When, Where, and What? A Benchmark for Accident Anticipation and Localization with Large Language Models

Anonymous Authors

1 IMPLEMENTATION DETAILS

For image preprocessing, we resize all input images to a size of 224×224 and feed them into a vision feature extractor, utilizing the MobileNetv2 model trained on ImageNet1K for feature extraction. We employ Cascade-rcnn for object detection on each image, obtaining up to 19 objects' bounding boxes. The images within these bounding boxes are also resized to 224×224 and subjected to feature extraction. The feature dimension of the output features is 1280. Within the Dynamic Object Attention mechanism, we first down sample the feature dimensions to 16 and set the number of iterations in the Dynamic Route to 8. A similar down sampling to 16 is applied in the Dual Vision Attention framework. For the loss function parameters, we set a decay coefficient $\lambda = 20$ and a loss function ratio coefficient $\eta = 10$. For the training parameters, we set the model learning rate to 1×10^{-4}, with a batch size of 16.

2 ABLATION STUDIES OF DYNAMIC OBJECT ATTENTION

2.1 Ablation Studies of the Sample Method

In diffusion models, the noise coefficient is related to the diffusion timestep, enabling a progressive effect between iterations. In the dynamic route approach, iterations replace the concept of timesteps. In Equation 1, the coefficient $\alpha^{(n)}$ is a discrete array concerning iteration n, and setting the value of $\alpha^{(n)}$ is crucial. While this value could be constant, doing so would maintain the same ratio between the result of the previous iteration $\mathcal{D}^{(n-1)}$ and the noise proportion ϵ across different iterations. Alternatively, employing a specific function to define $\alpha^{(n)}$ allows for dynamic adjustment in the relationship between $\mathcal{D}^{(n-1)}$ and ϵ with each iteration, resulting in varying noise intensities. We experimented with Linear, Cosine, and Sigmoid functions. As Table 1 illustrates, the Linear progression yields the best outcome, while the Cosine function performs the worst, with the AP value even slightly lower than the None scenario where $\alpha^{(n)}$ remains unchanged. Our analysis suggests that both Linear and Sigmoid functions are monotonically increasing within their domains, implying that the proportion of noise decreases as iterations progress. This aligns with the conventional understanding that higher initial noise levels enhance model robustness, but as iterations advance, the model requires optimization at a finer granularity, necessitating lower noise levels.

$$\mathcal{D}^{(n)} = \sqrt{\alpha^{(n)}}\mathcal{D}^{(n-1)} + \sqrt{1-\alpha^{(n)}}\epsilon \tag{1}$$

2.2 Ablation Studies of the Iterations

Ablation Studies of Dynamic Object Attention.

In the main text, we have presented experiments concerning varying numbers of iterations for the Dynamic Route during both training and testing phases. To further substantiate the significance of our conclusions, we have supplemented this with experiment result, as shown in Table 2. The experimental results further corroborate that a finite number of iterations yields optimal model performance, with excessive or insufficient iterations leading to overfitting or underfitting, respectively. Moreover, after training with multiple iterations, testing with a single iteration suffices to achieve satisfactory outcomes.

Table 1: Ablation studies of the dynamic object attention on different sample method.

Index	Method	Evaluation Metrics		
		AP(%)↑	mTTA(s)↑	AOLA↑
1	Linear	69.2	4.26	0.89
2	Cosine	65.4	4.21	0.82
3	Sigmoid	67.6	4.17	0.86
4	None	65.7	4.13	0.81

3 ABLATION STUDIES OF FEATURE EXTRACTOR

Traditional accident detection predominantly relies on utilizing VGG for feature extraction, resulting in slow extraction speeds and suboptimal performance. In response, we experimented with a variety of feature extractors and image preprocessing methods. As demonstrated in Table 5, we employed VGG-16, VGG-19, MobileNet, EfficientNet, ResNet101, Swin Transformer, and ViT for feature extraction. The results indicate that MobileNet delivers the best performance, coupled with minimal parameters and GFLOPS. Although a lower GFLOPS suggests higher computational efficiency, it demands more from GPUs and memory. Contrary to conventional wisdom, resizing images to dimensions higher than the standard 224 does not necessarily enhance performance; the optimal results were achieved with the traditional size of 224. This is presumably because conventional feature extractors are trained with images of 224×224 pixels, and resizing to dimensions beyond this threshold involves average pooling for dimension transformation, leading to the loss of critical information. Furthermore, excessively high feature dimensions are not conducive to accident detection and incur additional parameter and time costs.

