ADVersa: Abductive Driving Accident Video Understanding

Lei-Lei Li, Jianwu Fang, *Member, IEEE*, Junbin Xiao, Hongkai Yu, Chen Lv, *Senior Member, IEEE*, Jianru Xue, *Member, IEEE*, Zhengguo Li, *Fellow, IEEE*, and Tat-Seng Chua

***Abstract*—Understanding traffic accident scenes is a long-standing research for vision-based safe driving. It seeks to answer why accidents occur, how near-crash scenes develop, and what the key elements of an accident are. This research is challenging due to the scarcity and fragmentation of accident data, as well as the complex accident environments. To study this, we present a framework of Abductive Driving accident Video understanding (ADVersa), which infers a plausible visual and textual explanation for the absent near-crash scenes. ADVersa underscores three groups of tasks: 1) visual past recovery of near-crash scenes, 2) visual prediction of near-crash scenes, and 3) accident cause involved video synthesis. To support the study, we first contribute MM-AU, a novel dataset for Multi-Modal Accident video Understanding. MM-AU contains 11,727 in-the-wild driving accident videos with temporally aligned text descriptions, 2.23 million well-annotated object boxes, and 58,650 pairs of video-based accident cause texts. We then propose an Abductive CLIP model and a Contrastive Graph Video Pre-training (CGVP) model which exploit relation-aware cross-modal semantic learning to drive *spatially abductive* and *temporally abductive* accident video diffusion. Extensive experiments verify the superiority of ADVersa to the state-of-the-art approaches on different tasks, *i.e.*, historical near-crash video frame recovering, crashing video frame prediction, textual accident cause and category reasoning, normal-to-accident video synthesis and accident video editing. With these efforts, we hope this research can advance the progress on multimodal accident video understanding. The code, datasets, and more results are released at www.lotvsmmau.net.**

***Index Terms*—Safe-driving perception, traffic accident understanding, video diffusion model, scene-graph, abductive learning**

I. Introduction

AUTONOMOUS Vehicles (AV) are around the corner for practical use [1]. Yet, occasional traffic accidents are among the biggest obstacles to be crossed. To take a step further, it is urgent to comprehensively understand the whole story of a driving accident, such as why an accident occurs, how near-crash scenes develop, and what the key elements of the accident are. To this end, we present a framework of Abductive Driving accident Video understanding (**ADVersa**), which infers a plausible visual and textual explanation for the absent near-crash scenes. ADVersa underscores three groups of tasks: 1) **visual past recovery of near-crash scenes**, 2) **visual prediction of near-crash scenes**, and 3) **accident cause involved video synthesis**.

To facilitate the study on these tasks, we first construct **MM-AU**, a multi-modal dataset for driving accident video understanding. MM-AU contains **11,727** in-the-wild driving accident videos with 2.2 million frames. The videos are temporally aligned with the text descriptions of normal scene descriptions, accident causes, prevention solutions, and accident categories. To enable accident cause reasoning, **58.6K** pairs of video-based accident cause answers are annotated for 58 accident categories. Additionally, to support fine-grained and object-centric accident video understanding, we automatically obtain the segment maps, depth maps, and tracklets of **2.23M** manually annotated objects across about **463K** video frames.

Fig. 1 illustrates the tasks that MM-AU support. We divide a driving accident scenario with text and visual explanations into accident-free scene description (t_o), accident cause description (t_r), and accident category description (t_a), aligned with temporal accident-free window (V_o ■), the near-crash window (V_r ■), and the collision window (V_a ■), respectively. MM-AU can facilitate 10 traffic accident video understanding tasks: ① traffic object detection, ② accident video classification, ③ temporal accident detection and ④ anticipation, ⑤ critical object grounding, ⑥ accident cause answering, ⑦ key action grounding of accidents, ⑧ prevention advice evaluation, ⑨ multimodal accident video diffusion, and ⑩ abductive accident understanding.

While comprehensive, we focus on tasks ⑨-⑩ and design diffusion models to fulfill the *spatially and temporally* abductive understanding of driving accident videos. For other tasks, we provide benchmark results of established methods for future comparison. To be clear, we aim to infer a plausible visual or textual explanation for the *absent* near-crash or collision observations given incomplete visual premises (multimodal accident fragment observations). The absent multimodal scene inference implies a spatial-temporal **causal** understanding of a multimodal accident event.

In our implementation, we further partition ⑨-⑩ tasks into three groups of subtasks:

1) **Rec-CrashPast**→*Why an accident occurs*. It aims to recover the past near-crash frames in V_r or reason the

This work is supported by NSFC (W2411052 and 62273057).

L. Li, J. Fang, and J. Xue are with the State Key Laboratory of Human-Machine Hybrid Augmented Intelligence, Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University, Xi'an, 710049, China (670160532lileilei@gmail.com, fangjianwu@ieee.org, jrxue@mail.xjtu.edu.cn).

J. Xiao and T.S Chua are with the Department of Computer Science, National University of Singapore, Singapore ({junbin, dcscts}@nus.edu.sg).

C. Lv is with with the School of Mechanical and Aerospace Engineering, Nanyang Technological University, Singapore 639798 (e-mail: lyuchen@ntu.edu.sg).

H. Yu is with the Department of Electrical Engineering and Computer Science, Cleveland State University, Cleveland, OH 44115 USA (e-mail: h.yu19@csuohio.edu).

Z. Li is with the Institute for Infocomm Research, Agency for Science, Technology and Research (A*STAR), Singapore 138632 (e-mail: ezgli@i2r.a-star.edu.sg).

Fig. 1. **MM-AU** supports a variety of tasks for multimodal driving accident video understanding. We highlight the text descriptions for normal scene description (t_o), accident cause (t_r), and accident category (t_a), as well as temporal windows (**accident-free** V_o, **near-crash** V_r, and **collision** V_a windows) for different tasks.

accident cause text t_r from recovering frames conditioned by the observed collision scene premise [V_a, t_a].

2) **Pred-Crash**→*How the near-crash scenes develop*. It aims to predict the future crash frames in V_a or reason accident category descriptions t_a from prediction frames conditioned by the near-crash scene premise [V_r, t_r].
3) **Accident-Syn**→ *What key elements of an accident are*. It includes normal-to-accident video synthesis (N2A) and accident video editing (AEdit) conditioned by accident-free video frames, near-crash video frames, and accident-related text descriptions, respectively.

Notably, the tasks of **Rec-CrashPast** and **Pred-Crash** are conditioned by incomplete multimodal observations (*i.e.*, ***visual-text accident premises***), and **Accident-Syn** varies the visual content in observed frames.

