Abductive Ego-View Accident Video Understanding for Safe Driving Perception

Jianwu Fang[1†*], Lei-lei Li[2*], Junfei Zhou[2*], Junbin Xiao[3], Hongkai Yu[4], Chen Lv[5], Jianru Xue[1], and Tat-Seng Chua[3]

[1]Xi'an Jiaotong University [2]Chang'an University [3]National University of Singapore
[4]Cleveland State University [5]Nanyang Technological University

www.lotvsmmau.net *

Abstract

*We present **MM-AU**, a novel dataset for Multi-Modal Accident video Understanding. MM-AU contains 11,727 in-the-wild ego-view accident videos, each with temporally aligned text descriptions. We annotate over 2.23 million object boxes and 58,650 pairs of video-based accident reasons, covering 58 accident categories. MM-AU supports various accident understanding tasks, particularly multimodal video diffusion to understand accident cause-effect chains for safe driving. With MM-AU, we present an Abductive accident Video understanding framework for Safe Driving perception (**AdVersa-SD**). AdVersa-SD performs video diffusion via an Object-Centric Video Diffusion (OAVD) method which is driven by an abductive CLIP model. This model involves a contrastive interaction loss to learn the pair co-occurrence of normal, near-accident, accident frames with the corresponding text descriptions, such as accident reasons, prevention advice, and accident categories. OAVD enforces the object region learning while fixing the content of the original frame background in video generation, to find the dominant objects for certain accidents. Extensive experiments verify the abductive ability of AdVersa-SD and the superiority of OAVD against the state-of-the-art diffusion models. Additionally, we provide careful benchmark evaluations for object detection and accident reason answering since AdVersa-SD relies on precise object and accident reason information.*

1. Introduction

Autonomous Vehicles (AV) are around the corner for practical use [11]. Yet, occasionally emerging traffic accidents are among the biggest obstacles to be crossed. To make a step further, it is urgent to comprehensively understand the traffic accidents, such as telling what objects are involved, why an accident occurs and how to prevent it. Techniques that can answer these questions are of crucial importance for safe AV systems. So far, there is a lack of a large-scale dataset to develop such techniques.

Therefore, this paper constructs MM-AU, a multi-modal dataset for ego-view accident video understanding. MM-AU contains **11,727** in-the-wild ego-view accident videos. The videos are temporally aligned with the text descriptions of accident reasons, prevention solutions, and accident categories. In total, **58.6K** pairs of video-based Accident reason Answers (ArA) are annotated for 58 accident categories. Moreover, to enable object-centric accident video understanding, we annotate over **2.23M** object boxes for about **463K** video frames. As shown in Fig. 1, MM-AU can facilitate 8 tasks of traffic accident video understanding, and the models are required to infer ① what objects are involved, ② what kinds of accidents, ③ where and ④ when the accident will occur, ⑤ why the accident occurs, ⑥ what are the keys to accident reasons, ⑦ how to prevent it, and ⑧ multimodal accident video diffusion.

Different from previous works that concentrate on the former 4 basic tasks [3, 15, 24, 59], we advocate an Abductive Accident Video Understanding for Safe Driving (**AdVersa-SD**) perception by considering the accident reasons into the accident video understanding. For the ⑤-⑦ tasks, few works [37, 65] formulate Video Question Answering (VQA) problem to discern the accident reasons for given videos. However, abductive understanding of accident is more crucial to prevent collision. Hence, we present AdVersa-SD, which underscores a diffusion technique to bridge the visual content with specific accident reason texts.

Leveraging the text-vision CLIP model [44] and the video diffusion techniques [12, 57], we propose an abductive CLIP in AdVersa-SD with a constrastive interaction loss for the accident reason involved semantic co-occurrence learning within the text and video clips, such as the pairs of (■, t_r) and (■, t_a). To verify the abductive ability of abductive CLIP, we treat it as an engine to drive an Object-centric Accident Video Diffusion (OAVD) model by enforcing the learning of object locations and restricting the

*Equal contribution. †Corresponding author.

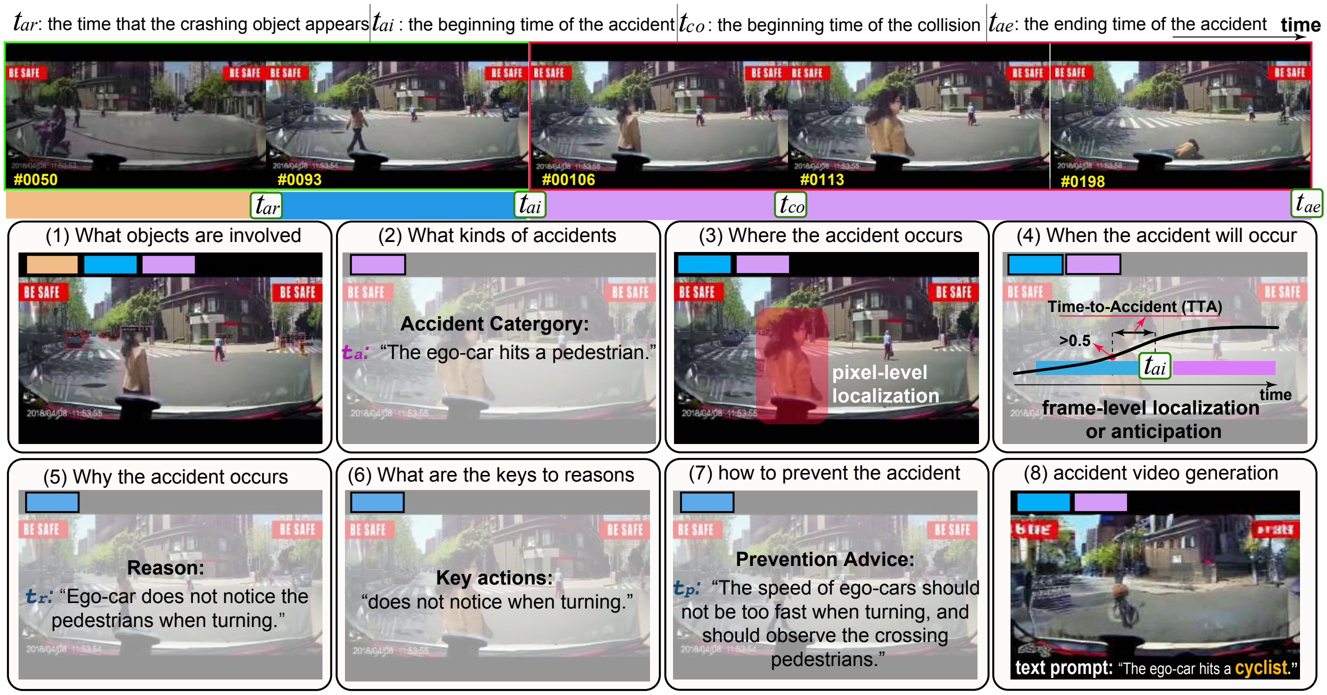

Figure 1. The ego-view multimodality accident video understanding tasks that **MM-AU** can support, where we highlight the text descriptions for accident reason (t_r), prevention advice (t_p), and accident category (t_a), as well as temporal windows (**accident-free**, **near-accident**, and **accident** windows) for different tasks.

influence of frame background for causal region generation. OAVD takes the stable diffusion model [46] as the baseline and extends it to the accident video diffusion by 3D convolution and well-designed spatial-temporal attention modules, as well as an object region masked latent representation reconstruction. This formulation is useful for finding the dominant objects of certain accidents irrelevant to the scene background. Thus, the **contributions** are threefold.

