Injecting Text into BEV Perception to Study Language-Conditioned Attention

Bhaswanth Ayapilla
Carnegie Mellon University
bayapill@andrew.cmu.edu

Kartik Agrawal
Carnegie Mellon University
Kartika@andrew.cmu.edu

Abstract

Recent work on language-conditioned driving systems, exemplified by EMMA (End-to-End Multimodal Model for Autonomous Driving by Waymo [2]), suggests that natural language can act as a high-level semantic interface for autonomous driving without modifying visual encoders or prediction heads. Motivated by this perspective, we examine whether injecting textual context into a standard BEV perception model, specifically BEVFormer, can influence its internal reasoning during perception. Instead of re-architecting the full end-to-end stack, we introduce a minimal modification: lightweight text embeddings are added to the BEV query set so that language participates directly in the attention process. This isolates the effect of language on spatial feature aggregation while keeping the input sensors, training objectives, and downstream consumers unchanged. Through controlled variants with and without language tokens, we analyze both quantitative outputs and attention behavior. ***Our findings show that while language does not improve raw 3D detection accuracy, it significantly enhances behavior-related metrics such as orientation, velocity, and attribute estimation, indicating that BEV transformers can absorb useful semantic priors through lightweight textual conditioning.***

1. Introduction

Recent work in multimodal autonomous driving, exemplified by EMMA (End-to-End Multimodal Model for Autonomous Driving by Waymo [2]), demonstrates that natural language can act as a powerful high-level interface for reasoning about driving scenes. By framing driving as a language-conditioned reasoning task, EMMA enables models to leverage broad world knowledge without redesigning perception modules. However, this raises a fundamental open question: *at what stage of a traditional perception–planning pipeline does language begin to exert meaningful influence?* Most existing multimodal systems inject language at the decision or planning level, leaving unclear whether language can shape the internal spatial reasoning processes within the perception module itself.

In parallel, modern autonomous driving perception has shifted toward bird's-eye-view (BEV) representations. Among BEV-based approaches, BEVFormer has emerged as a dominant architecture, leveraging spatiotemporal transformers to aggregate visual information over time through temporal self-attention and spatial cross-attention. Despite its geometric strengths, BEVFormer, like most perception systems, operates purely on visual cues and lacks the means to incorporate task-aware semantic knowledge. Autonomous systems increasingly require perception that prioritizes contextually relevant elements ("focus on crosswalk regions", "attend more to pedestrians near the ego vehicle") rather than treating all objects uniformly.

Motivated by this gap, we ask whether lightweight textual context, when injected directly into the BEV query set, can alter how BEVFormer aggregates spatial features. To isolate linguistic influence, we adopt a minimal intervention: we keep all sensors, losses, and detection heads unchanged, and simply allow language embeddings to participate in the transformer's attention. This design lets us study whether language biases BEV construction and internal attention patterns, without major architectural changes.

This investigation matters for three reasons.

1. It separates genuine language effects from changes introduced by restructuring the perception–planning pipeline.
2. Integrating text into BEV queries provides a direct lens on how semantic information interacts with spatial attention, which are insights that conventional perception models do not expose.
3. The approach is lightweight and compatible with existing BEVFormer deployments, suggesting potential for controllable and interpretable BEV perception.

2. Related Work

Early autonomous driving perception systems relied on image-space detectors such as Faster R-CNN [8] and RetinaNet [8], but these architectures struggle with scale distortions and perspective inconsistencies across camera views. To overcome this, BEV-based fusion frameworks such as LSS [6], BEVFormer [3], and PETRv2 [5] emerged as the

dominant paradigm by lifting multi-view features into a spatially aligned bird's-eye grid, providing a natural interface for downstream planning and scene reasoning.

BEVFormer [3] advances this line of work by introducing a spatiotemporal transformer that constructs a persistent BEV representation using temporal self-attention and spatial cross-attention. This allows BEVFormer to aggregate multi-camera information over time without requiring depth supervision, and has made it a strong foundation for camera-only 3D detection. Complementary approaches such as PETRv2 [5] integrate explicit 3D positional encodings and unify camera features into a geometry-aware latent space, improving localization accuracy and highlighting the trend toward token-based BEV fusion architectures.

In parallel, multimodal learning in driving has explored the integration of language with vision. Works such as Talk2Car [1] and LAV [4] show that language can guide referential perception and instruction-following in interactive settings. More recently, EMMA [2] demonstrated that recasting driving as a VQA-style reasoning problem allows large multimodal models to leverage pretrained world knowledge for downstream driving decisions without specialized task heads.

However, most multimodal efforts either (i) inject language at the decision level (LAV, EMMA), or (ii) build task-specific fusion modules on top of the visual backbone. There is little work isolating whether language can influence the internal operation of a standard BEV perception transformer itself, without altering the end-to-end stack. This gap motivates our controlled study: a minimal intervention in which language embeddings are injected into BEVFormer's query set to examine how textual context interacts with BEV spatial reasoning.