~~In a traditional thought, these tasks may be~~ explored separately by different models [2], [3], [4]. ~~In this work,~~ we pursue a unified abductive formulation and provide a package for possible downstream safe-driving tasks, such as visual action decision [5], [6] or traffic accident anticipation [7], [8]. To fulfill this, we take the fast-paced video diffusion as the core pipeline and introduce a spatial-temporal structured representation to constrain the co-occurrence learning of text-video semantics attentively. Specifically, our framework ADVersa is built on three key components: 1) Relation-aware visual semantic alignment of accident scenes (§IV-A), which is optimized via contrastive learning between relevant (positive) and irrelevant (negative) accident causes (Abductive CLIP), and between text and video clips of graph representation (Contrastive Graph Video Pre-training (CGVP)). 2) With Abductive CLIP and CGVP, we take them as the engines to drive an Object-centric Accident Video Diffusion (§IV-B) and a Relation-centric Accident Video Diffusion (RAVD) (§IV-C) for achieving *spatially* (masking objects) and *temporally* (masking frames) abductive accident video understanding, respectively. 3) RAVD is further fine-tuned for accident cause and category reasoning (§IV-D).

This paper is an extension of our previous work in CVPR2024 [9]. We further propose a unified abductive formulation in ADVersa. The extended **contributions** are as follows.

(1) We provide object tracklets for the large-scale multimodal driving accident understanding dataset MM-AU, along with depth maps and segment instance maps for all accident videos. The dataset can support diverse accident video understanding tasks for safe driving.

(2) We formulate a unified abductive driving accident video understanding on three subtask groups, including 1) visual past recovery of near-crash scenes, 2) visual prediction of crash scenes, and 3) accident video synthesis.

(3) We propose a Contrastive Graph Video Pre-training (CGVP) model for vision-relation alignment, which drives *temporally* abductive accident video diffusion and accident cause and category reasoning, for enhancing important region semantic learning and alignment.

(4) Extensive experiments of ADVersa on three subtask groups verify the superiority of ADVersa against other counterparts, resulting in a thorough benchmark for abductive driving accident video understanding.

The remainder of this paper is organized as follows. Sec. II reviews and discusses the related work on accident video understanding and visual abductive reasoning. The annotation details are presented in Sec. III. Sec. IV presents the methodologies of ADVersa and the experimental evaluations are shown in Sec. V. Conclusion and future insights are discussed in Sec. VI.

II. Related Work

A. Multimodal Driving Accident Video Understanding

For a long time, research on monocular accident video understanding has focused primarily on tasks such as object detection, accident detection [10], [11], [12], [13], accident anticipation [7], and accident video classification [14], [15]. These tasks commonly aim to localize anomaly objects and temporal segments where accidents occur [16], [17], [18], or to predict the likelihood of future accidents for early warning [19], [20], [21], [7]. With the emergence of large language models [22] and large multimodal models [23], [24], multimodal driving accident video understanding is becoming a hot spot in the safe-driving field. Generally, this field mainly concentrates on video question answering of traffic accidents [25], dense accident video captioning [26], accident video diffusion [27], [28], or the unified formulation of these tasks [29].

Accident Cause Answering: Accident cause answering infers the most possible cause choices given a complete video sequence along with candidate causes. The milestone work SUTD-TrafficQA [30] formulates the cause explanation and prevention advice of accidents by the question-answering (QA) framework. Based on this, Liu *et al.* [31] explores the cross-modal causal relation to fulfill the traffic accident cause answering. We believe QA frameworks [32], [33] 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.

Dense Accident Video Captioning: To effectively understand what elements and their relations are in a traffic accident video, dense video captioning aims to describe the evolving content of scenes through fine-grained natural language [34], [35]. Some typical works [36], [37], [2] release a traffic video captioning benchmark in real-world driving accident or risky scenarios. Based on this, Zhou *et al.* [38] exploits domain knowledge embedded in human-annotated sentences through a cross-modal Transformer to enhance captioning quality. Despite these benchmarks bringing advances in dense accident video captioning, existing approaches still rely on static sentence-level supervision and fall short of capturing multi-phase interactions and behavioral changes that occur over extended time windows. To address this, MLLM-based frameworks [39], [26] unify an end-to-end workflow by aligning generated descriptions with accident videos, enabling more adaptive and interactive descriptions grounded in both semantics and temporal context. Although these captioning works answer to "*what is happening*", the verbose descriptions may obscure decision-critical cues and lack causal links to key objects and actions. These limitations call for unified frameworks that seamlessly integrate dense captioning with causal reasoning for accident video understanding.

Accident Video Diffusion: This is a new topic that leverages the popular video diffusion models but needs to consider the specific accident evolution rules in the diffusion process. For a long time, traffic accident video generation relied on various simulators, such as CARLA simulator [40], Carsim [41], etc. The Bird's Eye-View (BEV)-map-based participant collision setting is a popular solution to model the accident evolution rules [42], [43], [44]. Recently, accident video diffusion develops fast for freely generating visual and expected driving accident scenes conditioned by accident-related text descriptions, such as the object-centric diffusion models [9], [45], while the scene structure and style distortion remains a tricky problem. Therefore, this work presents an abductive framework to make a joint optimization of accident video diffusion and accident cause reasoning, with the aid of spatial-temporal scene graph representation. Moreover, generating realistic accident videos requires not only visual plausibility but also sensitivity to causal entities with controllable semantics. The works [27], [28] enhance traffic accident video synthesis by integrating causal reasoning, semantic control, and cognitively informed conditioning, enabling both counterfactual generation and fine-grained alignment with causally critical entities.

Unified Accident Video Understanding: A growing body of works [46], [47], [48], [49] have demonstrated the advantages of unified frameworks that couple with video generation and understanding in general-purpose scenarios. However, these works rarely address safety-critical scenarios like traffic accidents. AVD2 [29] is an underway attempt to explore a framework for accident video generation and text reasoning that highlights the potential of combining generation and understanding to enhance accident video understanding. Nevertheless, AVD2 adopts a parallel formulation that lacks a unified spatial-temporal reasoning and scene-consistent alignment with the synthetic accident videos. Differently, our ADVersa framework is the first attempt at unified abductive accident understanding.

B. Ego-view Driving Accident Understanding Datasets

Tab. I presents the comparison of available ego-view accident video datasets. DAD [50] is the pioneering dataset, where each video clip is trimmed with ten accident frames at the end of each clip. This setting is also adopted in the CCD datasets [19] with a total of 50 frames for each clip. A3D [10] and DoTA [11] are used for unsupervised ego-view accident detection [10], [11], [51]. Specifically, the DADA-2000 dataset [13] annotates the extra driver attention data. Because of the difficulty in collecting enough accident videos in the real world, some works leverage the simulation tool to synthesize the virtual accident videos or object tracklets, such as GTACrash [52], VIENA^2[53], DeepAccident [54], and CTAD [15]. 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. CTA [55] and SUTD-TrafficQA [30] explore the role of text information in driving accident video understanding. RoadSocial [56] further introduces a large-scale VideoQA dataset that captures diverse road events from social media narratives. However, these works still focus on the video captioning or VideoQA tasks.

More recently, several new datasets have emerged with more specialized modalities and targeted objectives. COOOL [58] addresses the Out-of-Distribution (OOD) challenge by introducing a benchmark tailored for open set recognition, open vocabulary, and domain adaptation in hazard detection.

TABLE I
ATTRIBUTE COMPARISON OF DRIVING ACCIDENT VIDEO DATASETS.