(1) A new large-scale ego-view multimodality accident understanding dataset, *i.e.*, MM-AU, is created, which will facilitate the more promising abductive understanding for safe driving perception. (2) We present AdVersa-SD, an abductive accident video understanding framework, to learn the dominant reason-occurrence elements of accidents within text-video pairs. (3) Within AdVersa-SD, we propose an Object-centric Accident Video Diffusion (OAVD) driven by the abductive CLIP to attempt to explicitly explore the causal-effect chain of accident occurrence, and positive results are obtained.

2. Related Work

2.1. Ego-View Accident Video Understanding

Accident Detection: Accident detection in ego-view videos aims to localize the spatial regions and temporal frame windows where the accident occurs. Because of the drastic change in the shape, location, and relations of the participants, the key problems of accident detection are to extract robust appearance or motion features for the representation of video frames, spatial-temporal video volumes, or trajectories. Commonly, the frame consistency [18, 43, 49, 73], location consistency [21, 32, 48, 52], and scene context consistency (*e.g.*, the object interactions) [13, 47, 56] are modeled to find the accident window or regions. Up to now, unsupervised location or frame prediction has been a common choice for model designing. For example, DoTA [66, 67], the typical ego-view accident detection method, computes the Standard Deviation (STD) of predicted locations of the pre-detected objects.

Accident Anticipation: Accident anticipation aims to forecast the probability and prefers an early warning of future accidents based on the complex scene structure modeling in video frames [2, 25, 40]. The earliness is maintained by taking the exponential loss [3, 5, 22, 51, 59] to penalize the positive occurrence of the accident. Different from accident detection, most accident anticipation works need to provide the accident window annotation to fulfill supervised learning. One groundbreaking work by Chan *et al.* [5] models a Dynamic-Spatial-Attention Recurrent Neural Network (DSA-RNN) to correlate the temporal consistency of road participants' tracklets, which is extended by the works [26, 27] to compute the riskiness of video frames or objects. To boost the explainability of accident anticipation, Bao and Kong [3] develop a Deep Reinforced accident anticipation with Visual Explanation (DRIVE) assisted by driver attention [15] and obtain a significant improvement.

Table 1. Attribute comparison of ego-view accident video datasets.

Datasets	Years	#Clips	#Frames	Bboxes	Tracklet	TA.	TT	R/S
DAD [5]	2016	1,750	175K		✓	✓		R
A3D [67]	2019	3,757	208K			✓		R
GTACrash [28]	2019	7,720	-			✓		S
VIENA^2 [1]	2019	**15,000**	**2.25M**			✓		S
CTA [68]	2020	1,935	-			✓	✓	R
CCD [2]	2021	1,381	75K		✓	✓		R
TRA [36]	2022	560	-			✓		R
DADA-2000 [15]	2022	2000	658k			✓		R
DoTA [66]	2022	5,586	732K	partial		✓		R
ROL [27]	2023	1000	100K			✓		R
DeepAccident [60]	2023	-	57k			✓		S
CTAD [39]	2023	1,100	-			✓		S
MM-AU	2023	**11,727**	**2.19M**	✓		✓	✓	R

Bboxes: bounding boxes of objects, **TA.**: temporal annotation of the accident, **TT**: text descriptions, **R/S**: real or synthetic datasets.

Table 2. Static attributes of ego-view accident video datasets.

Datasets	weather condition				occasion situations				
	sunny	rainy	snowy	foggy	highway	urban	rural	mountain	tunnel
CCD [2]	1,306	61	14	0	148	725	502	5	1
A3D [67]	2,990	251	474	42	225	2,458	720	328	26
DADA-2000 [15]	1,860	130	10	-	**1,420**	380	180	-	20
DoTA [66]	4,920	341	313	12	617	3,656	1,148	145	20
MM-AU	**10,116**	**761**	**793**	**57**	1082	**7,563**	**2,548**	**484**	**50**

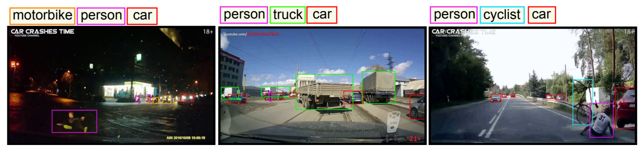

Figure 2. Some samples of object annotation in MM-AU.

Accident Classification: Because of the video data limitation of different accident categories, there is a paucity of research on ego-view video-based accident classification, and many works concentrate on the surveillance view with limited image set [17, 29, 39]. Kang *et al.* [24] propose a Vision Transformer-Traffic Accident (ViT-TA) model that classifies the ego-view traffic accident scenes, highlighting key objects through attention maps to increase the objectivity and reliability of functional scenes.

The aforementioned works focus on the monocular vision modal, while the spatial or temporal causal part of the accident video is hard to learn effectively owning to the complex evolution of accidents.

Accident Reason Answering: Closely related to this work, You and Han [68] investigate the causal-effect recognition of accident scenarios, and build the class taxonomy of traffic accidents. Besides, SUTD-TrafficQA [65] formulates the reason explanation and prevention advice of accidents by the Question-Answering (QA) framework, which involves the reasoning of dynamic and complex traffic scenes. Based on this, Liu *et al.* [37] reason the cross-modal causal relation to fulfill the traffic accident reason answering. We believe QA frameworks [64, 71] can provide a direct understanding for telling why the accident occurs. However, there is no explicit double-check solution to verify what key elements (*e.g.*, specific actions or objects) are dominant for subsequent accidents.

2.2. Ego-View Accident Understanding Datasets

The community has realized the importance of accident video understanding for safe driving perception, and some benchmarks have been released in recent years. Tab. 1 presents the attribute comparison of available ego-view accident video datasets. DAD [5] is the pioneering dataset, where each video clip is trimmed with 10 accident frames at the end of each clip. This setting is also adopted in the CCD datasets [2] with a total of 50 frames for each clip. A3D [67] and DoTA [66] are used for unsupervised ego-view accident detection [13, 66, 67]. Specially, the DADA-2000 dataset [15] annotates the extra driver attention data. Because of the difficulty to collect enough accident videos in real world, some work leverage the simulation tool to synthesize the virtual accident videos or object tracklets, such as GTACrash [28], VIENA^2 [1], DeepAccident [60], and CTAD [39]. However, the real-synthetic data domain gap is a tough nut to crack because it is rather hard to project the natural evolution process of accidents in the simulation tools. Besides CTA [68], the vision modal is concentrated, and the meaningful text descriptions are not explored.

3. MM-AU Dataset

The videos in MM-AU are collected from the publicly available ego-view accident datasets, such as CCD [2], A3D [67], DoTA [66], and DADA-2000 [15], and various video stream sites[1]. As presented in Tab. 2, the weather conditions and occasion situations are various and our MM-AU owns the largest sample scale. In total, **11,727** videos with **2,195,613** frames are collected and annotated. All videos are annotated the text descriptions, accident windows, and accident time stamps. To the best of our knowledge, MM-AU is the largest and most fine-grained ego-view multi-modal accident dataset. The annotation process of MM-AU is depicted as follows.

Accident Window Annotation: Leveraging the annotation criteria of DoTA [66], the accident window is labeled by 5 volunteers, and the final frame indexes are determined by average operation. The temporal annotation contains the beginning time of the accident t_{ai}, the end time of the accident t_{ae}, and the beginning time of the collision t_{co}. The frame ratio distributions within different windows of [0,t_{ai}], [t_{ai},t_{co}], [t_{ai},t_{ae}], [t_{co},t_{ae}], and [t_{ae}, end] are shown in Fig. 3(a). It can be seen that, in many videos, many accidents end in the last frame. The accident window of [t_{ai},t_{ae}] of most videos occupies half of the video length, which is

[1]https://www.youtube.com, https://www.bilibili.com, and https://v.qq.com.

