# MTA: Multimodal Task Alignment for BEV Perception and Captioning

Supplementary Material

## 6. Experimental Details

### 6.1. Detection Metric Details

The nuScenes detection task [2] involves detecting 10 object classes in 3D space, estimating bounding boxes, attributes (e.g. sitting vs. standing), and velocities.

**Average Precision Metric.** The mean Average Precision (mAP) metric is used, where a match is defined by thresholding the 2D center distance $d$ on the ground plane rather than using intersection over union (IOU). This decouples detection from object size and orientation. mAP is calculated as:

$$\text{mAP} = \frac{1}{|\mathbb{C}||\mathbb{D}|} \sum_{c \in \mathbb{C}} \sum_{d \in \mathbb{D}} \text{AP}_{c,d} \tag{9}$$

where $\mathbb{C}$ is the set of classes, $\mathbb{D} = \{0.5, 1, 2, 4\}$ meters are the matching thresholds, and $\text{AP}_{c,d}$ is the AP for class $c$ at threshold $d$.

**True Positive Metrics.** Several True Positive (TP) metrics are defined for each prediction matched with a ground truth box using a center distance of $d = 2$m:

- Average Translation Error (ATE): Euclidean center distance in 2D (meters)
- Average Scale Error (ASE): 3D IOU after aligning orientation and translation $(1 - \text{IOU})$
- Average Orientation Error (AOE): Smallest yaw angle difference (radians)
- Average Velocity Error (AVE): L2 norm of the 2D velocity differences (m/s)
- Average Attribute Error (AAE): $1 -$ accuracy of attribute classification

The mean TP metric over all classes is:

$$\text{mTP} = \frac{1}{|\mathbb{C}|} \sum_{c \in \mathbb{C}} \text{TP}_c \tag{10}$$

**nuScenes Detection Score.** The nuScenes Detection Score (NDS) consolidates the AP and TP metrics into a single scalar value:

$$\text{NDS} = \frac{1}{10} \left[ 5 \cdot \text{mAP} + \sum_{\text{mTP} \in \mathbb{TP}} (1 - \min(1, \text{mTP})) \right], \tag{11}$$

where $\mathbb{TP}$ is the set of the five mean True Positive metrics. Half of the NDS is based on detection performance (mAP), while the other half measures the quality of the detections in terms of box location, size, orientation, attributes, and velocity.

### 6.2. Further Implementation Details

In our implementation, we adopt the Bench2DriveZoo[1] [16] environment for BEVFormer [22]. The image backbone in the captioning module is ResNet50 [11]. The BEV feature size is $50 \times 50$ with 256 channels. The learnable cross-modal prompt tokens $\mathcal{P}$ in DCA have dimensions $\mathbb{R}^{N \times D}$, where $N = 4096$ and $D = 1024$.

For the Llama 3 model[2], the hidden size is 2048, and the feed-forward network (FFN) hidden size is 3072. The model has 32 attention heads and eight key-value heads. The number of hidden layers is 16. We use the PyTorch data type `bfloat16` for efficient computation. The activation function used is the Sigmoid Linear Unit (SiLU), defined as:

$$\text{silu}(x) = x \cdot \sigma(x), \tag{12}$$

where $\sigma(x)$ is the logistic sigmoid.

The Relation Q-Former [18] has a hidden size of 768 and consists of $L = 8$ layers with 16 attention heads each. For the Llama-Adapter [49], we insert learnable adaptation prompts into all the Llama layers except for the first layer. The prompting length for each layer is set to 10. The CLIP text encoder checkpoint is obtained from HuggingFace[3]. All the experiments are conducted on 2 H100 GPUs.

## 7. More Quantitative Results

We provide comprehensive versions of the ablation tables presented in the main paper. Tab. 11 is an extended version of Tab. 4 and shows the ablation study on the contributions of the BLA and DCA mechanisms. The results confirm that the combination of both BLA and DCA achieves the best overall performance, highlighting the significance of multimodal alignment in enhancing both captioning and perception tasks.

Tab. 12 is an extended version of Tab. 5 that presents the results of ablating the BLA attachment layer $\ell$. The findings confirm that aligning at the middle layers of the Q-Former yields the best performance, striking a balance between allowing the detection embedding to interact with the BEV features and providing sufficient capacity for mapping to the MLLM space.