Datasets	Years	#Clips	#Frames	Bboxes	D-S	Tracklet	TA.	TT	R/S
DAD [50]	2016	1,750	175K			✓	✓		R
A3D [10]	2019	3,757	208K				✓		R
GTACrash [52]	2019	7,720	-				✓		S
VIENA^2 [53]	2019	**15,000**	**2.25M**				✓		S
CTA [55]	2020	1,935	-				✓	✓	R
CCD [19]	2021	1,381	75K			✓	✓		R
TRA [57]	2022	560	-				✓		R
DADA-2000 [13]	2022	2000	658k				✓		R
DoTA [11]	2022	5,586	732K	partial			✓		R
SUTD-TrafficQA [30]	2021	9480	142K					✓	R
ROL [12]	2023	1000	100K				✓		R
DeepAccident [54]	2023	-	57k				✓		S
CTAD [15]	2023	1,100	-				✓		S
COOOL [58]	2024	200	-	✓					R
CycleCrash [59]	2025	3,000	436K				✓	✓	R
RoadSocial [56]	2025	13,200	14M					✓	R
AV-TAU[18]	2025	29,865	3.16M				✓	✓	R
MM-AU-1 [9]	2024	**11,727**	**2.19M**	✓			✓	✓	R
MM-AU-2	2025	**11,727**	**2.19M**	✓	✓	✓	✓	✓	R

Bboxes: object bounding boxes, **TA.**: temporal annotation of the accident, **TT**: text descriptions, **R/S**: real or synthetic datasets, **D-S**: depth and segment maps.

CycleCrash [59] focuses on cycling-involved scenarios with fine-grained annotations, supporting video understanding tasks such as crash annotation and accident prediction. AV-TAU [18] introduces the audio modality for the first time in the understanding of traffic accidents captured in surveillance and driving scenes by pairing traffic videos with synchronized sound cues, allowing vision-acoustic reasoning. In this work, our MM-AU dataset supports more tasks from single-modal perception to multimodal understanding, and this work explores the abductive understanding framework that covers four groups of visual abductive understanding tasks.

C. Visual Abductive Reasoning

Visual Abductive Reasoning (VAR) infers the most plausible visual explanation of an incomplete video observation or a visual phenomenon by visual and textual modalities in our lives [60]. This is an instinct of humans, but challenging in contemporary AI systems because of the fine-grained and deep understanding of domain context and common sense. The Region Conditioned Adaptation (RCA) solution [61] is proposed for VAR for image-level understanding, which models a hybrid parameter-efficient fine-tuning approach that equips the frozen CLIP with the ability to infer hidden explanations from regional visual hints. The video-based VAR extracts the intricate and detailed conceptual knowledge embedded within the observed video, and the underlying text description for incomplete observation is generated given the visual premises of causal-linked intervals [62]. Liang *et al.* [4] establishes the concept of video-based VAR and models a causal-and-cascaded reasoning Transformer to capture the causal structure of visual observations. MECD [63] further extends VAR to uncover causal relationships between multiple events distributed chronologically over long video sequences, moving beyond single-event reasoning paradigms, and CARVE [64] formulates it as identifying the root-cause trigger event by a causal event-level graph for modeling inter-event dependencies.

More fine-grained, reasoning the multimodal action chains abductively is underscored by the works [65], [66], [67].

TABLE II
STATIC ATTRIBUTES OF DRIVING ACCIDENT VIDEO DATASETS.

Datasets	weather condition				occasion situations				
	sunny	rainy	snowy	foggy	highway	urban	rural	mountain	tunnel
CCD [19]	1,306	61	14	0	148	725	502	5	1
A3D [10]	2,990	251	474	42	225	2,458	720	328	26
DADA-2000 [13]	1,860	130	10	-	**1,420**	380	180	-	20
DoTA [11]	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**

Amongst, the work [65] grounds the most plausible event spatio-temporally and infers the hypothesis of the subsequent action chain that can best explain the language premise. Unlike previous works that rely on supervised learning over fixed-length observations, VIDSE [68] introduces a training-free framework that first deduces high-level goals through LLM and then selects supporting visual evidence from long videos, thus decoupling reasoning from perception and improving interpretability in open-vocabulary settings. In general video or image understanding, the reasoned plausible event or explanations easily involve ambiguity. In contrast, the driving accident video domain implies more specific scene knowledge for plausible accident causes or absent visual observations.

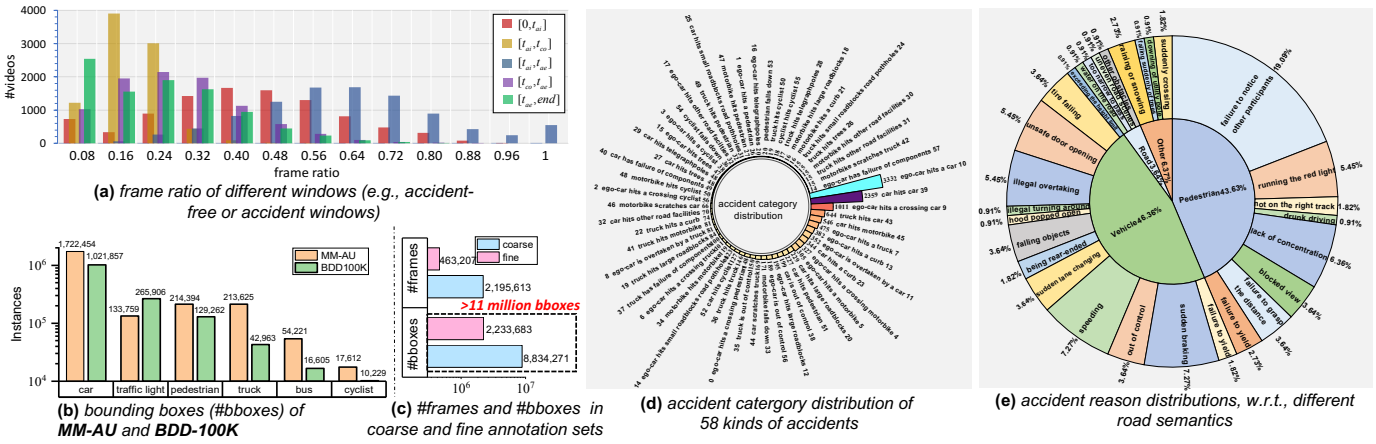
III. MM-AU DATASET

The videos in MM-AU are collected from the publicly available ego-view accident datasets, such as CCD [19], A3D [10], DoTA [11], and DADA-2000 [13], and various video stream sites[1]. As presented in Tab. II, 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 with text descriptions, accident windows, and accident time stamps. To our best knowledge, MM-AU is the largest and the most fine-grained ego-view multi-modal accident dataset to date, covering the tasks from object perception to scene understanding. The annotation process is depicted as follows.