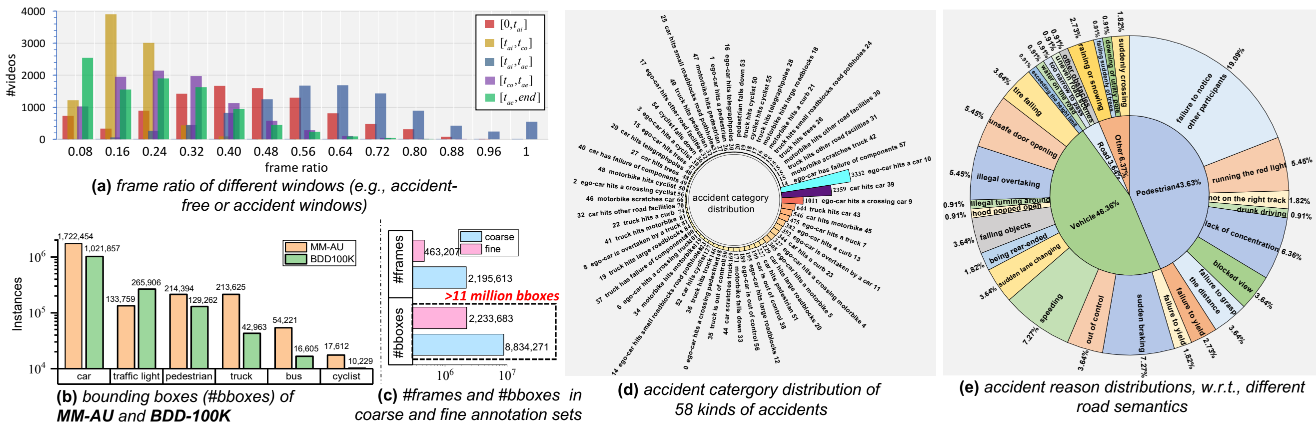

Figure 3. The annotation attribute statistics in MM-AU for the temporal, object, and text annotations. Better viewed in zoomed-in mode.

useful for model training in accident video understanding.

Object Detection Annotation: To facilitate the object-centric accident video understanding, we annotate 7 classes of road participants (*i.e., cars, traffic lights, pedestrians, trucks, buses, motorbikes, and cyclists*) in MM-AU. To fulfill an efficient annotation, we firstly employ the YOLOX [16] detector (pre-trained on the COCO dataset [35]), to initially detect the objects in the raw MM-AU videos. Secondly, MM-AU has a fine annotation set and a coarse annotation set of bounding boxes (#bboxes). For the fine annotation set, we took three months to manually correct the wrong detections using LabelImg every five frames by ten volunteers, and **2,233,683** bounding boxes within 463,207 frames are obtained. Each bounding box is double-checked for the final confirmation, and some samples are shown in Fig. 2. As for the coarse annotation set, we utilize a state-of-the-art (SOTA) DiffusionDet [7][2] to obtain the object bboxes for the remainder of frames in MM-AU. Fig. 3(b) presents the #bboxes on different road participants with the comparison to BDD-100K [69], and Fig. 3(c) shows the #frames and #bboxes on the fine and coarse annotation sets.

Text Description Annotation: Different from previous ego-view accident video datasets, MM-AU annotates three kinds of text descriptions: accident reason, prevention advice, and accident category descriptions. Accident category description is certainly aligned to the accident window $[t_{ai}, t_{ae}]$, while the reason and prevention advice description are aligned to the near-accident window $[t_{ai}-40, t_{ai}]$ for enhancing the vision-text contrastive learning between the near-accident windows and the accident reason or prevention texts. The descriptions and the video sequences do not show a unique correlation, and each description sentence usually correlates with many videos because of the co-occurrence. Similar to [14], based on the road layout, road user types, and their movement actions, we annotate 58 description sentences for accident categories, and their sample distribution is shown in Fig. 3(d).

We annotate 110 pairs of sentences for accident reason and prevention advice descriptions correlating to four kinds of road semantics, *i.e.*, pedestrian-centric, vehicle-centric, road-centric, and others (environmental issue). Fig. 3(e) shows the accident reason distribution concerning different road semantics. It is clear that the "*failure to notice other participants*" is dominant for pedestrian-centric accident reasons, and "*speeding*", "*sudden braking*", and "*illegal overtaking*" are the main kinds of vehicle-centric accident reasons. Following the form of Video Question Answering (VideoQA) task [64], we provide an Accident reason Answering (ArA) task while there is only one question "`What is the reason for the accident in this video?`". For each accident reason of a video, we further provide four reasonable distractors[3] to form a multi-choice ArA task, and the distractor reasons are all unrelated to the target accident video. We obtain **58,650** ArA pairs in MM-AU.

4. AdVersa-SD

This section presents our AdVersa-SD with the abductive text-video coherent learning for ego-view accident video understanding. As aforementioned, we partition each accident video into three video segments, *i.e.*, the normal video segment V_o, the near-accident segment V_r, and the accident segment V_a. Correspondingly, we annotate the text descriptions of the accident reason t_r, the prevention advice t_p, and the accident category t_a. To be clear, we define a denotation of text-video Co-oCcurrence Pair (**Co-**

[2]DiffusionDet is experimentally compared with 11 state-of-the-art detectors in Tab. 3 after fine-tuning it on the fine annotation set.

[3](1) general distractor: distracted driving, speeding, extreme weather, etc. (2) location distractor: sudden overtaking or lane-changing in the main road, too fast in turning, running the red light in intersection, etc. (3) keyword distractor: motorcycle→ cyclist, car→ ego-car, decelerate→ accelerate, stop→ start, etc. (4) random distractor: random select one reason description in the reason set.

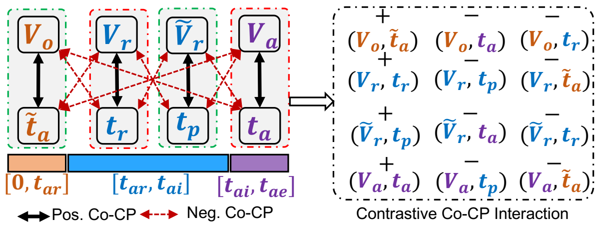

Figure 4. The structure of **Abductive CLIP** contains four interaction groups with one positive Co-CP and two negative Co-CPs for each interaction group, where t_{ar}=t_{ai}-40.

CP) to represent the natural co-occurrence of video clip and text description, *e.g.*, (V_r, t_r), and (V_a, t_a).

4.1. Abductive CLIP

For abductive video understanding, we propose an Abductive CLIP for AdVersa-SD to fulfill the coherent semantic learning within different Co-CPs. The structure of Abductive CLIP is illustrated in Fig. 4. We create two virtual Co-CPs for training Abductive CLIP, *i.e.*, the (V_o, $\tilde{t}_a$) and ($\tilde{V}_r$, t_p). (V_o, $\tilde{t}_a$) represents the Co-CP of antonymous accident category description $\tilde{t}_a$ and normal video clip V_o, while ($\tilde{V}_r$, t_p) denotes the Co-CP of $\tilde{V}_r$ and the prevention advice description t_p. Notably, $\tilde{t}_a$ is obtained by adding the antonym of verbs to the accident category description, such as "does/do not", "are not", etc. In addition, to make a dissipation process from the near-accident state to the normal state, $\tilde{V}_r$ is obtained by reverse frame rearrangement of V_r.