The impact of the BLA objective and DCA objective are investigated in Tab. 13 (extended from Tab. 6) and Tab. 14 (extended from Tab. 7), respectively. The results further

[1] Bench2DriveZoo: https://github.com/Thinklab-SJTU/Bench2DriveZoo

[2] https://huggingface.co/meta-llama/Llama-3.2-1B

[3] https://huggingface.co/openai/clip-vit-base-patch32

confirm that the Mean Squared Error (MSE) objective is most effective for BLA, while the CLIP Contrastive Loss (CLIP) is the best-performing objective for DCA.

**BEV Encoder** To investigate the effect of different input modalities, we conduct experiments using both LiDAR point clouds and camera images as inputs. For processing the LiDAR point clouds, we adopt the BEVFusion-tiny [26][4] encoder, which transforms the multi-view camera images and LiDAR point clouds into unified spatial-temporal BEV representations $F_0^{\text{BEV}}$. The results are provided in Tab. 10. Similar to the BEVFormer setting, the proposed MTA significantly outperforms both the TOD3Cap and BEVFusion methods in detection and captioning tasks. Notably, MTA achieves a $9.7\%$ improvement in the C@0.5 captioning metric and a $7.2\%$ improvement in the mAP detection metric compared to TOD3Cap. These results demonstrate the effectiveness of MTA in enhancing the performance of both tasks, even when using different input modalities.

**Object Proposal Order** MTA inherently handles object proposal ordering through direct conditioning. In MTA, each caption is generated based on its corresponding object proposal from the perception task head, ensuring a consistent one-to-one matching between detections and captions. To validate the importance of this ordered conditioning, we conduct an experiment where we shuffle the object proposals before feeding them into the captioning module. As shown in Tab. 8, shuffling the object proposals leads to a significant performance degradation across all captioning metrics.

| Object Proposals | C@0.5↑ | B-4@0.5↑ | M@0.5↑ | R@0.5↑ |
|---|---|---|---|---|
| Ordered | **118.7** | **47.6** | **48.4** | **66.2** |
| Shuffled | 45.9 | 29.1 | 33.8 | 40.4 |

Table 8. **Effect of object proposal order on dense captioning performance.** Shuffling the object proposals leads to a significant degradation across all metrics.

**Effect of Number of Prompt Tokens ($N$)** In the Detection-Captioning Alignment (DCA) module, we introduce $N$ learnable prompt tokens to serve as a shared embedding space for aligning the detection and captioning outputs. To investigate the impact of $N$ on the performance of MTA, we conduct an ablation study with different values of $N$, as shown in Tab. 9. We observe that setting $N = 4096$ achieves the best overall performance, providing a good balance between the captioning and perception metrics. Increasing $N$ to 8192 slightly improves the captioning performance but leads to a slight degradation in the perception metrics. On the other hand, decreasing $N$ to 1024 results in a drop in performance across all metrics.

| # Prompt Tokens ($N$) | Captioning | | Perception | |
|---|---|---|---|---|
| | C@0.5↑ | B-4@0.5↑ | NDS↑ | mAP↑ |
| 1024 | 111.3 | 46.1 | 38.4 | 27.2 |
| 4096 | 113.6 | 46.4 | 38.7 | 27.5 |
| 8192 | 114.1 | 46.7 | 38.3 | 27.3 |

Table 9. **Ablation study on the number of learnable prompt tokens ($N$) in the Detection-Captioning Alignment (DCA) module.** $N = 4096$ is the default setting and achieves the best overall performance.

**Effect of Number of Object Queries ($K$)** Considering the memory burden and optimization difficulty of the LLM given the dynamic number of objects present in the scene, we set a limit on the number of object queries ($K$) fed into the LLM at once, following [18]. We examine the effect of this cap by comparing the performance across different values of $K$, as shown in Tab. 15. Empirically, we found that the overall results increase with $K$, and the best overall performance is achieved at $K = 32$. However, if the cap is set too large (e.g., $K = 64$), the performance degrades. Therefore, in other experiments, we set $K = 32$. It is worth noting that in the original TOD3Cap implementation, the authors set $K = 2$. However, with the proposed MTA framework, the performance consistently surpasses that of TOD3Cap across different values of $K$, including $K = 2$, demonstrating the effectiveness of our approach in leveraging a larger number of object queries for improved captioning and perception performance.