Accident Window Annotation: Leveraging the annotation criteria of DoTA [11], 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. 2(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 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 [69] detector (pre-trained on the COCO dataset [70]), to initially detect the objects in the raw MM-AU videos. Secondly, MM-AU has a fine

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



(a) *frame ratio of different windows (e.g., accident-free or accident windows)*

(b) *bounding boxes (#bboxes) of* ***MM-AU*** *and* ***BDD-100K***

(c) *#frames and #bboxes in coarse and fine annotation sets*

(d) *accident catergory distribution of 58 kinds of accidents*

(e) *accident reason distributions, w.r.t., different road semantics*

Fig. 2. The annotation attribute statistics in MM-AU for the temporal, object, and text annotations. (Zoom in for better view.)

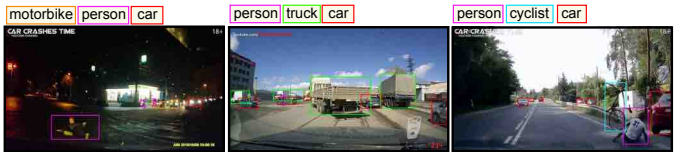


Fig. 3. Examples of object annotations in MM-AU.

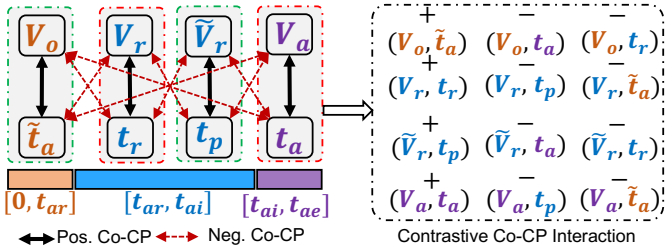


Fig. 4. **Abductive CLIP** contains four interaction groups, with one positive Co-CP and two negative Co-CPs for each interaction group. t_{ar}=t_{ai}-40.

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 were obtained. Each bounding box is double-checked for the final confirmation, and some samples are shown in Fig. 3. As for the coarse annotation set, we utilize a state-of-the-art (SOTA) DiffusionDet [71] to obtain the object bboxes for the remainder of frames in MM-AU. Fig. 2(b) presents the #bboxes on different road participants compared to BDD-100K [72], and Fig. 2(c) shows the #frames and #bboxes on the fine and coarse sets.

Depth and Segment Maps: This work annotates the semantic instance and depth maps using the excellent SegmentAnything Model (SAM) [73] and DepthAnything Model (DAM) [74]. We take the High-quality Tracking [75] approach to track each object instance with a consistent identity across frames, where each object identity is initialized at the video frame with accurate object annotations, where **83,965** tracklets for every 5 frames are obtained.

Text Description Annotation: Different from previous ego-view accident video datasets, MM-AU annotates three kinds of text descriptions: accident cause, 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 cause 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 [76], 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. 2(d).

We annotate 110 pairs of sentences for accident cause and prevention advice descriptions correlating to four kinds of road semantics, *i.e.*, pedestrian-centric, vehicle-centric, road-centric, and others (environmental issue). Fig. 2(e) shows the accident cause distribution concerning different road semantics. It is clear that the "*failure to notice other participants*" is dominant for pedestrian-centric accident causes, and "*speeding*", "*sudden braking*", and "*illegal overtaking*" are the main kinds of vehicle-centric accident causes. Following the form of the video question answering (VideoQA) task [33], MMAU can support Accident cause Answering (AcA) task while there is only one question "`What is the cause for the accident in this video?`". For each accident cause of a video, we further provide four reasonable distractors [2] to form a multi-choice AcA task, and the distractor reasons are all unrelated to the target accident video. We obtain **58,650** AcA pairs in MM-AU.

IV. ADVERSA

This section presents our ADVersa with the text-video coherent learning for ego-view accident video understanding. As aforementioned, we partition each accident video into three segments, *i.e.*, the normal segment V_o, the near-crash segment V_r, and the collision segment V_a. Correspondingly,

[2](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 cause description in the cause set.

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

Based on this denotation, ADVersa aims to infer the absent but most plausible visual and text Co-CPs given an observed Co-CP, *i.e.*, abductively recovering, predicting the absent near-crash scenes paired with the accident cause or category descriptions. However, this formulation is challenging because of the easily overlooked subtle, local, but critical objects that dominate the accident video understanding, reflected by the multi-modal semantic alignment.

A. Relation-wise Visual Semantic Alignment of Accident Scenes

1) Abductive CLIP: For abductive video understanding, we propose an Abductive CLIP for AdVersa to fulfill the coherent semantic learning within different Co-CPs. Abductive CLIP aims 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 antonyms of verbs to the accident category description, such as "does/do not", "are not", etc. In addition, because there is no realistic clip with the accident dissipation process, $\tilde{V}_r$ is obtained by reverse frame rearrangement of V_r from the near-crash state to the normal state.

Certainly, because each video segment may have a different number of video frames, we create Co-CPs by coupling the text description with the randomly selected 16 successive frames within a 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 [77] as the backbone. Subsequently, we input each Co-CP into XCLIP to obtain the feature embedding of the video clip feature $\boldsymbol{z}_v$ and the text feature $\boldsymbol{z}_t$. 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:

$$\mathcal{L}_{\text{CILoss}} = -\sum_{s=1}^{B} \log \frac{E(\boldsymbol{z}^p_{v_s}, \boldsymbol{z}^p_{t_s})}{\mathcal{K}},$$
$$\mathcal{K} = \sum_{b=1}^{B} [E(\boldsymbol{z}^p_{v_s}, \boldsymbol{z}^p_{t_b}) + E(\boldsymbol{z}^p_{v_s}, \boldsymbol{z}^{n_1}_{t_{b\neq s}}) + E(\boldsymbol{z}^p_{v_s}, \boldsymbol{z}^{n_2}_{t_{b\neq s}})], \tag{1}$$

where $E(\boldsymbol{z}_v, \boldsymbol{z}_t) = e^{\boldsymbol{z}_v{}^\top \boldsymbol{z}_t/\tau}$ computes the coherence degree of video clip feature $\boldsymbol{z}_v$ and text feature $\boldsymbol{z}_t$. 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, s and b are the sample indexes in each batchsize.

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

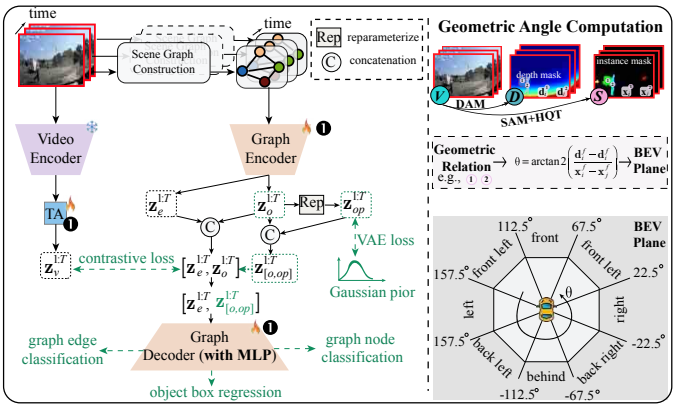


Fig. 5. The pipeline of **CGVP** and the binning scheme for object depth relations. The frame-level scene graphs $\mathcal{G}^{1:T}$ are encoded by a Graph Encoder to output edge features $\boldsymbol{z}_e^{1:T}$ and object node features $\boldsymbol{z}_o^{1:T}$ over T frames. Then $\boldsymbol{z}_o^{1:T}$ is regularized to match a Gaussian prior by the VAE loss, facilitating compatibility with the diffusion latent Gaussian sampling, and further fused with $\boldsymbol{z}_e^{1:T}$ for scene graph reconstruction by a Graph Decoder. Moreover, the video-level representation $\boldsymbol{z}_v^{1:T}$ and the global graph representation $[\boldsymbol{z}_e^{1:T}, \boldsymbol{z}_{o,op}^{1:T}]$ are aligned by a contrastive loss to ensure visual and scene-graph alignment.