Certainly, because each video segment may have different numbers of video frames, we create Co-CPs by coupling the text description with the randomly selected 16 successive frames within the certain video segment. Consequently, the training for Abductive CLIP is straightforward by enhanced difference learning of the embeddings of Co-CPs.

Contrastive Interaction Loss: Abductive CLIP takes the XCLIP model [42] as the backbone. Subsequently, we input each Co-CP into XCLIP to obtain the feature embedding of the video clip feature $\mathbf{z}_v$ and the text feature $\mathbf{z}_t$. To achieve the purpose stated in Fig. 4, we provide a Contrastive Interaction Loss (CILoss) to make the interactive Co-CP learning. The CILoss for different interaction groups of Co-CPs is the same, and consistently defined as:

$$\mathcal{L}_{\text{CILoss}} = -\sum_{i=1}^{B} \log \frac{E(\mathbf{z}^p_{v_i}, \mathbf{z}^p_{t_i})}{\mathcal{K}}, \quad , \qquad (1)$$

$$\mathcal{K} = \sum_{j=1}^{B} [E(\mathbf{z}^p_{v_i}, \mathbf{z}^p_{t_j}) + E(\mathbf{z}^p_{v_i}, \mathbf{z}^{n_1}_{t_{j\neq i}}) + E(\mathbf{z}^p_{v_i}, \mathbf{z}^{n_2}_{t_{j\neq i}})],$$

where $E(\mathbf{z}_v, \mathbf{z}_t) = e^{\mathbf{z}_v{}^T \mathbf{z}_t / \tau}$ computes the coherence degree of video clip feature $\mathbf{z}_v$ and text feature $\mathbf{z}_t$. B denotes the batchsize scale, the upscripts p and n_1/n_2 refer to the *Pos. CoCP* and the *Neg. CoCPs*. τ is a learnable hyper-parameter, i and j are the sample indexes in each batchsize.

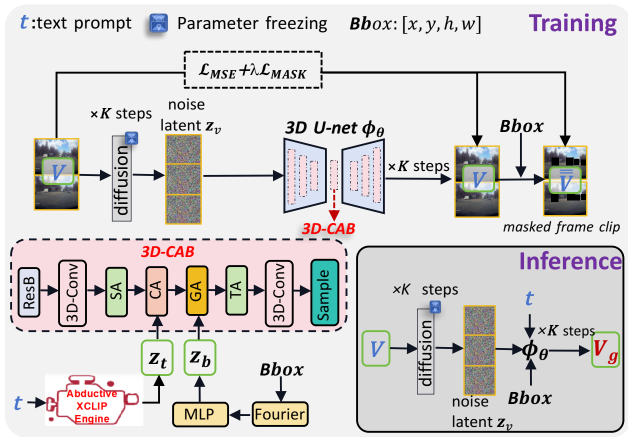

Figure 5. The structure of **OAVD**, where V and t are the video frame clip and the corresponding text prompt. Object bounding boxes $Bbox$ in each video frame can be obtained by object detectors. $\overline{\overline{V}}$ is the masked frame clip where the pixels in object regions are set to 0. Different attention modules, *i.e.*, SA, CA, and TA, follow the ones of [61] while with a dense multi-head attention.

CILoss aims to enhance the coherence between the text description and video frames in Co-CP by enlarging the distance of text descriptions or video frames with the ones in negative Co-CPs. Abductive CLIP is optimized by minimizing the summation of four kinds of $\mathcal{L}^{o,r,p,a}_{\text{CILoss}}$.

4.2. Extension to Accident Video Diffusion

To verify Abductive CLIP, this work treats it as an engine to drive the accident video diffusion task for finding the dominant objects in accident occurrence. Because traffic accidents are commonly caused by the irregular or sudden movement of road participants, the video diffusion model should have the ability for object-level representation. As shown in Fig. 5, we propose an Object-centric Accident Video Diffusion model (OAVD) which takes the Latent Diffusion Model (LDM) [46] as the baseline and extends it to the video diffusion with the input of Co-CPs and K steps of forward and reverse diffusion process.

The structure of OAVD becomes similar to Tune-A-Video work [61], while differently, the 3D-CAB block of the 3D U-net module in Fig. 5 is redesigned and contains the 3D-Conv layers (with four layers with the kernel size of (3,1,1)), a text-video Cross-Attention (**CA**) layer, a Spatial Attention (**SA**) layer, a Temporal Attention (**TA**) layer and a Gated self-attention (**GA**) [34] for the frame correlation and object location consideration. To be capable of object-level video diffusion, we further add a masked representation reconstruction path on frame-level reconstruction. The **inference** phase has the same input form as the OAVD training and generates the new frame clip V_g.

Masked Video Frame Diffusion: Our aim is to learn the 3D U-net by forward adding noise on the clean latent representation of raw video frames and inverse denoising the noise $\mathbf{z}_v$ with K time steps, conditioned by the text prompt t and the bounding boxes. This formulation enables the derivation of the mask video representation $\mathbf{z}_{mask}$ of $\overline{\overline{V}}$ to fix the frame background details in diffusion process and fulfill the object-centric video generation. The optimization of 3D U-net is achieved by minimizing:

$$\begin{aligned}&\mathbb{E}_{V,\mathbf{e}\sim\mathcal{N}(0,I),k,\mathbf{z}_t,\mathbf{z}_b,\overline{\overline{V}}}\Big\{\|\mathbf{e}-\phi_\theta(\mathbf{z}_v,k,\mathbf{z}_t,\mathbf{z}_b)\|_2^2+\\ &\lambda\|\mathbf{e}(\mathbf{1}-\mathbf{z}_{mask})-\phi_\theta(\mathbf{z}_v,k,\mathbf{z}_t,\mathbf{z}_b)(\mathbf{1}-\mathbf{z}_{mask})\|_1^1\Big\},\end{aligned} \tag{2}$$

where the first term is Mean Square Error ($\mathcal{L}_{MSE}$) and the second term denotes the reconstruction loss of the Masked Latent Representation ($\mathcal{L}_{MASK}$). $k \in [1,\dots,K]$ denotes the diffusion step (K=1000), $\mathbf{e}$ is the ground-truth noise representation in each diffusion step. ϕ_θ is the 3D U-net to be optimized which contains the parameters of 3D Cross-Attention Blocks (**3D-CAB**) with down-sample and up-sample layers (**Sample**). λ=0.5 is a parameter for balancing the weights of $\mathcal{L}_{MSE}$ and $\mathcal{L}_{MASK}$. $\mathbf{1}$ is an identity tensor with the same size of $\mathbf{e}$, and the masked noise representation $\mathbf{z}_{mask}$ is obtained by Denoising Diffusion Probabilistic Model (DDPM) Scheduler [20] on the binarization of a latent representation $\mathbf{z}_l$ ($\mathbf{z}_l$ = VAE($\mathbf{z}_v$)) through the Variational Autoencoder (VAE) in LDM [46] as:

$$\begin{aligned}\mathbf{z}_{mask} &= \text{DDPM Scheduler}(\mathbf{m}_{(z)},k,\mathbf{e})),\\ \mathbf{m}_{(z)} &= \begin{cases}0 & \text{if } \mathbf{z}_l < 0.5,\\ 1\end{cases}\end{aligned} \tag{3}$$

Gated Bbox Representation: The key insight is to involve object bounding boxes to enhance the causal object region learning concerning the related text words, which is useful for eliminating the influence of the frame background and explicitly checking the role of certain text words for subsequent accident situations. Inspired by the Gated self-Attention (GA) [34], the location embedding $\mathbf{z}_b$, collaborating with the output of CA in 3D-CAB, is obtained from Bbox by MLP layers with the Fourier embedding [41]:

$$\mathbf{z}_b = \text{MLP}(\text{Fourier}(Bbox)). \tag{4}$$

OAVD Inference: The OAVD inference stage inputs the Co-CPs while Denosing Diffusion Implict Model (DDIM) scheduler [50] is taken on the trained 3D U-net ϕ_θ conditioned by the text prompt t and $Bbox$. The frame clip V within the Co-CPs are fed into the inference stage with the same dimension to the generated V_g.