## 8. More Qualitative Results

We provide additional qualitative results in Fig. 3 and Fig. 4. These results show that the proposed MTA approach provides improved detection and captioning capabilities. Specifically, in Fig. 3, TOD3Cap incorrectly describes object 1 (a trafficcone) as being positioned behind the car rather than in front of it in the caption. Similarly, in Fig. 4, misidentifies object 1 (a silver car) as a yellow bicycle in the caption, illustrating a heightened risk of hallucination. In contrast, the proposed MTA approach provides more accurate descriptions and object localization within the scene.

[4]We use the tiny version of BEVFusion implemented in https://github.com/jxbbb/TOD3Cap/blob/main/tod3cap_camera/projects/configs/bevformer/bevfusion_tiny_stage3.py

| Method | Captioning | | | | Perception | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | C@0.25/0.5↑ | B-4@0.25/0.5↑ | M@0.25/0.5↑ | R@0.25/0.5↑ | NDS↑ | mAP↑ | mATE↓ | mASE↓ | mAOE↓ | mAVE↓ | mAAE↓ |
| BEVFusion [26] | - | - | - | - | 38.2 | 34.6 | 0.634 | 0.293 | 0.716 | 1.084 | 0.267 |
| TOD3Cap [18] | 112.4/108.6 | 47.8/46.0 | 49.6/47.7 | 67.8/65.3 | 38.5 | 34.7 | 0.643 | 0.294 | 0.687 | 1.067 | **0.257** |
| **MTA (Ours)** | **123.6/119.1** | **49.7/48.1** | **50.8/49.1** | **69.5/67.2** | **40.1** | **37.2** | **0.629** | **0.287** | **0.669** | **1.011** | 0.262 |

Table 10. **Performance comparison of BEVFusion, TOD3Cap, and the proposed MTA when using both LiDAR point clouds and camera images as inputs.** The tiny version of BEVFusion is used as the BEV encoder. MTA significantly outperforms the baselines in both captioning and perception tasks, achieving a 9.7% improvement in C@0.5 and a 7.2% improvement in mAP over TOD3Cap.

| Method | Captioning | | | | Perception | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | C@0.25/0.5↑ | B-4@0.25/0.5↑ | M@0.25/0.5↑ | R@0.25/0.5↑ | NDS↑ | mAP↑ | mATE↓ | mASE↓ | mAOE↓ | mAVE↓ | mAAE↓ |
| TOD3Cap | 113.1/108.7 | 48.7/46.7 | 49.8/47.8 | 68.0/65.3 | 37.7 | 26.6 | 0.895 | 0.290 | 0.584 | 0.570 | 0.219 |
| + BLA (ours) | 115.8/111.9 | 48.3/46.4 | 49.3/48.0 | 68.1/65.4 | **38.9** | 27.7 | 0.885 | 0.287 | 0.577 | 0.543 | 0.201 |
| + DCA (ours) | 117.9/113.6 | 48.5/46.4 | 49.4/48.0 | 68.2/65.3 | 38.7 | 27.5 | 0.882 | 0.284 | 0.598 | 0.531 | 0.213 |
| + MTA (ours) | **122.8/118.7** | **49.4/47.6** | 50.3/48.4 | 68.7/66.2 | **38.9** | **27.9** | 0.878 | 0.285 | 0.595 | 0.541 | 0.213 |

Table 11. **Ablation study on the contributions of the BLA and DCA mechanisms.** Both the BLA and DCA mechanisms independently enhance overall performance over the TOD3Cap baseline. Combining both modules in the MTA framework yields the highest performance.

| BLA Layer | Captioning | | | | Perception | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | C@0.25/0.5↑ | B-4@0.25/0.5↑ | M@0.25/0.5↑ | R@0.25/0.5↑ | NDS↑ | mAP↑ | mATE↓ | mASE↓ | mAOE↓ | mAVE↓ | mAAE↓ |
| 1 (first) | 115.3/111.2 | **48.4/46.4** | 49.3/48.0 | 68.0/65.2 | 38.4 | 27.3 | 0.891 | 0.284 | 0.581 | 0.550 | 0.216 |
| 4 (middle) | **115.8/111.9** | 48.3/**46.4** | 49.3/48.0 | 68.1/65.4 | **38.9** | **27.7** | 0.885 | 0.287 | 0.577 | 0.543 | 0.201 |
| 8 (last) | 111.4/107.7 | 47.7/45.8 | 49.0/47.6 | 67.6/64.8 | 37.9 | 27.1 | 0.895 | 0.285 | 0.596 | 0.566 | 0.219 |