3. Proposed Method

3.1. Overall Architecture

Modern BEV transformer architectures provide a structured and geometry-consistent representation that is well suited for spatial reasoning in autonomous driving. Because they operate on a unified top-down grid and use attention mechanisms to fuse multi-view camera features, they also offer a natural interface for introducing additional modalities such as language. Before describing our language-conditioning approach, we summarize the baseline BEVFormer architecture on which our method builds.

Multi-View Feature Extraction. Images from six cameras are encoded using a ResNet–FPN backbone to produce multi-scale feature maps that serve as visual tokens for attention.

BEV Transformer Encoder. A grid of learnable BEV queries represents the spatial area surrounding the ego vehicle. These queries are updated through:

- Temporal Self-Attention, which aligns the current BEV with past frames to provide temporal consistency, and
- Spatial Cross-Attention, which projects BEV queries into each camera view to sample image features relevant to each spatial cell.

Together, these modules generate a stable, geometry-aware BEV feature map.

Detection Head. The final BEV representation is decoded into 3D bounding boxes, orientations, velocities, and attributes. Since this head operates only on BEV features, any effect of language must manifest during the BEV construction process itself.

Motivation for Language Injection. Because BEVFormer's BEV queries drive the spatial and temporal aggregation of information, modifying these tokens provides a clean and interpretable interface for introducing high-level semantic priors. This motivates our methodology, which explores lightweight ways of injecting language into BEVFormer to analyze how textual context alters internal spatial reasoning.

3.2. Methodology

Building directly on this architecture, our methodology introduces lightweight textual tokens into the BEV query sequence to examine how language interacts with BEVFormer's attention-driven feature construction. The purpose is not to improve raw detection accuracy, but to isolate whether language can bias spatial reasoning within an otherwise standard BEV perception model.

Rationale. BEV transformers provide a unified coordinate frame in which all spatial relationships are explicit. Injecting language at the BEV query level therefore allows semantic cues, such as descriptions of scene context or object behavior, to influence the same queries responsible for aggregating visual information. This single-point intervention creates a controlled environment for studying language-conditioned attention without modifying sensors, supervision signals, or detection heads.

Language Priors. For each frame, we generate a short textual description such as *"Wide urban intersection with potential pedestrian crossings"*, using either a language model or rule-based heuristics from scene geometry. This text is embedded using a frozen CLIP [7] text encoder or a lightweight learnable embedding layer. The resulting language tokens are concatenated with the BEV queries before being passed to the transformer.

Attention Interaction. During both temporal self-attention and spatial cross-attention, the language tokens participate in the same attention operations as the BEV

queries. They compete for image features, influence spatial aggregation, and propagate contextual priors across time. This enables us to analyze whether high-level textual priors influence spatial attention and detection focus, even when the language content is weakly or randomly informative.

Ablation Settings. To systematically study the influence of language, we evaluate three conceptual variants:

1. Baseline BEVFormer with no language tokens
2. Static Language Tokens appended to the BEV query sequence
3. Dynamic Reasoning Tokens that evolve through the transformer layers

This methodology creates a clear and interpretable pathway for injecting language into BEV perception. By modifying only the BEV query set, we preserve the rest of the pipeline while enabling a targeted study of language-conditioned attention within BEVFormer.

3.3. Implementation Details

Our implementation extends the BEVFormer-Tiny architecture by injecting language features into different components of the perception pipeline. We use the nuScenes-Text annotations, which provide three textual descriptions for each object per frame, refined by a large language model (LLM) and stripped of viewpoint-dependent references (*e.g., "left", "right", "near the ego"*). These descriptions capture semantic cues such as behavior, attributes, intent, and scene context.

Language Embedding Extraction. We generate a 256-dimensional embedding for each sample token. Each cleaned textual description is passed through a frozen sentence encoder, and multiple descriptions are averaged. A 2-layer MLP projects the embeddings into the BEVFormer latent dimension before fusion.

Fusion Mechanisms. We experiment with several injection strategies:

- **BEV Grid Bias**: Add language bias directly to BEV grid queries.
- **Object Query Bias**: Add language to decoder object queries.
- **MLP-Aligned Bias**: Learn a projection before injection.
- **Light Cross-Attention**: Treat language as key/value attending to object queries.
- **Multi-Point Fusion**: Combine BEV bias, query bias, and cross-attention.
- **FiLM Conditioning**: Modulate BEV features using learned scale–shift parameters.

Training Setup. All models are fine-tuned from the official BEVFormer-Tiny pretrained weights. We use the AdamW optimizer with a base learning rate of 2×10^{-4} and cosine annealing schedule. All experiments are conducted on the nuScenes-mini dataset for rapid iteration.

4. Experiments

We evaluate the effect of language-conditioned fusion on 3D object detection performance and on key behavior-related metrics such as velocity, orientation, and attribute estimation. All experiments use the BEVFormer-Tiny backbone for comparability against prior work.