5. Experiments

5.1. Experimental Details

AdVersa-SD takes as input the object bounding boxes and accident reasons. As a prerequisite, we first carry out a benchmark evaluation for the Object Detection (**OD**) and Accident reason Answering (**ArA**) tasks, which is of crucial importance to video diffusion. Second, we evaluate our **AdVersa-SD** with extensive video diffusion experiments.

(1) OD Task: We select 11 state-of-the-art detectors to be presented in Tab. 3 in the OD benchmark evaluation. All detectors used the corresponding architectures provided in MMDetection [6] and MMYOLO [4] while keeping important hyperparameters equal, such as batch size, initial learning rate, and epochs. All training and inference are implemented on three GeForce RTX 3090s. All the detectors are pre-trained on the BDD-100K dataset [69], and fine-tuned with the training set of fine object annotations. Notably, to check the OD performance in accident window, we provide two versions of detectors fine-tuned on the training set coming from whole frame windows (abbrev., V1-Train [, ,]) and the ones fine-tuned on the training set of accident-free windows (abbrev., V2-Train [,]).

We use Average Precision (**AP50**) and Average Recall (**AR**) [35] to evaluate the detection results with the threshold of 50% detection score.

(2) ArA Task: We follow the task of multi-choice Video Question Answering (VQA) to formulate the ArA task while the question is "`What is the reason for the accident in this video?`". The performance is measured by the **accuracy**, *i.e.*, the percentage of questions that are correctly answered.

(3) Abductive Video Diffusion Task: In AdVersa-SD, we aim to abductive ego-view accident understanding conditioned by the descriptions of accident reasons or prevention advice. Hence, based on the input form of OAVD in AdVersa-SD, we input the object bboxes and Co-CPs of (V_r, t_r) and (V_r, t_p) in the evaluations. Two state-of-the-art video diffusion models including the DDIM inversion-based Tune-A-Video (TAV) [61] and the training-free ControlVideo (CVideo) [72] are selected. We generate 1500 clips (with 16 frames/clip) for all diffusion experiments, where the object bboxes are pre-detected by DiffusionDet [7]. Similar to previous video diffusion models, Fréchet Video Distance (**FVD**) [55] is taken for quality evaluation of the synthetic video clips. We also use the CLIP score (**CLIP**$_S$) [61] to measure the alignment degree between text prompts and video frames.

In the evaluation, 6000 pairs of Co-CPs in MM-AU are adopted to train the AdVersa-SD and the Tune-A-Video model. The learning rate of Abductive CLIP in AdVersa-SD is $1e-6$ with the batchsize of 2 and trained with 30000 iteration steps. The learning rate of OAVD is $5e-6$ with the same batchsize and trained with 8000 iteration steps. Adam optimizer is adopted with the default β_1 = 0.9 and $\beta_2 = 0.999$ on a platform with 2 GeForce RTX 3090s.

[4] https://github.com/open-mmlab/mmyolo

Table 3. The results of V1-Train [, ,]) and V2-Train [,]) for 11 state-of-the-art detectors on the MM-AU.

Detectors	Years	V1-Train [, ,] val. [, ,] mAP50	AR	test. [, ,] mAP50	AR	test. [ ] mAP50	AR	V2-Train [,] test. [,] mAP50	AR	test. [ ] mAP50	AR	Anchor	GFlops	#Params.
FasterRCNN [45]	2015	0.674	0.634	0.666	0.623	0.664	0.620	0.544	0.524	0.497	0.509	✓	0.19T	41.38M
CornerNet [30]	2018	0.495	0.625	0.485	0.619	0.483	0.624	0.436	0.563	0.456	0.598		0.71T	201M
CascadeRPN [58]	2019	0.662	0.699	0.664	0.696	0.649	0.689	0.579	0.663	0.532	0.624	✓	0.18T	41.97M
CenterNet [10]	2019	0.054	0.238	0.051	0.233	0.047	0.224	0.161	0.260	0.155	0.257		20.38G	14.21M
DETR [4]	2020	0.367	0.407	0.377	0.403	0.363	0.403	0.275	0.329	0.254	0.318		44.55G	28.83M
EfficientNet [53]	2020	0.310	0.417	0.310	0.412	0.293	0.404	0.075	0.128	0.073	0.133		57.28G	18.46M
Deformable-DETR [74]	2021	0.660	0.671	0.661	0.668	0.652	0.663	0.626	0.631	0.587	0.626		0.18T	40.1M
YOLOx [16]	2021	0.673	0.709	0.672	0.698	0.670	0.698	0.563	0.627	0.540	0.626		13.33G	8.94M
YOLOv5s [23]	2021	**0.757**	**0.766**	**0.748**	**0.764**	**0.743**	**0.761**	0.660	0.716	0.636	0.712	✓	8.13G	12.35M
DiffusionDet [7]	2023	0.731	0.749	0.733	0.745	0.718	0.738	**0.701**	**0.729**	**0.660**	**0.716**		-	26.82M
YOLOv8 [54]	2023	0.716	0.754	0.715	0.753	0.717	0.755	0.606	0.702	0.597	0.703		14.28G	11.14M

Notes: To ensure that the distribution of accident windows is essentially the same on the training, validation, and testing sets, we divide the bounding box annotations in the ratio of 7:1.5:1.5. The results of GFlops and #Params. are reported by the tools in MMDetection and MMYOLO of the testing phase.

5.2. Result Analysis

(1) OD Evaluations: Tab. 3 presents the detection results of 11 state-of-the-art detectors. From the results, we can see all the detectors generate a degradation for the accident window test. CenterNet [10] and EfficientNet [53] show limited ability for the OD task in the accident scenarios and the metric values decrease significantly for V2-Train mode. As claimed by previous research, pure Transformer-based detectors, such as DETR [4], demonstrate limited performance. Deformable-DETR has improved performance but is still not better than the CNN-based ones for traffic accident cases. YOLOv5s [23] and DiffusionDet [7] are the two leading approaches for the OD task. However, from the results difference obtained by V1-Train [, ,] and V2-Train [,], DiffusionDet [7] shows superior performance to the testing set of accident window in V2-Train. It indicates that diffusion-based object detection may be more robust with better generalization ability. More qualitative results can be viewed in the supplemental file.

(2) ArA Evaluations: We present in Tab. 4 and Fig. 6 the performances of the state-of-the-art on the ArA task. We carefully select the baseline methods to include temporal relation network (HCRN [31]), graph transformer network (VGT [63], CoVGT [64]), cross-modal pre-trained transformers (ClipBERT [33]) and those using large language models (LMMs) (FrozenGQA [62] and SeViLA [70]). The methods also include frame-centric and more fine-grained object-centric video representations. Our key observations are: LLM-based methods, such as SeViLA[5] which uses Flan T5-XL (3B vs less than 1B of other competitors) [8], show absolute advantage in this task, surpassing the second-ranked method CoVGT by 7% to 9%. Furthermore, fine-grained visual representations, *e.g.*, region or object level, are key for higher performances. We speculate that the videos are all taken on the road about traffic accidents and thus a coarse frame-level representation is insufficient to discern various accident reasons.