Table 12. **Ablating the BLA attachment layer $\ell$.** Aligning at the middle layer achieves the best performance.

| BLA Objective | Captioning | | | | Perception | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | C@0.25/0.5↑ | B-4@0.25/0.5↑ | M@0.25/0.5↑ | R@0.25/0.5↑ | NDS↑ | mAP↑ | mATE↓ | mASE↓ | mAOE↓ | mAVE↓ | mAAE↓ |
| CLIP | 114.7/110.6 | **48.4/46.4** | 49.3/47.9 | 68.1/65.3 | 38.8 | 27.6 | 0.879 | 0.288 | 0.578 | 0.538 | 0.214 |
| MSE | **115.8/111.9** | 48.3/**46.4** | 49.3/48.0 | 68.1/65.4 | **38.9** | **27.7** | 0.885 | 0.287 | 0.577 | 0.543 | 0.201 |
| Cos. Sim. | 114.2/110.3 | 48.3/46.3 | 49.3/48.0 | 68.1/65.3 | 38.5 | 27.3 | 0.903 | 0.286 | 0.590 | 0.530 | 0.211 |

Table 13. **Ablating the BLA objective.** The MSE loss achieves the best overall performance.

| DCA Objective | Captioning | | | | Perception | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | C@0.25/0.5↑ | B-4@0.25/0.5↑ | M@0.25/0.5↑ | R@0.25/0.5↑ | NDS↑ | mAP↑ | mATE↓ | mASE↓ | mAOE↓ | mAVE↓ | mAAE↓ |
| CLIP | **117.9/113.6** | **48.5/46.4** | 49.4/48.0 | 68.2/65.3 | **38.7** | **27.5** | 0.882 | 0.284 | 0.598 | 0.531 | 0.213 |
| MSE | 116.4/112.4 | 48.3/46.3 | 49.3/47.9 | 67.9/65.1 | 38.0 | 27.2 | 0.895 | 0.282 | 0.606 | 0.561 | 0.215 |
| Cos. Sim. | 117.3/113.1 | 48.4/**46.4** | 49.3/47.9 | 67.9/65.1 | 38.3 | **27.5** | 0.892 | 0.285 | 0.610 | 0.545 | 0.215 |

Table 14. **Ablating the DCA Objective.** The CLIP loss achieves the best results.

| Method | K | Captioning | | | | Perception | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | C@0.25/0.5↑ | B-4@0.25/0.5↑ | M@0.25/0.5↑ | R@0.25/0.5↑ | NDS↑ | mAP↑ | mATE↓ | mASE↓ | mAOE↓ | mAVE↓ | mAAE↓ |
| TOD3Cap | 2 | 113.1/108.7 | 48.7/46.7 | 49.8/47.8 | 68.0/65.3 | 37.7 | 26.6 | 0.895 | 0.290 | 0.584 | 0.570 | 0.219 |
| MTA (**ours**) | 2 | 119.4/115.3 | 48.9/46.8 | 49.9/47.8 | 68.2/65.2 | 38.6 | 27.2 | 0.883 | 0.284 | 0.581 | 0.537 | 0.215 |
| | 4 | 118.5/114.8 | 48.8/46.9 | 50.0/48.0 | 68.2/65.5 | 38.5 | 27.3 | 0.885 | 0.287 | 0.569 | 0.568 | 0.204 |
| | 8 | 121.3/117.5 | 49.2/47.2 | 50.1/48.1 | 68.5/65.7 | 38.3 | 27.1 | 0.884 | 0.286 | 0.593 | 0.546 | 0.215 |
| | 16 | 119.9/116.0 | 48.9/46.8 | 50.1/48.0 | 68.5/65.6 | 38.5 | 27.7 | 0.901 | 0.289 | 0.578 | 0.562 | 0.205 |
| | 32 | 122.8/118.7 | 49.4/47.6 | 50.3/48.4 | 68.7/66.2 | 38.9 | 27.9 | 0.878 | 0.285 | 0.595 | 0.541 | 0.213 |
| | 64 | 116.1/112.1 | 48.2/46.3 | 49.7/47.7 | 67.8/65.1 | 38.8 | 27.4 | 0.901 | 0.288 | 0.577 | 0.514 | 0.211 |

Table 15. **Ablating the object captioning sample number $K$.** $K = 32$ achieves the best overall performance.