4 ABLATION STUDIES OF THE LOSS FUNCTIONS

4.1 Ablation Studies of the Scaling Coefficient

The scaling coefficient η balances the ratio between the score loss L_S and the anticipation loss L_A, controlling the proportion between these two loss functions. During the training process of our model, we keep other variables constant and experiment with different values of η. As shown in Table 3, the experimental outcomes indicate

Table 2: Complete ablation studies of the Dynamic Object Attention on iterations. Num-iteration means the number of iterations that Dynamic Route used during the training and testing process. TC means the time consumption during training. During the training process, the time consumption by the model with Num-iteration=1 is set as a baseline of 1.

Index	Num-iteration		Evaluation Metrics			
	Train	Test	AP(%)↑	mTTA(s)↑	AOLA↑	TC(%)↓
1	2	2	63.1	3.95	0.82	1.02
2	3	3	64.9	4.01	0.84	1.03
3	4	4	66.8	4.10	0.85	1.04
4	5	5	68.4	4.19	0.88	1.05
5	6	6	69.2	4.26	0.89	1.07
6	7	7	68.8	4.26	0.89	1.10
7	8	8	68.7	4.28	0.88	1.12
8	9	9	68.3	4.24	0.86	1.13
9	10	10	67.4	4.23	0.86	1.15
10	5	1	65.8	4.03	0.80	-
11	5	2	66.7	4.10	0.82	-
12	5	3	67.5	4.14	0.85	-
13	5	4	68.0	4.17	0.87	-
14	5	5	68.4	4.19	0.88	-
15	6	1	66.4	4.16	0.82	-
16	6	2	67.1	4.20	0.85	-
17	6	3	68.3	4.22	0.87	-
18	6	4	68.9	4.23	0.88	-
19	6	5	69.0	4.25	0.88	-
20	6	6	69.2	4.26	0.89	-
21	7	1	66.8	4.14	0.79	-
22	7	2	67.3	4.17	0.82	-
23	7	3	67.7	4.20	0.85	-
24	7	4	68.2	4.22	0.87	-
25	7	5	68.5	4.23	0.87	-
26	7	6	68.6	4.25	0.88	-
27	7	7	68.8	4.26	0.89	-

Table 3: Ablation studies of the Scaling Coefficient η.

Index	η	Evaluation Metrics		
		AP(%)↑	mTTA(s)↑	AOLA↑
1	0.01	67.0	4.35	0.85
2	0.1	67.9	4.27	0.84
3	1	68.4	4.29	0.87
4	10	69.2	4.26	0.89
5	100	69.6	4.11	0.85
6	1000	70.3	3.79	0.86

Table 4: Ablation studies of the Decay Coefficient λ.

Index	λ	Evaluation Metrics		
		AP(%)↑	mTTA(s)↑	AOLA↑
1	5	67.9	4.39	0.82
2	10	68.6	4.25	0.86
3	20	69.2	4.26	0.89
4	50	69.4	3.95	0.89
5	100	69.5	3.72	0.87

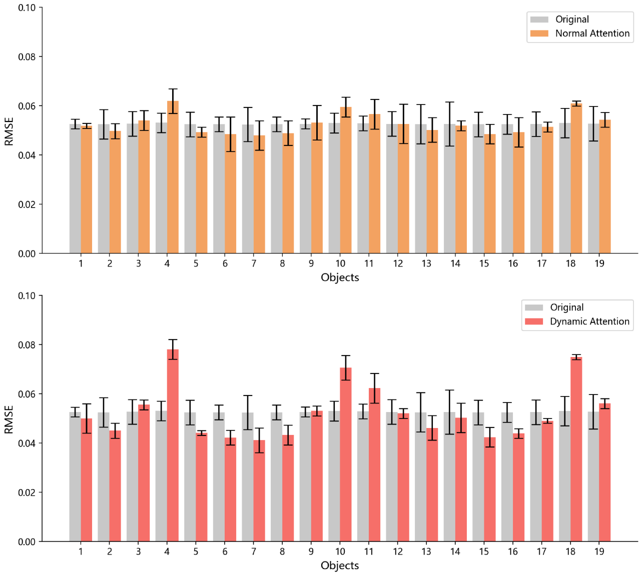


Figure 1: Visualization of the Attention Allocation. "Original" represents the object features, "Normal Attention" represents the object-aware features processed by a normal attention mechanism, "Dynamic Attention" represents the object-aware features processed by Dynamic Object Attention proposed in this paper.

that a lower η emphasizes score loss, making the model prioritize optimizing the probability of incident occurrence at each moment. This leads to an increase in mean Time to Accident (mTTA) and a decrease in Average Precision (AP); conversely, a higher η emphasizes anticipation loss, inclining the model towards optimizing the accuracy of predicting incident occurrence in videos, which results in an increase in AP and a decrease in mTTA. Comparably, changes in η have a minor effect on the accuracy of incident prediction.

