for multi-view 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 16688–16698.

[21] Yue Liu, Yunjie Tian, Yuzhong Zhao, Hongtian Yu, Lingxi Xie, Yaowei Wang, Qixiang Ye, Jianbin Jiao, and Yunfan Liu. 2024. Vmamba: Visual state space model. *Advances in neural information processing systems* 37 (2024), 103031–103063.

[22] Chen Min, Dawei Zhao, Liang Xiao, Jian Zhao, Xinli Xu, Zheng Zhu, Lei Jin, Jianshu Li, Yulan Guo, Junliang Xing, et al. 2024. Driveworld: 4d pre-trained scene understanding via world models for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 15522–15533.

[23] Liliang Ren, Yang Liu, Yadong Lu, Yelong Shen, Chen Liang, and Weizhu Chen. 2024. Samba: Simple hybrid state space models for efficient unlimited context language modeling. *arXiv preprint arXiv:2406.07522* (2024).

[24] Jimmy TH Smith, Andrew Warrington, and Scott W Linderman. 2022. Simplified state space layers for sequence modeling. *arXiv preprint arXiv:2208.04933* (2022).

[25] Haoxin Wang, Xiaolong Tu, Hongyu Ke, Huirong Chai, Dawei Chen, and Kyungtae Han. 2025. lm-Meter: Unveiling Runtime Inference Latency for On-Device Language Models. *arXiv preprint arXiv:2510.06126* (2025).

[26] Haoxin Wang, Ziran Wang, Dawei Chen, Qiang Liu, Hongyu Ke, and Kyungtae KT Han. 2023. Metamobility: Connecting future mobility with the metaverse. *IEEE Vehicular Technology Magazine* 18, 3 (2023), 69–79.

[27] Tingqi Wang, Xu Zheng, Lei Gao, Tianqi Wan, and Ling Tian. 2023. SC-FGCL: Self-adaptive cluster-based federal graph contrastive learning. *IEEE Open Journal of the Computer Society* 4 (2023), 13–22.

[28] Tingqi Wang, Xu Zheng, Jinchuan Zhang, and Ling Tian. 2024. Federal graph contrastive learning with secure cross-device validation. *IEEE Transactions on Mobile Computing* (2024).

[29] Yue Wang, Vitor Campagnolo Guizilini, Tianyuan Zhang, Yilun Wang, Hang Zhao, and Justin Solomon. 2022. Detr3d: 3d object detection from multi-view images via 3d-to-2d queries. In *Conference on robot learning*. PMLR, 180–191.

[30] Chenyu Yang, Yuntao Chen, Hao Tian, Chenxin Tao, Xizhou Zhu, Zhaoxiang Zhang, Gao Huang, Hongyang Li, Yu Qiao, Lewei Lu, et al. 2023. Bevformer v2: Adapting modern image backbones to bird's-eye-view recognition via perspective supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 17830–17839.

[31] Yu Yang, Jianbiao Mei, Yukai Ma, Siliang Du, Wenqing Chen, Yijie Qian, Yuxiang Feng, and Yong Liu. 2025. Driving in the occupancy world: Vision-centric 4d occupancy forecasting and planning via world models for autonomous driving. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 39. 9327–9335.

[32] Jinqing Zhang, Yanan Zhang, Yunlong Qi, Zehua Fu, Qingjie Liu, and Yunhong Wang. 2025. Geobev: Learning geometric bev representation for multi-view 3d object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 39. 9960–9968.

[33] Wenzhao Zheng, Weiliang Chen, Yuanhui Huang, Borui Zhang, Yueqi Duan, and Jiwen Lu. 2024. Occworld: Learning a 3d occupancy world model for autonomous driving. In *European conference on computer vision*. Springer, 55–72.

[34] Wenzhao Zheng, Ruiqi Song, Xianda Guo, Chenming Zhang, and Long Chen. 2024. Genad: Generative end-to-end autonomous driving. In *European Conference on Computer Vision*. Springer, 87–104.

[35] Lianghui Zhu, Bencheng Liao, Qian Zhang, Xinlong Wang, Wenyu Liu, and Xinggang Wang. 2024. Vision mamba: Efficient visual representation learning with bidirectional state space model. *arXiv preprint arXiv:2401.09417* (2024).

[36] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. 2020. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159* (2020).