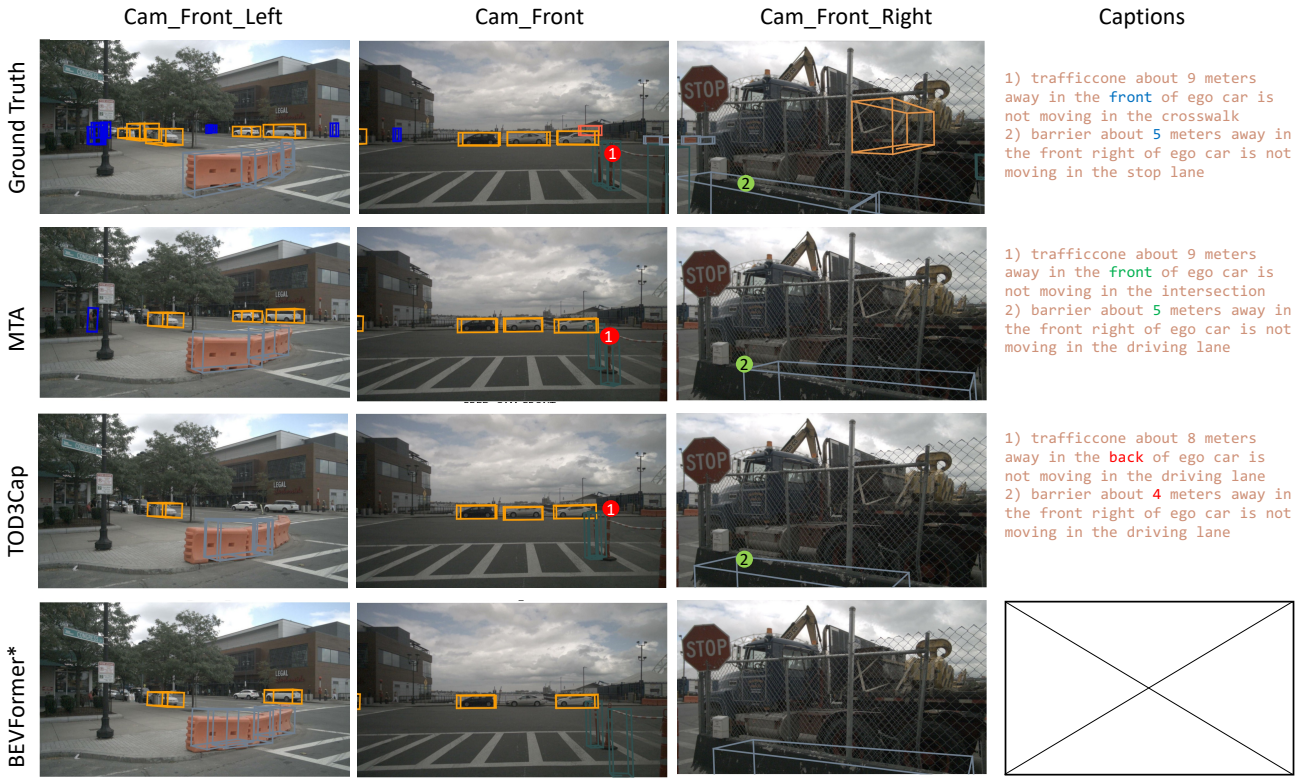


Figure 3. Qualitative results comparing the proposed MTA with baseline methods on nuScenes and TOD3Cap datasets. Visualization results show that MTA shows improved alignment with ground-truth detections over the counterpart methods. Captioning results show that the proposed MTA generates captions that are more accurate in terms of description and spatial orientation of objects over the TOD3Cap. Unlike MTA, TOD3Cap perceives object 1 (a trafficcone) to be positioned in the back of the car instead of front in the caption, illustrating a heightened risk of hallucination. *We note that BEVFormer is only suited for perception tasks, thus caption is not provided.