4.2 Ablation Studies of the Decay Coefficient

The parameter λ in the loss function L_S represents the decay size in the exponent, controlling the decay rate of the loss function over time. Theoretically, a larger ratio coefficient implies greater weight variance across different time points. Excessive variance may cause the model to predominantly optimize probabilities close to the incident, resulting in a lower mean Time to Accident (mTTA); conversely, minimal variance can lead to the model equally focusing on information from all time points, reducing the accuracy of accident prediction. As depicted in Table 4, the final experimental results align with our hypothesis. Optimal performance is achieved when $\lambda = 20$.

Table 5: Comparison of models for the best AP on DAD datasets. Bold values represent the best performance of each category. Dimension means the output feature dimension of the backbone. Bounding box means the size of each bounding box in object detection. Resize means the size of each image will resize before feature extraction. Dimension means the output dimension of the extracted features. Params (M) represents the number of parameters in the model, while GFLOPS represents the number of floating-point operations per second the model performs during inference or training. A higher GFLOPS value indicates a higher computational demand on the GPU and memory.

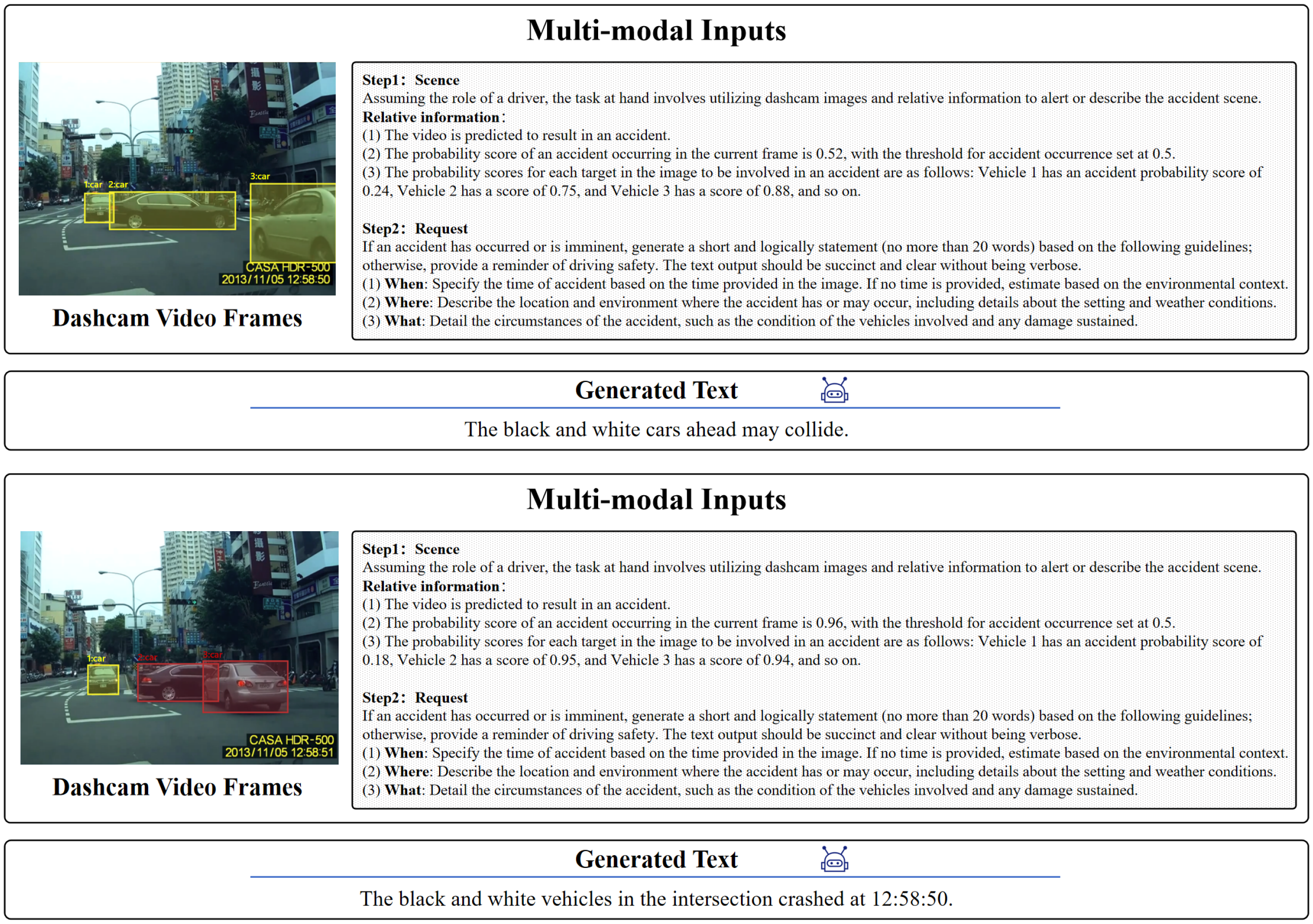
Index	Backbone	Resize	Bounding box	Dimension	Params (M)↓	GFLOPS↓	AP(%)↑	mTTA(s)↑	AOLA↑
1	VGG-16	224×224	224×224	512	138.4	15.47	61.8	3.91	0.77
2	VGG-16	224×224	224×224	1280	138.4	15.47	63.5	3.96	0.80
3	VGG-16	224×224	224×224	2048	138.4	15.47	63.3	3.92	0.79
4	VGG-16	224×224	224×224	4096	138.4	15.47	62.9	3.90	0.77
5	VGG-16	384×384	384×384	1280	138.4	15.47	62.7	3.94	0.78
6	VGG-16	384×384	224×224	1280	138.4	15.47	62.4	3.88	0.74
7	VGG-16	512×512	224×224	1280	138.4	15.47	58.1	4.07	0.75
8	VGG-16	1280×720	224×224	1280	138.4	15.47	56.8	4.15	0.75
9	VGG-19	224×224	224×224	1280	143.7	19.63	64.2	3.99	0.82
10	MobileNetv3	224×224	224×224	1280	5.5	**0.22**	68.9	4.22	0.87
11	MobileNetv2	224×224	224×224	512	**3.5**	0.3	67.6	4.12	0.84
12	MobileNetv2	224×224	224×224	1280	**3.5**	0.3	**69.2**	4.26	**0.89**
13	MobileNetv2	224×224	224×224	4096	**3.5**	0.3	68.5	**4.33**	0.82
14	MobileNetv2	384×384	224×224	1280	**3.5**	0.3	65.4	4.16	0.84
15	MobileNetv2	384×384	384×384	1280	**3.5**	0.3	66.9	4.20	0.88
16	MobileNetv2	512×512	224×224	1280	**3.5**	0.3	64.1	4.05	0.80
17	MobileNetv2	1280×720	224×224	1280	**3.5**	0.3	61.3	3.89	0.75
18	EfficentNet B0	224×224	224×224	1280	5.3	0.39	64.7	4.11	0.81
19	EfficentNet B7	224×224	224×224	1280	66.3	37.75	66.8	4.19	0.83
20	ResNet101	224×224	224×224	1280	44.5	7.8	62.8	4.14	0.78
21	Swin Transformer (S)	224×224	224×224	1280	48.6	8.74	59.4	4.05	0.72
22	Vision Transformer	224×224	224×224	1280	88.2	4.41	60.6	4.10	0.76