2) Contrastive Graph Video Pre-Training (CGVP): Besides the text-vision alignment, this work also explores the alignment between the scene relation and videos, and the Contrastive Graph Video Pre-Training (CGVP) relies on scene graph construction and visual-graph alignment.

This work constructs the scene-graph $\boldsymbol{G}^f = (\boldsymbol{O}^f, \boldsymbol{E}^f)$ with N_o nodes and N_r edges for the f^{th} frame in a video clip using a semi-automatic strategy. The node $\boldsymbol{O}^f$ contains the object semantic labels and the pixel coordinate ($\boldsymbol{x}$) of all objects in the f^{th} frame, such as the pedestrians, cars, etc. The edge $\boldsymbol{E}^f$ infers the geometric depth relation between different objects, where the geometric depth relation between two objects is computed by projecting the relative spatial position's angle to birds'-eye-view (BEV). As shown in Fig. 5, the geometric angle between two objects are computed by $\arctan2(\frac{\boldsymbol{d}_j^f - \boldsymbol{d}_i^f}{\boldsymbol{x}_j^f - \boldsymbol{x}_i^f})$, where $(\boldsymbol{x}_i^f, \boldsymbol{d}_i^f)$ and $(\boldsymbol{x}_j^f, \boldsymbol{d}_j^f)$ denote the average lateral coordinates and depth values of object nodes o_i^f and o_j^f, respectively. We discretize the geometric angle into 8 angular bins.

Node and Edge Representation: For simplicity, we omit frame index f, batch size, and denote the i^{th} object node in the graph as $o_i \in \boldsymbol{O}$, and relations as $e_{ij} \in \boldsymbol{E}$. Then the relation triplets $\Delta_i = (o_i, e_{ij}, o_j)$ are constructed for all (i, j) pairs in $\boldsymbol{G}$. In Δ_i, each object node o_i is embedded by:

$$\boldsymbol{z}_{o_i} = \boldsymbol{z}_{\text{clip}}(c_i) \oplus \boldsymbol{z}_{\text{learn}}(c_i) \oplus \boldsymbol{z}_{\text{box}}(\boldsymbol{b}_i), \tag{2}$$

where c_i is the object class label, $\boldsymbol{z}_{\text{clip}}(c_i) \in \mathbb{R}^{N_o \times 1024}$ is the pre-trained embedding of object semantics by Abductive CLIP. $\boldsymbol{z}_{\text{learn}}(c_i) \in \mathbb{R}^{N_o \times 64}$ denotes a learnable object class embedding, $\boldsymbol{z}_{\text{box}}(\boldsymbol{b}_i) \in \mathbb{R}^{N_o \times 64}$ is the coordinate embedding of o_i, $\boldsymbol{z}_{o_i} \in \mathbb{R}^{N_o \times 1152}$, and $\oplus$ denotes feature concatenation.

The edge embedding $\boldsymbol{z}_{e_{ij}} \in \mathbb{R}^{N_r \times 1152}$ is modeled by the geometric bin index bin_{ij} (ranging from 1 to 8) and relative location semantics rs_{ij}, *e.g.*, *"in front of"*, *"behind"*, etc., for two linked objects, denoted as:

$$\boldsymbol{z}_{e_{ij}} = \boldsymbol{z}_{\text{clip}}(rs_{ij}) \oplus \boldsymbol{z}_{\text{learn}}(bin_{ij}), \tag{3}$$

where $\boldsymbol{z}_{\text{clip}}(rs_{ij}) \in \mathbb{R}^{N_r \times 1024}$ is also embedded by Abductive-CLIP, and $\boldsymbol{z}_{\text{learn}}(bin_{ij}) \in \mathbb{R}^{N_r \times 128}$ is a learnable relation class embedding. To maintain a consistent scene graph relation across different frames, we use HQ-Track [75] to preserve object temporal identity and aggregate the scene graphs $\{\boldsymbol{G}^f\}_{f=1}^{F}$ across frames into a linked global scene graph $\mathcal{G} = (\mathcal{O}, \mathcal{E})$. The global scene graph is then encoded by L layers of encoder-decoder structured triplet-GCN [78] to optimize the node and edge representations $\boldsymbol{z}_{\text{o}}^{(L)} \in \mathbb{R}^{|\mathcal{O}| \times d_o}$ and $\boldsymbol{z}_{e}^{(L)} \in \mathbb{R}^{|\mathcal{E}| \times d_e}$ in the end layer (the L^{th} layer), respectively. We set $L = 5$, $d_o = 64$, and $d_e = 128$ in this work.

Scene Graph Representation Learning: To faithfully represent a scene graph, we define the scene graph reconstruction loss $\mathcal{L}_{\text{SG}}$ for graph node and edge classification and a regression loss $\mathcal{L}_{\text{reg}}$ for bounding box prediction. These joint losses ensure both semantic and structural fidelity in the reconstructed graph.

In addition, to reconcile the following accident video diffusion and inspired by the Gaussian noise assumption in video diffusion models [79], we learn a Gaussian distribution in latent space over possible scene graph structures to handle the graph perturbations in driving scenes. Therefore, the graph node representation $\boldsymbol{z}_{\text{o}}^{(L)}$ is used to predict the scene graph latent distribution $\mathcal{N}(\boldsymbol{\mu}, \log \boldsymbol{\sigma}^2)$ by two separate Multiple Layers of Perceptrons (MLPs), where $\boldsymbol{\mu}$ is the mean and $\log \boldsymbol{\sigma}^2$ denotes the logarithmic variance, capturing the central tendency and uncertainty of the scene graph representation, respectively. In the decoding stage of the scene graph, we obtain the re-parameterized graph node representation $\boldsymbol{z}_{sg} \in \mathbb{R}^{N_o \times 64}$ by:

$$\boldsymbol{z}_{sg} = \boldsymbol{\mu} + \boldsymbol{\sigma} \odot \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I}), \tag{4}$$

where $\boldsymbol{\sigma} = \exp(0.5 \log \boldsymbol{\sigma}^2)$ is the standard deviation, $\boldsymbol{\epsilon}$ is the random noise, and $\odot$ is element-wise multiplication. $\boldsymbol{z}_{sg}$ is then integrated into the scene structure, yielding a new triplet set $\Delta_i' = (\hat{\boldsymbol{z}}_{o_i}, \boldsymbol{z}_{e_{ij}}, \hat{\boldsymbol{z}}_{o_j}), (i,j) \in \mathcal{E}$. Each object embedding $\hat{\boldsymbol{z}}_{o_{i/j}}$ is re-computed as $\hat{\boldsymbol{z}}_{o_{i/j}} = \boldsymbol{z}_{\text{clip}}(c_{i/j}) \oplus \boldsymbol{z}_{\text{learn}}(c_{i/j}) \oplus \boldsymbol{z}_{sg}^{i/j}$, incorporating both object semantics and latent variables. The decoder in triplet GCN takes Δ_i' as input and reconstructs the scene graph using the Kullback-Leibler divergence $\mathcal{L}_{\text{KL}}$ to regularize the latent distribution $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\sigma}^2)$ toward the Gaussian prior by $\text{KL}(\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\sigma}^2) || (\mathcal{N}(0, \mathbf{I}))$, adopting the cross-entropy $\mathcal{L}_{\text{SG}}$ and the regression loss $\mathcal{L}_{\text{reg}}$ to classify the nodes, edges of the scene graph and regress the bounding boxes.