[5]We take the pre-trained SeViLA and fine-tune it on the ArA training set by 10 epochs with a learning rate of 1e-5.

Table 4. The Accident Reason Answering (ArA) Accuracy (Acc. %) on the validation (val.) and testing (test.) set of MM-AU by 6 SOTA methods whose size of learnable parameters is provided.

Methods	Years	Acc (val.)	Acc (test.)	V.	T.	Params.(M)
HCRN [31]	2020	65.81	64.65	F	G	42
ClipBERT [33]	2021	72.09	72.71	F	B	137
VGT [63]	2022	68.40	68.66	O	B	143
FrozenGQA [62]	2023	77.10	77.01	F	D	30
CoVGT [64]	2023	81.70	79.97	O	R	159
SeViLA [70]	2023	**89.26**	**89.02**	O	F	108

F: frame-centric representations; O: object-centric representations; V.:Vision; T.:Text; G: GloVe; B: BERT [9]; D: DeBERTa [19]; R: RoBERTa [38]; F: Flan T5 [8]. The ratio of training, validation, and testing set is 7:1:2.

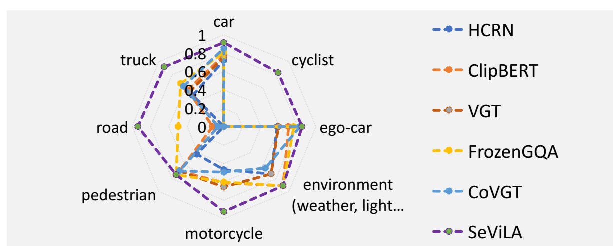

Figure 6. Accident Reason Answering (ArA) Accuracy (Acc. %), w.r.t., different accident participants on the testing set of MM-AU.

An in-depth analysis in Fig. 6 shows that the methods' performances vary a lot across different accident participant types. Generally, SeViLA outperforms other methods in all scenarios. Yet, all methods perform closely when the accident reason is car-related. Curiously, we find that most of the methods fail to identify the cyclists in the accidents except for SeViLA. The reason could be that cyclists are the least frequent participants (as shown in Fig. 3 (b)) in accidents. Thus, it calls for commonsense knowledge (carried in LLMs) to find the related reasons.

(3) Diffusion Evaluations: Here, we evaluate the AdVersa-SD, and the importance of Abductive-CLIP and

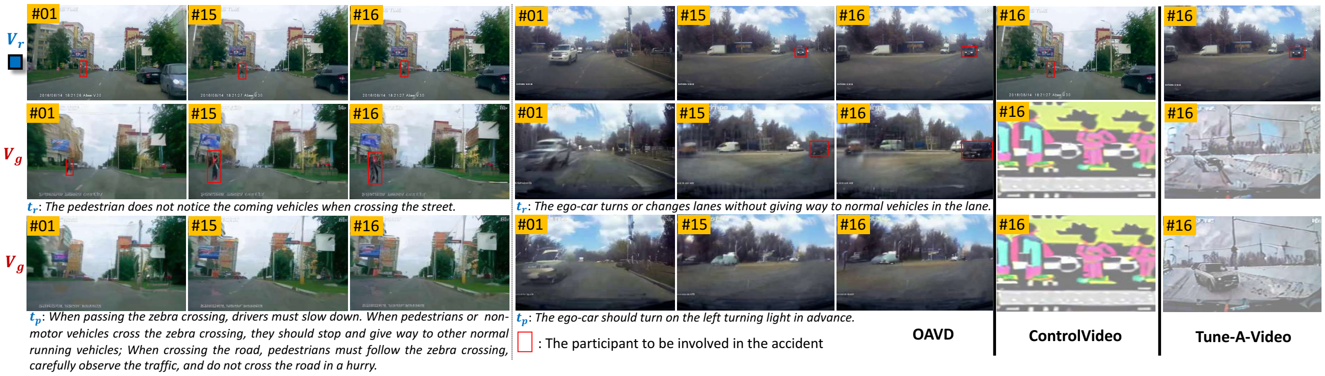

Figure 7. Some results (V_g) inputting the Co-CPs of (V_r, t_r) or (V_r, t_p). From the generation of OAVD, the participants to be involved in accidents appear in advance when giving the accident reason prompt t_r, while the accident objects disappear when providing the prevention advice prompt t_p. ControlVideo and Tune-A-Video are given the same t_r and t_p prompt with OAVD, respectively. However, the artifacts and unrelated content are generated, and the phenomena of "*appear in advance* " or "*disappear*" do not occur.

Table 5. Results with SOTA diffusion models and our OAVD driven by varying CLIP models, where ↓ and ↑ prefer a lower and larger value, respectively. FPS: frames/second (tested on a single GeForce RTX 3090).

Method	TAV [61]	CVideo [72]	OAVD (CLIP [44])	OAVD (S-CLIP)*	OAVD (A-CLIP)*
$CLIP_S$ ↑	21.77	22.51	21.9	27.14	**27.24**
FVD ↓	9545.6	12275.2	10122.5	5372.3	**5238.1**
FPS ↑	1.7	0.5	1.7	1.2	1.2

***: with the input of bounding boxes.**

Bboxes. To verify Abductive-CLIP in AdVersa-SD, we take two baselines: 1) the original CLIP model [44] and 2) a "Sequential-CLIP (S-CLIP)" that only maintains the positive Co-CPs of the Abductive-CLIP (A-CLIP) (see Fig. 4).

Abductive Ability Check of AdVesrsa-SD. Fig. 7 visually presents video diffusion results of accident scenarios given the descriptions of accident reason t_r and prevention advice t_p, respectively. Curiously, OAVD can make the accident participant (*i.e.*, the pedestrian or the black car) appear in advance given t_r, and eliminate the accident participants provided t_p. It indicates that our AdVesrsa-SD catches the dominant object representation for the accident occurrence. Contrarily, ControlVideo and Tune-A-Video generate irrelevant styles with worse performance than OAVD, as listed in Tab. 5, which shows that accident knowledge is scarce in this field but rather crucial for the accident video diffusion models. Tab. 5 shows the results of different diffusion models.

Roles of different CLIP Models. Tab. 5 also presents the results of OAVD with varying CLIP models. The results show that our Abductive-CLIP can generate better text-video semantic alignment than the original CLIP model and the Sequential-CLIP trained on our MM-AU. It indicates that the contrastive interaction loss of the text-video pairs, *i.e.*, Co-CPs, is important to discern the key semantic information within text and videos.

Figure 8. A visualization for the importance of bounding box.

Role of Bboxes From Tab. 5 and Fig. 8, the advantages of Bboxes are demonstrated with clearer and more detailed content in the generated frames. In addition, object-involved video diffusion can facilitate the key object region learning and maintain the details of the frames better than the version without Bbox input. More ablation studies on the role of bboxes can be viewed in the supplemental file.

Besides, our OAVD also can flexibly generate any accident videos with the input of only object boxes or the accident category descriptions (see the supplemental file).

6. Conclusion

This work presents a precious large-scale ego-view multi-modal accident dataset (MM-AU) for safe driving perception which provides the temporal, object, and text annotations for fine-grained accident video understanding. Within MM-AU, the evaluations of the state-of-the-art methods on object detection and accident reason answering tasks are carefully conducted. Based on MM-AU, we present AdVersa-SD to fulfill an abductive accident video understanding, where an Object-centric Accident Video Diffusion (OAVD) driven by an Abductive-CLIP model is proposed. Extensive experiments verify that AdVersa-SD shows promising ability for the abductive accident video understanding and generates superior video diffusion performance to two state-of-the-art diffusion models.