5 VISUALIZATION OF THE ATTENTION ALLOCATION

To further elucidate the efficacy of our proposed Dynamic Object Attention mechanism, we examined a specific scenario to compare the allocation of target features with no attention mechanism, traditional attention mechanism, and Dynamic Object Attention applied. For ease of comparison, we first normalized the number of targets using softmax across the dimension of target quantity, followed by summing up the feature dimensions for each target to derive the feature value of each target. As illustrated in Figure 1, compared to traditional attention mechanisms, employing Dynamic Object Attention facilitates a superior differentiation among diverse features, enabling the learning of more profound feature content.

6 PROMPT ENGINEERING

As observed from Figures 2 and 3, the generality of large models like GPT-4 does not afford them the specificity required for tasks such as accident anticipation, resulting in imprecise detections. Fine-tuning such large language models presents challenges in data collection and incurs significant costs. Therefore, we have employed a strategy where smaller models guide larger ones, as demonstrated, yielding superior performance compared to GPT-4. Specifically, we incorporate the outputs of the smaller model as part of the prompt to guide the multimodal large language model on certain key information. Additionally, we process the original images to annotate the corresponding vehicles, enabling the larger model to better recognize objects.



Step1: Scence
Assuming the role of a driver, the task at hand involves utilizing dashcam images and relative information to alert or describe the accident scene.
Relative information:
(1) The video is predicted to result in an accident.
(2) The probability score of an accident occurring in the current frame is 0.52, with the threshold for accident occurrence set at 0.5.
(3) The probability scores for each target in the image to be involved in an accident are as follows: Vehicle 1 has an accident probability score of 0.24, Vehicle 2 has a score of 0.75, and Vehicle 3 has a score of 0.88, and so on.