Visual-Graph Alignment: We formulate the visual-graph alignment in a segment-level contrastive strategy. Specifically, following Co-CPs illustrated in Fig. 4, we construct four types of visual-graph Co-Occurrence Pairs (VG-Co-CPs), *e.g.*, [$\mathcal{G}_o$, V_o], [$\mathcal{G}_r$, V_r], [$\mathcal{G}_a$, V_a], and [$\tilde{\mathcal{G}}_r$, $\tilde{V}_r$], where $\tilde{\mathcal{G}}_r$ is obtained by reverse frame rearrangement of $\mathcal{G}_r$.

This design allows us to learn visual-graph alignment under both normal and accident contexts. Therefore, we extract the video feature $\boldsymbol{z}_v^{\text{temp}} \in \mathbb{R}^{d_v}$ for each VG-Co-CPs from the video encoder [80], followed by a Temporal Attention (**TA**) mechanism with an average pooling and a linear layer over $\boldsymbol{z}_o^{(L)}$ and $\boldsymbol{z}_e^{(L)}$ to obtain the global embedding $\boldsymbol{z}_{\text{gb}} \in \mathbb{R}^{d_v}$ that captures the holistic scene structure, where d_v is set to 640. Each VG-Co-CP is then represented as a paired embedding $(\boldsymbol{z}_v^{\text{temp}}, \boldsymbol{z}_{\text{gb}})$. To align video and scene graph representations, we use an InfoNCE contrastive loss:

$$\mathcal{L}_{\text{con}} = -\sum_{s=1}^{B} \log \frac{E(\boldsymbol{z}_{v_{(s)}}^{\text{temp}}, \boldsymbol{z}_{\text{gb}_{(s)}})}{\sum_{b=1}^{B} E(\boldsymbol{z}_{v_{(s)}}^{\text{temp}}, \boldsymbol{z}_{\text{gb}_{(b \neq s)}})}, \tag{5}$$

where $E(\boldsymbol{z}_v^{temp}, \boldsymbol{z}_{\text{gb}}) = \exp\left(\boldsymbol{z}_v^{temp^\top} \boldsymbol{z}_{\text{gb}} / \tau\right)$ denotes the exponential similarity between video and graph embeddings, and τ is a learnable temperature parameter set to 0.07. For each video feature $\boldsymbol{z}_{v_{(s)}}^{\text{temp}}$, its paired graph embedding $\boldsymbol{z}_{\text{gb}_{(s)}}$ is treated as the positive sample, while all others in the batch act as negatives. The loss for scene CGVP is formulated as:

$$\mathcal{L}_{\text{total}} = \lambda_{\text{KL}} \mathcal{L}_{\text{KL}} + \lambda_{\text{SG}} \mathcal{L}_{\text{SG}} + \lambda_{\text{con}} \mathcal{L}_{\text{con}}, \tag{6}$$

where $\lambda_{\text{KL}}, \lambda_{\text{SG}}, \lambda_{\text{con}}$ are balancing hyperparameters, which are set to 0.1, 1, and 1, respectively.

After achieving the pre-training of text, vision, and scene graph alignment, *i.e.*, with the pre-trained Abductive CLIP and CGVP, we take them as the engines to drive the spatially and temporally abductive accident video diffusion.

B. Spatially Abductive Accident Video Diffusion

To verify Abductive CLIP, this work treats it as an engine to drive a spatially abductive accident video diffusion for finding the dominant but masked objects in accident occurrence, because of the common accident cause of the irregular or sudden movement of road participants. As shown in Fig. 6, we propose an Object-centric Accident Video Diffusion model (ADVersa-OAVD) which takes the Latent Diffusion Model (LDM) [79] and extends it to video diffusion with the input of Co-CPs and K forward and reverse diffusion steps.

The structure of ADVersa-OAVD becomes similar to Tune-A-Video work [81], while differently, the 3D-CAB block of the 3D U-net module in Fig. 6 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**) [84] 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 the frame-level reconstruction. The **inference** phase has the same input form as the OAVD training and generates the new frame clip V_g.

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

$$\begin{aligned} &\mathbb{E}_{V, \boldsymbol{e} \sim \mathcal{N}(0, \boldsymbol{I}), k, \boldsymbol{z}_t, \boldsymbol{z}_b, \overline{\overline{V}}} \Big\{ \|\boldsymbol{e} - \phi_{\theta_v}(\boldsymbol{z}_v, k, \boldsymbol{z}_t, \boldsymbol{z}_b)\|_2^2 + \\ &\lambda \|\boldsymbol{e}(\mathbf{1} - \boldsymbol{z}_{mask}) - \phi_{\theta_v}(\boldsymbol{z}_v, k, \boldsymbol{z}_t, \boldsymbol{z}_b)(\mathbf{1} - \boldsymbol{z}_{mask})\|_1^1 \Big\}, \end{aligned} \tag{7}$$

where the first term is Mean Squared Error ($\mathcal{L}_{\text{MSE}}$) and the second term denotes the reconstruction loss of the Masked