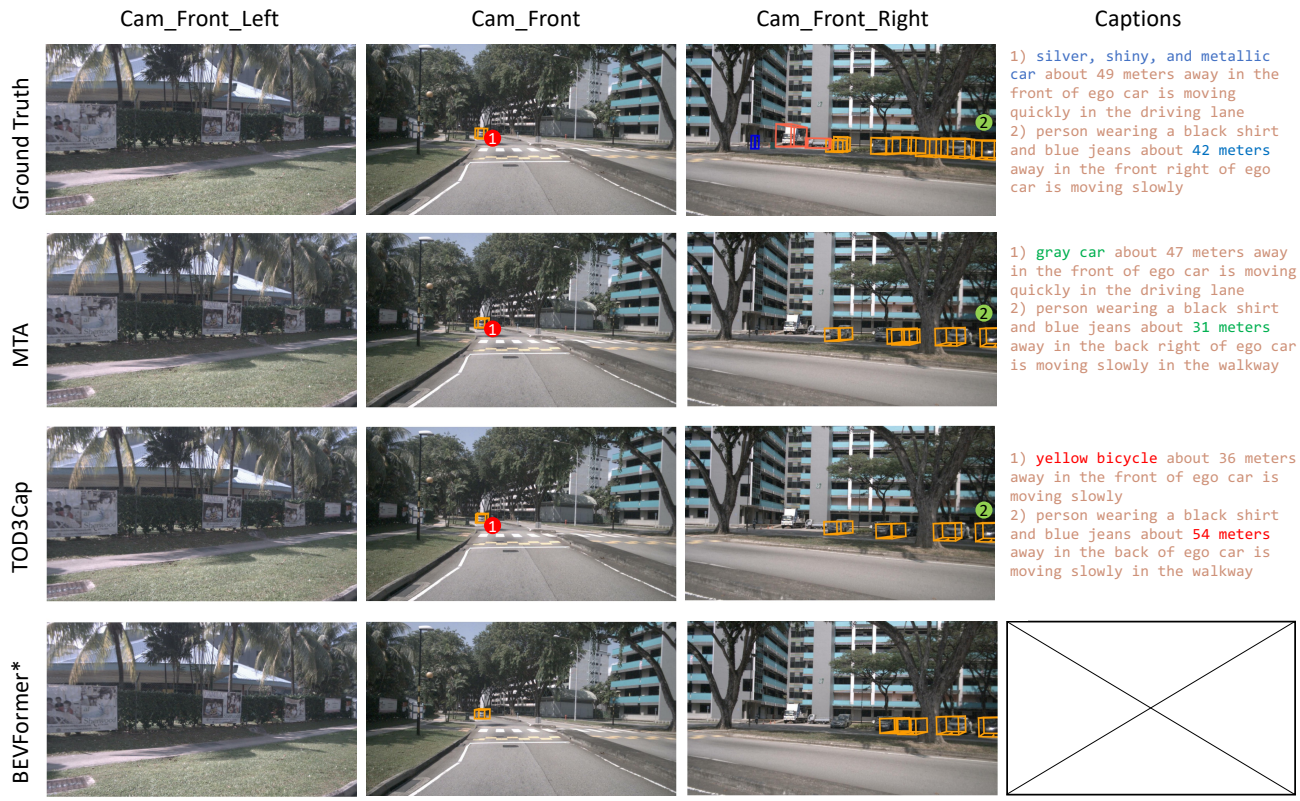


Figure 4. Qualitative results comparing the proposed MTA with baseline methods on nuScenes and TOD3Cap datasets. Visualization results show that MTA shows improved alignment with ground-truth detections over the counterpart methods. Captioning results show that the proposed MTA generates captions that are more accurate in terms of description and spatial orientation of objects over the TOD3Cap. Unlike MTA, TOD3Cap perceives object 1 (a silver car) to be a yellow bicycle in the caption, illustrating a heightened risk of hallucination. *We note that BEVFormer is only suited for perception tasks, thus caption is not provided.