Acknowledgements: This research is supported by NSFC 62036008 and 62273057. We also thank Kuan Yang for the accident reason annotation.

References

[1] Mohammad Sadegh Aliakbarian, Fatemeh Sadat Saleh, Mathieu Salzmann, Basura Fernando, Lars Petersson, and Lars Andersson. VIENA: A driving anticipation dataset. In *ACCV*, pages 449–466, 2019. 3

[2] Wentao Bao, Qi Yu, and Yu Kong. Uncertainty-based traffic accident anticipation with spatio-temporal relational learning. In *ACM MM*, pages 2682–2690, 2020. 2, 3

[3] Wentao Bao, Qi Yu, and Yu Kong. DRIVE: deep reinforced accident anticipation with visual explanation. In *ICCV*, pages 7599–7608, 2021. 1, 2

[4] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, pages 213–229, 2020. 7, 2, 3, 4

[5] Fu-Hsiang Chan, Yu-Ting Chen, Yu Xiang, and Min Sun. Anticipating accidents in dashcam videos. In *ACCV*, volume 10114, pages 136–153, 2016. 2, 3

[6] Kai Chen et al. MMDetection: Open mmlab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*, 2019. 6

[7] Shoufa Chen, Peize Sun, Yibing Song, and Ping Luo. Diffusiondet: Diffusion model for object detection. In *ICCV*, pages 19830–19843, 2023. 4, 6, 7, 2, 3

[8] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*, 2022. 7

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 7

[10] Kaiwen Duan, Song Bai, Lingxi Xie, Honggang Qi, Qingming Huang, and Qi Tian. Centernet: Keypoint triplets for object detection. In *ICCV*, pages 6569–6578, 2019. 7, 2, 3

[11] Editorial. Safe driving cars. *Nat. Mach. Intell.*, 4:95–96, 2022. 1

[12] Patrick Esser, Johnathan Chiu, Parmida Atighehchian, Jonathan Granskog, and Anastasis Germanidis. Structure and content-guided video synthesis with diffusion models. In *ICCV*, pages 7346–7356, 2023. 1

[13] Jianwu Fang, Jiahuan Qiao, Jie Bai, Hongkai Yu, and Jianru Xue. Traffic accident detection via self-supervised consistency learning in driving scenarios. *IEEE Trans. Intell. Transp. Syst.*, 23(7):9601–9614, 2022. 2, 3

[14] Jianwu Fang, Dingxin Yan, Jiahuan Qiao, Jianru Xue, He Wang, and Sen Li. DADA-2000: can driving accident be predicted by driver attention? Analyzed by A benchmark. In *ITSC*, pages 4303–4309, 2019. 4

[15] Jianwu Fang, Dingxin Yan, Jiahuan Qiao, Jianru Xue, and Hongkai Yu. DADA: driver attention prediction in driving accident scenarios. *IEEE Trans. Intell. Transp. Syst.*, 23(6):4959–4971, 2022. 1, 2, 3

[16] Zheng Ge, Songtao Liu, Feng Wang, Zeming Li, and Jian Sun. YOLOX: exceeding YOLO series in 2021. *CoRR*, abs/2107.08430, 2021. 4, 7, 2, 3

[17] Sreyan Ghosh, Sherwin Joseph Sunny, and Rohan Roney. Accident detection using convolutional neural networks. In *IconDSC*, pages 1–6, 2019. 3

[18] Feten Hajri and Hajer Fradi. Vision transformers for road accident detection from dashboard cameras. In *AVSS*, pages 1–8, 2022. 2

[19] Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*, 2020. 7

[20] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *NeurIPS*, 33:6840–6851, 2020. 6

[21] Hongyu Hu, Qi Wang, Ming Cheng, and Zhenhai Gao. Cost-sensitive semi-supervised deep learning to assess driving risk by application of naturalistic vehicle trajectories. *Expert Syst. Appl.*, 178:115041, 2021. 2

[22] Ashesh Jain, Avi Singh, Hema Swetha Koppula, Shane Soh, and Ashutosh Saxena. Recurrent neural networks for driver activity anticipation via sensory-fusion architecture. In *ICRA*, pages 3118–3125, 2016. 2

[23] Glenn Jocher et al. ultralytics/yolov5: v6.2 - YOLOv5 Classification Models, Apple M1, Reproducibility, ClearML and Deci.ai integrations, 2022. 7, 2, 3, 4

[24] Minhee Kang, Wooseop Lee, Keeyeon Hwang, and Young Yoon. Vision transformer for detecting critical situations and extracting functional scenario for automated vehicle safety assessment. *Sustainability*, 14(15):9680, 2022. 1, 3

[25] Muhammad Monjurul Karim, Yu Li, Ruwen Qin, and Zhaozheng Yin. A dynamic spatial-temporal attention network for early anticipation of traffic accidents. *IEEE Trans. Intell. Transp. Syst.*, 23(7):9590–9600, 2022. 2

[26] Muhammad Monjurul Karim, Yu Li, Ruwen Qin, and Zhaozheng Yin. A dynamic spatial-temporal attention network for early anticipation of traffic accidents. *IEEE Trans. Intell. Transp. Syst.*, 23(7):9590–9600, 2022. 2

[27] Muhammad Monjurul Karim, Zhaozheng Yin, and Ruwen Qin. An attention-guided multistream feature fusion network for early localization of risky traffic agents in driving videos. *IEEE Trans. Intell. Veh.*, 9(1):1792 – 1803, 2024. 2, 3

[28] Hoon Kim, Kangwook Lee, Gyeongjo Hwang, and Changho Suh. Crash to not crash: Learn to identify dangerous vehicles using a simulator. In *AAAI*, pages 978–985, 2019. 3

[29] Bulbula Kumeda et al. Vehicle accident and traffic classification using deep convolutional neural networks. In *International Computer Conference on Wavelet Active Media Technology and Information Processing*, pages 323–328, 2019. 3

[30] Hei Law and Jia Deng. Cornernet: Detecting objects as paired keypoints. In *ECCV*, pages 765–781, 2018. 7, 2, 3, 4

[31] Thao Minh Le, Vuong Le, Svetha Venkatesh, and Truyen Tran. Hierarchical conditional relation networks for video question answering. In *CVPR*, pages 9972–9981, 2020. 7, 4

[32] Trung-Nghia Le, Shintaro Ono, Akihiro Sugimoto, and Hiroshi Kawasaki. Attention R-CNN for accident detection. In *IV*, pages 313–320, 2020. 2

[33] Jie Lei, Linjie Li, Luowei Zhou, Zhe Gan, Tamara L Berg, Mohit Bansal, and Jingjing Liu. Less is more: Clipbert for

video-and-language learning via sparse sampling. In *CVPR*, pages 7331–7341, 2021. 7

[34] Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. Gligen: Open-set grounded text-to-image generation. In *CVPR*, pages 22511–22521, 2023. 5, 6, 1

[35] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common objects in context. In *ECCV*, pages 740–755, 2014. 4, 6

[36] Chunsheng Liu, Zijian Li, Faliang Chang, Shuang Li, and Jincan Xie. Temporal shift and spatial attention-based two-stream network for traffic risk assessment. *IEEE Trans. Intell. Transp. Syst.*, 23(8):12518–12530, 2022. 3