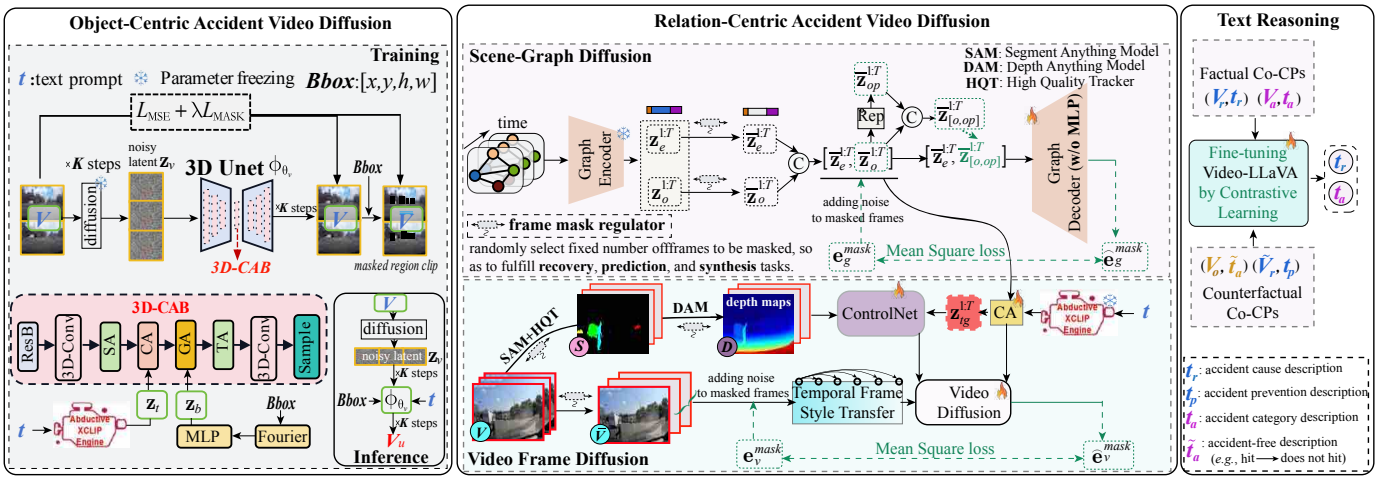


Fig. 6. The architectures of ADVersa-OAVD, ADVersa-RAVD, and Text Reasoning. In ADVersa-OAVD, V and t are the video frame clip and its paired text prompt. Object bounding boxes $Bbox$ in each video frame can be obtained by object detectors. $\overline{\overline{V}}$ is the masked frame clip where the pixels in object regions are set to 0. Different attention modules, *i.e.*, SA, CA, and TA, follow the ones of [81] while with a dense multi-head attention. In ADVersa-RAVD, to facilitate a relation-wise accident video diffusion, we jointly optimize the scene-graph diffusion and video frame diffusion, where the scene-graph diffusion branch operates in the latent space by adding noise $\boldsymbol{e}_g^{mask}$ directly into object node features $\boldsymbol{z}_o^{1:T}$ with T frames. The video diffusion branch receives noise $\boldsymbol{e}_v^{\text{mask}}$ that is initialized by a temporal frame style transfer operation (to be described in §IV-C). To enhance relation-wise visual and text alignment, the global graph representation is fused with the text via a Cross-Attention (CA) module to output enhanced scene-graph textual embeddings $\boldsymbol{z}_{tg}^{1:T}$. $\boldsymbol{z}_{tg}^{1:T}$ is then used as a conditional input to the video diffusion model through ControlNet [82], along with depth map and semantic instance mask conditions. **Text Reasoning**: We perform text reasoning by fine-tuning a Video-LLaVA [83] and ADVersa-RAVD with contrastive factual↔counterfactual Co-CPs learning ((V_r, t_r)↔($\tilde{V}_r$, t_p), and (V_a, t_a)↔($\tilde{V}_r$ t_p)).

Latent Representation ($\mathcal{L}_{\text{MASK}}$). $k \in [1, \ldots, K]$ denotes the diffusion step (K=1000), $\boldsymbol{e}$ is the ground-truth noise representation in each diffusion step. ϕ_{θ_v} 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 as $\boldsymbol{e}$, and the masked noise representation $\mathbf{z}_{mask}$ is obtained by Denoising Diffusion Probabilistic Model (DDPM) Scheduler [85] on the binarization of a latent representation $\boldsymbol{z}_l$ ($\boldsymbol{z}_l = \text{VAE}(\boldsymbol{z}_v)$) through the Variational Autoencoder (VAE) in LDM [79] as:

$$\begin{aligned} \boldsymbol{z}_{mask} &= \text{DDPM Scheduler}(\boldsymbol{m}_{(z)}, k, \boldsymbol{e})), \\ \boldsymbol{m}_{(z)} &= \begin{cases} 0 & \text{if } \boldsymbol{z}_l < 0.5, \\ 1 & \text{if } \boldsymbol{z}_l \geq 0.5. \end{cases} \end{aligned} \tag{8}$$

Gated Bbox Representation: The key insight is to involve object bounding boxes to enhance the important 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) [84], the location embedding $\boldsymbol{z}_b$, collaborating with the output of CA in 3D-CAB, is obtained from Bbox by MLP layers with the Fourier embedding [86]:

$$\boldsymbol{z}_b = \text{MLP}(\text{Fourier}(Bbox)). \tag{9}$$

C. Temporally Abductive Accident Video Diffusion

In addition to the spatially abductive accident video diffusion, we also take the pre-trained Abductive-CLIP and CGVP as engines to drive temporally abductive accident video diffusion. To achieve this goal, we propose a Relation-Centric Accident Video Diffusion model (ADVersa-RAVD). ADVersa-RAVD is also based on the 3D-CAB backbone and involves a causal-consistent visual prior transfer module in the scene-graph reconciled accident video diffusion, as shown in Fig. 6.

Assume we observe an incomplete accident video clip V_c with P frames optionally paired with a text description t_c sampled from the aforementioned Co-CPs, our objective is to infer the absent V_u with Q frames by accident video diffusion model ϕ_{θ_v}, where V_c and V_u belong to the originally complete accident video clip V with $T = P + Q$ frames. With scene graph $\mathcal{G}_c$, ϕ_{θ_v} also modeled by **3D-CAB** (without GA) is optimized by:

$$\mathbb{E}_{V_u, \boldsymbol{e}_v^{mask} \sim \mathcal{N}(0,\boldsymbol{I}), k, \boldsymbol{z}_{tg}, \boldsymbol{z}_{s_c}, \boldsymbol{z}_{d_c}} \| \boldsymbol{e}_v^{mask} - \phi_{\theta_v}(\boldsymbol{z}_u, k, \boldsymbol{z}_{tg}, \boldsymbol{z}_{s_c}, \boldsymbol{z}_{d_c}) \|_2^2, \tag{10}$$

$\boldsymbol{e}_v^{mask} \sim \mathcal{N}(0, \boldsymbol{I})$ is Gaussian noise added in masked frames, $\boldsymbol{z}_{tg} \in \mathbb{R}^{b \times l \times 1024}$ is the conditions of the scene-graph enhanced textual embedding, where b is the batch size and l the prompt length (default 77). $\boldsymbol{z}_{s_c} \in \mathbb{R}^{(b \times P) \times c \times h \times w}$ and $\boldsymbol{z}_{d_c} \in \mathbb{R}^{(b \times P) \times c \times h \times w}$ are the feature embeddings of segment maps $\boldsymbol{S}_c$ and depth maps, where $c = 320$ is the channel dimension, and $h = w = 28$ are the height and width, respectively. $\boldsymbol{D}_c$ and $\boldsymbol{S}_c$ are regulated by a frame mask regulator $s(t_m, M)$, where M ($M < (P + Q)$) is the number of masked frames and t_m indexes the first masked frame. We use ControlNet [82] to integrate these conditions. $\boldsymbol{z}_u$ to be described in the following is obtained by transferring a causal-consistent visual prior from V_c. In particular, $\boldsymbol{z}_{tg}$ is the key that represents the scene relation for ADVersa-RAVD.