4.1. Datasets

nuScenes-mini. We use nuScenes-mini for all ablation experiments. It contains 10 scenes with full sensor calibration, annotations, and motion attributes. We use the official train/val splits.

nuScenes-Text. To incorporate semantic priors, we use the nuScenes-Text dataset [9], which provides detailed natural-language annotations for each object instance across time. The descriptions were refined using an LLM to remove viewpoint-dependent phrases and improve clarity while preserving semantic features such as:

- agent behavior (*e.g., "slowing down", "changing lanes"*),
- appearance cues (*"large white truck", "small blue car"*),
- motion attributes (*"turning left", "accelerating"*),
- scene context (*"near intersection", "at pedestrian crossing"*).

We average multiple descriptions per object to obtain a single embedding per sample token.

4.2. Results

Table 1 summarizes the performance of different fusion strategies. While language does not improve raw 3D box detection (mAP), it significantly enhances behavior-related metrics that contribute to the nuScenes detection score (NDS).

Key Findings. Language improves semantic consistency and behavior understanding rather than raw mAP. Our best model (FiLM + Cross-Attn + BEV Bias) yields strong gains in motion and attribute metrics. These results confirm that language introduces strong priors that improve behavioral reasoning in 3D perception models.

5. Future Implications

Our results show that lightweight textual cues can meaningfully influence BEVFormer's behavioral outputs, improving

Method	mAP	NDS	Notes
BEV Grid Bias	0.1881	0.2121	Language bias added to BEV embedding
Object Query Bias	0.1697	0.2046	Injected into query content only
MLP-Aligned Bias	0.1688	0.2046	2-layer MLP improves stability
Light Cross-Attention	0.1717	0.1984	Language attends to decoder queries
Cross-Attn + BEV Bias + More Tokens	0.1744	0.2121	Multi-point injection
Cross-Attn + BEV Bias + More Tokens + FiLM	**0.1865**	**0.2683**	Best language-aware model (small LR + slow warmup)
Baseline (No Language)	**0.2429**	**0.2865**	Highest raw detection accuracy

Table 1. Comparison of language-conditioned fusion strategies on nuScenes-mini.

orientation, velocity, and attribute estimation, despite not increasing raw mAP. These findings suggest several promising directions for future work.

Interpretable and Steerable Perception. Language tokens change how the BEV transformer distributes its attention, giving us a direct way to see and influence what the model focuses on. This makes the perception system easier to understand and control, since its behavior can be guided or examined using straightforward, human-readable descriptions instead of opaque feature patterns.

Interactive and Task-Aware Perception Systems. If textual cues consistently bias spatial aggregation, future driving systems could support interactive prompts such as "focus on the crosswalk" or "prioritize small moving agents." This creates a bridge between high-level intent and low-level perception, enabling perception modules that can be dynamically reweighted based on the driving objective.

Lightweight Alternative to Full Multimodal Fusion. Our minimal intervention demonstrates that useful semantic conditioning does not require redesigning the entire end-to-end pipeline. A language-aware BEV encoder could act as a plug-in module for existing autonomy stacks, functioning during simulation-time analysis, dataset debugging, or safety-critical overrides where additional semantic priors are beneficial.

Ultimately, this work suggests that BEV transformers are not only geometric fusion modules but can also serve as semantically aligned spatial token processors. Incorporating lightweight textual priors into BEV construction opens the door to controllable, interpretable, and more context-aware perception systems for autonomous driving.

References

[1] Thierry Deruyttere, Simon Vandenhende, Dusan Grujicic, Luc Van Gool, and Marie-Francine Moens. Talk2car: Taking control of your self-driving car. *arXiv preprint arXiv:1909.10838*, 2019. 2

[2] Jyh-Jing Hwang, Runsheng Xu, Hubert Lin, Wei-Chih Hung, Jingwei Ji, Kristy Choi, Di Huang, Tong He, Paul Covington, Benjamin Sapp, Yin Zhou, James Guo, Dragomir Anguelov, and Mingxing Tan. Emma: End-to-end multimodal model for autonomous driving. *arXiv preprint arXiv:2410.23262*, 2024. 1, 2

[3] Z Li, W Wang, H Li, E Xie, C Sima, T Lu, Q Yu, and J Dai. Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers. arxiv 2022. *arXiv preprint arXiv:2203.17270*. 1, 2

[4] Shih-Yang Liu, Chien-Yi Wang, Hongxu Yin, Pavlo Molchanov, Yu-Chiang Frank Wang, Kwang-Ting Cheng, and Min-Hung Chen. Dora: Weight-decomposed low-rank adaptation. In *Forty-first International Conference on Machine Learning*, 2024. 2

[5] 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*, pages 3262–3272, 2023. 1, 2

[6] 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*, pages 194–210. Springer, 2020. 1

[7] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 2

[8] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015. 1

[9] Manolis Savva, Dequan Wang, Wenqi Shao, Jiageng Mao, Kunyang Sun, Dahua Lin, et al. nuscenes-text: A large-scale multimodal dataset for language-guided 3d perception. In *CVPR*, 2024. 3