Step2: Request
If an accident has occurred or is imminent, generate a short and logically statement (no more than 20 words) based on the following guidelines; otherwise, provide a reminder of driving safety. The text output should be succinct and clear without being verbose.
(1) **When**: Specify the time of accident based on the time provided in the image. If no time is provided, estimate based on the environmental context.
(2) **Where**: Describe the location and environment where the accident has or may occur, including details about the setting and weather conditions.
(3) **What**: Detail the circumstances of the accident, such as the condition of the vehicles involved and any damage sustained.

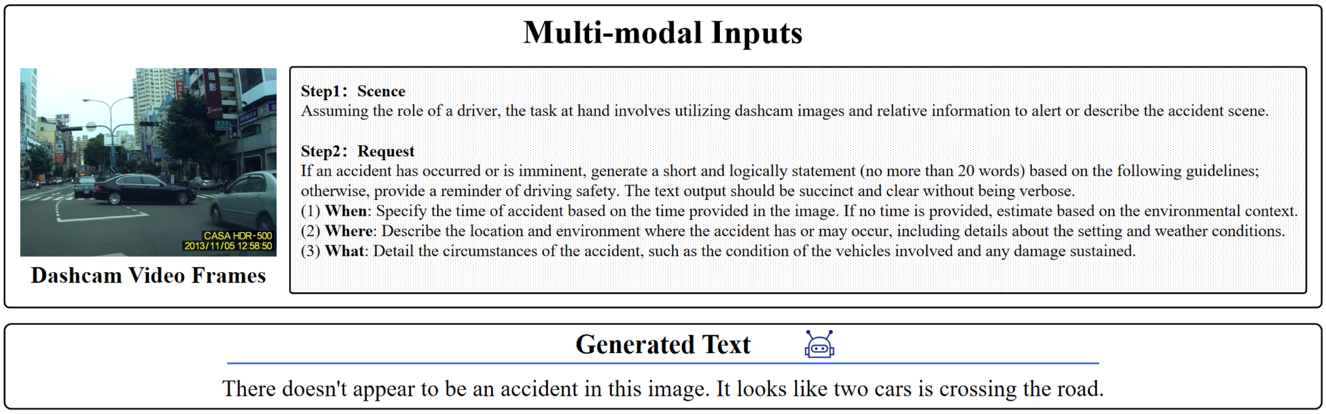
The black and white cars ahead may collide.

Step1: Scence
Assuming the role of a driver, the task at hand involves utilizing dashcam images and relative information to alert or describe the accident scene.
Relative information:
(1) The video is predicted to result in an accident.
(2) The probability score of an accident occurring in the current frame is 0.96, with the threshold for accident occurrence set at 0.5.
(3) The probability scores for each target in the image to be involved in an accident are as follows: Vehicle 1 has an accident probability score of 0.18, Vehicle 2 has a score of 0.95, and Vehicle 3 has a score of 0.94, and so on.

Step2: Request
If an accident has occurred or is imminent, generate a short and logically statement (no more than 20 words) based on the following guidelines; otherwise, provide a reminder of driving safety. The text output should be succinct and clear without being verbose.
(1) **When**: Specify the time of accident based on the time provided in the image. If no time is provided, estimate based on the environmental context.
(2) **Where**: Describe the location and environment where the accident has or may occur, including details about the setting and weather conditions.
(3) **What**: Detail the circumstances of the accident, such as the condition of the vehicles involved and any damage sustained.

The black and white vehicles in the intersection crashed at 12:58:50.

Figure 2: Detailed design of prompts using our model.



Step1: Scence
Assuming the role of a driver, the task at hand involves utilizing dashcam images and relative information to alert or describe the accident scene.

Step2: Request
If an accident has occurred or is imminent, generate a short and logically statement (no more than 20 words) based on the following guidelines; otherwise, provide a reminder of driving safety. The text output should be succinct and clear without being verbose.
(1) **When**: Specify the time of accident based on the time provided in the image. If no time is provided, estimate based on the environmental context.
(2) **Where**: Describe the location and environment where the accident has or may occur, including details about the setting and weather conditions.
(3) **What**: Detail the circumstances of the accident, such as the condition of the vehicles involved and any damage sustained.

There doesn't appear to be an accident in this image. It looks like two cars is crossing the road.

Figure 3: Detailed design of prompts using GPT4.