[37] Yang Liu, Guanbin Li, and Liang Lin. Cross-modal causal relational reasoning for event-level visual question answering. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(10):11624–11641, 2023. 1, 3

[38] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019. 7

[39] Haohan Luo and Feng Wang. A simulation-based framework for urban traffic accident detection. In *ICASSP*, pages 1–5, 2023. 3

[40] Arnav Vaibhav Malawade, Shih-Yuan Yu, Brandon Hsu, Deepan Muthirayan, Pramod P. Khargonekar, and Mohammad Abdullah Al Faruque. Spatiotemporal scene-graph embedding for autonomous vehicle collision prediction. *IEEE Internet Things J.*, 9(12):9379–9388, 2022. 2

[41] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021. 6

[42] Bolin Ni, Houwen Peng, Minghao Chen, Songyang Zhang, Gaofeng Meng, Jianlong Fu, Shiming Xiang, and Haibin Ling. Expanding language-image pretrained models for general video recognition. In *ECCV*, pages 1–18. Springer, 2022. 5

[43] Karishma Pawar and Vahida Attar. Deep learning based detection and localization of road accidents from traffic surveillance videos. *ICT Express*, 8(3):379–387, 2022. 2

[44] 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 *ICML*, pages 8748–8763, 2021. 1, 8

[45] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. *NeurIPS*, 28, 2015. 7, 2, 3

[46] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10684–10695, 2022. 2, 5, 6, 1

[47] Debaditya Roy, Tetsuhiro Ishizaka, C Krishna Mohan, and Atsushi Fukuda. Detection of collision-prone vehicle behavior at intersections using siamese interaction lstm. *IEEE Trans. Intell. Transp. Syst.*, 23(4):3137–3147, 2020. 2

[48] Kelathodi Kumaran Santhosh, Debi Prosad Dogra, Partha Pratim Roy, and Adway Mitra. Vehicular trajectory classification and traffic anomaly detection in videos using a hybrid cnn-vae architecture. *IEEE Trans. Intell. Transp. Syst.*, 23(8):11891–11902, 2021. 2

[49] Dinesh Singh and Chalavadi Krishna Mohan. Deep spatio-temporal representation for detection of road accidents using stacked autoencoder. *IEEE Trans. Intell. Transp. Syst.*, 20(3):879–887, 2019. 2

[50] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *ICLR*, 2021. 6

[51] Tomoyuki Suzuki, Hirokatsu Kataoka, Yoshimitsu Aoki, and Yutaka Satoh. Anticipating traffic accidents with adaptive loss and large-scale incident DB. In *CVPR*, pages 3521–3529, 2018. 2

[52] Leonardo Taccari, Francesco Sambo, Luca Bravi, Samuele Salti, Leonardo Sarti, Matteo Simoncini, and Alessandro Lori. Classification of crash and near-crash events from dashcam videos and telematics. In *ITSC*, pages 2460–2465, 2018. 2

[53] Mingxing Tan and Quoc V. Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *ICML*, volume 97, pages 6105–6114, 2019. 7, 2, 3

[54] Ultralytics. Ultralytics github repository. https://github.com/ultralytics/ultralytics, November 2023. 7, 2, 3

[55] Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*, 2018. 6, 8

[56] Thakare Kamalakar Vijay, Debi Prosad Dogra, Heeseung Choi, Gi Pyo Nam, and Ig-Jae Kim. Detection of road accidents using synthetically generated multi-perspective accident videos. *IEEE Trans. Intell. Transp. Syst.*, 24(2):1926–1935, 2023. 2

[57] Vikram Voleti, Alexia Jolicoeur-Martineau, and Chris Pal. Mcvd-masked conditional video diffusion for prediction, generation, and interpolation. *NeurIPS*, 35:23371–23385, 2022. 1

[58] Thang Vu, Hyunjun Jang, Trung X. Pham, and Chang Dong Yoo. Cascade RPN: delving into high-quality region proposal network with adaptive convolution. In *NeurIPS*, pages 1430–1440, 2019. 7, 2, 3

[59] Tianhang Wang, Kai Chen, Guang Chen, Bin Li, Zhijun Li, Zhengfa Liu, and Changjun Jiang. GSC: A graph and spatio-temporal continuity based framework for accident anticipation. *IEEE Trans. Intell. Veh.*, 9(1):2249–2261, 2023. 1, 2

[60] Tianqi Wang, Sukmin Kim, Wenxuan Ji, Enze Xie, Chongjian Ge, Junsong Chen, Zhenguo Li, and Ping Luo. Deepaccident: A motion and accident prediction benchmark for V2X autonomous driving. *CoRR*, abs/2304.01168, 2023. 3

[61] Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Stan Weixian Lei, Yuchao Gu, Yufei Shi, Wynne Hsu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. Tune-a-Video: One-shot tuning of image diffusion models for text-to-video generation. In *ICCV*, pages 7623–7633, 2023. 5, 6, 8, 1, 7

[62] Junbin Xiao, Angela Yao, Yicong Li, and Tat Seng Chua. Can i trust your answer? visually grounded video question answering. *arXiv preprint arXiv:2309.01327*, 2023. 7

[63] Junbin Xiao, Pan Zhou, Tat-Seng Chua, and Shuicheng Yan. Video graph transformer for video question answering. In *ECCV*, pages 39–58, 2022. 7

[64] Junbin Xiao, Pan Zhou, Angela Yao, Yicong Li, Richang Hong, Shuicheng Yan, and Tat-Seng Chua. Contrastive video question answering via video graph transformer. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(11):13265–13280, 2023. 3, 4, 7

[65] Li Xu, He Huang, and Jun Liu. SUTD-TrafficQA: A question answering benchmark and an efficient network for video reasoning over traffic events. In *CVPR*, pages 9878–9888, 2021. 1, 3

[66] Yu Yao, Xizi Wang, Mingze Xu, Zelin Pu, Yuchen Wang, Ella M. Atkins, and David J. Crandall. Dota: Unsupervised detection of traffic anomaly in driving videos. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(1):444–459, 2023. 2, 3

[67] Yu Yao, Mingze Xu, Yuchen Wang, David J. Crandall, and Ella M. Atkins. Unsupervised traffic accident detection in first-person videos. In *IROS*, pages 273–280, 2019. 2, 3

[68] Tackgeun You and Bohyung Han. Traffic accident benchmark for causality recognition. In *ECCV*, volume 12352, pages 540–556, 2020. 3

[69] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. BDD100K: A diverse driving dataset for heterogeneous multitask learning. In *CVPR*, pages 2633–2642, 2020. 4, 6

[70] Shoubin Yu, Jaemin Cho, Prateek Yadav, and Mohit Bansal. Self-chained image-language model for video localization and question answering. *NeurIPS*, pages 76749–76771, 2023. 7, 4

[71] Chuanqi Zang, Hanqing Wang, Mingtao Pei, and Wei Liang. Discovering the real association: Multimodal causal reasoning in video question answering. In *CVPR*, pages 19027–19036, 2023. 3

[72] Yabo Zhang, Yuxiang Wei, Dongsheng Jiang, Xiaopeng Zhang, Wangmeng Zuo, and Qi Tian. Controlvideo: Training-free controllable text-to-video generation. *arXiv preprint arXiv:2305.13077*, 2023. 6, 8, 7

[73] Zhili Zhou, Xiaohua Dong, Zhetao Li, Keping Yu, Chun Ding, and Yimin Yang. Spatio-temporal feature encoding for traffic accident detection in vanet environment. *IEEE Trans. Intell. Transp. Syst.*, 23(10):19772–19781, 2022. 2

[74] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. In *ICLR*, 2021. 7, 2, 3