Frame Mask Regulator: As the task definition, different t_m associated with past recovery (t_m=1), or future prediction (t_m=$P + Q - M$) of the near-crash driving scene, respectively.

1) Causal-Consistent Visual Prior Transfer: Since the scene diffusion is modeled to fulfill the recovering and predicting tasks reflecting the true scene evolution, the causal-consistent

visual generation is preferred. To achieve this goal, we introduce a style-conditioned prior transfer operation that transfers distributional cues from observed segments to the noisy latent embeddings of the masked frames. This operation ameliorates the perceptual discontinuities or flickering artifacts associated with the observed context. We perform this style transfer to modulate the reference of each frame in the diffusion process, defined as:

$$\tilde{\boldsymbol{z}}^{f}_{e_u} = \gamma\, \boldsymbol{z}^{f}_{c} + (1-\eta_f)\bar{\boldsymbol{z}}^{f}_{e_u}, \tag{11}$$

where γ is a small fixed blending factor (empirically set to 0.1), $\boldsymbol{z}^{f}_{c}$ is the task-specific style embedding derived from observed segments V_c, and η_f is a temporal decay weight controlling transfer strength over time.

- *Recovering:* The frames in $\boldsymbol{z}^{f}_{c}$ are selected from V_a, increasing η_f linearly from current to past frames.
- *Prediction:* The frames in $\boldsymbol{z}^{f}_{c}$ are sampled in V_r, decreasing η_f linearly from current to future frames.

This mechanism ensures that the denoising process of each target frame is guided by temporal and causally consistent aligned visual priors, improving visual generation controllability. Then the style-conditioned noisy latent $\tilde{\boldsymbol{z}}_{e_u}$ at the k^{th} step is denoised by video diffusion model ϕ_{θ_v}.

2) Reconciling Scene-Graph to Accident Video Diffusion: With the pre-trained CGVP, we also model a scene graph diffusion pipeline (Fig. 6) and attentively inject the latent representation of the scene graph into the video latent representation for achieving a relation-wise accident video diffusion.

In the *forward scene-graph diffusion process*, all object nodes in the masked frames are also regulated by the frame mask regulator. We add mask Gaussian noise $\boldsymbol{e}^{mask}_{g}$ to the scene graph node $\boldsymbol{z}_o$ at k^{th} step and obtain a noisy latent node representation $\bar{\boldsymbol{z}}^{k}_{o}$:

$$\bar{\boldsymbol{z}}^{k}_{o} = \sqrt{\hat{\beta}_k}\boldsymbol{z}^{0}_{o} + \sqrt{1-\hat{\beta}_k}\boldsymbol{e}^{mask}_{g}, \hat{\beta}_k = \prod_{i=1}^{k}\alpha_i, \alpha_k = 1-\beta_k, \tag{12}$$

β_k is a parameterized schedule, and $\boldsymbol{z}^{0}_{o}$ is the pure node embedding. Then the *reverse process* recovers clean $\boldsymbol{z}_o$ by the following sampler step-by-step:

$$\bar{\boldsymbol{z}}^{k-1}_{o} = \mu_{\hat{\boldsymbol{e}}^{mask}_{g}}(\bar{\boldsymbol{z}}_o, k) + \sigma_k \boldsymbol{e}^{mask}_{g}, \tag{13}$$

where $\mu_{\hat{\boldsymbol{e}}^{mask}_{g}}(\bar{\boldsymbol{z}}^{k}_{o}, k) = \frac{1}{\sqrt{\hat{\beta}_k}}\left(\bar{\boldsymbol{z}}^{k}_{o} - \frac{\beta_k}{\sqrt{1-\hat{\beta}_k}}\phi_{\theta_g}(k, \bar{\boldsymbol{z}}^{k}_{o}, \boldsymbol{z}_e)\right)$, and σ_k is the noise variance at the k^{th} step. Similarly to Eq. 7, the scene-graph diffusion process is optimized by:

$$\mathbb{E}_{\boldsymbol{z}_o, \boldsymbol{e}^{mask}_{g}\sim\mathcal{N}(0,\boldsymbol{I})}\|\boldsymbol{e}^{mask}_{g} - \phi_{\theta_g}(k, \bar{\boldsymbol{z}}_o, k, \boldsymbol{z}_e)\|^2_2, \tag{14}$$

where ϕ_{θ_g} is the pre-trained triplet-GCN decoder [87] without MLP layers, $\boldsymbol{z}_o$ and $\boldsymbol{z}_e$ are encoded by the pre-trained triplet-GCN [78] in **CGVP**, respectively. Additionally, the noisy latent node and edge representations are projected into a global graph representation $\boldsymbol{z}_{\text{sg}}$, which interacts with the original text embedding $\boldsymbol{z}_{t_c}$ via a Cross-Attention (**CA**) to produce the enhanced scene-graph textual embedding $\boldsymbol{z}_{tg}$.

Finally, the scene-graph diffusion and video diffusion are jointly optimized under the following total loss:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{VD}} + \lambda_{\text{SGD}}\mathcal{L}_{\text{SGD}}, \tag{15}$$

where λ_{SGD} (set 0.3) balances the diffusion of scene-graphs and video frames.

D. Abductive Text Reasoning of Accident Videos

After optimizing ADVersa-RAVD, we further fine-tune it by a counterfactual reasoning stage to enhance the visual abductive ability of ADVersa in accident scenes.

In this stage, the Multi-modal Large Language Models (MLLM) can be used and fine-tuned jointly with RAVD by leveraging both factual and counterfactual pair-wise Co-CPs. In this work, the pair-wise Co-CPs (V_r, t_r)↔($\tilde{V}_r$, t_p), and (V_a, t_a)↔($\tilde{V}_r$, t_p) are naturally selected as factual↔counterfactual Co-CPs in the fine-tuning stage. These factual and counterfactual Co-CPs are contrastively learned to enhance the alignment between generated videos V_u and the text t_u further.

Contrastive Counterfactual Loss: We use **Video-LLaVA-7B** [83] as an attempt, and input factual pair-wise Co-CPs and counterfactual pair-wise Co-CPs to obtain the feature embeddings of the video $\boldsymbol{z}^{u}_{v} \in \mathbb{R}^{b\times 4096}$ and text $\boldsymbol{z}^{u}_{t} \in \mathbb{R}^{b\times 4096}$. Then, a Contrastive Counterfactual Loss (CCLoss) aligns video and text features within the related Co-CPs:

$$\mathcal{L}_{\text{CCLoss}} = -\frac{1}{SN}\sum_{i=1}^{SN}\log\frac{\exp(\text{sim}(\boldsymbol{z}^{u(i)}_{v}, \boldsymbol{z}^{u(i)}_{t})/\tau)}{\sum_{j=1}^{N}\boldsymbol{I}_{[y_j=y_i]}\cdot\exp(\text{sim}(\boldsymbol{z}^{u(i)}_{v}, \boldsymbol{z}^{u(j)}_{t})/\tau)}, \tag{16}$$

where $\text{sim}(\cdot)$ denotes cosine similarity, τ (set to 0.07) is a learnable temperature parameter, and $\boldsymbol{I}_{[y_j=y_i]}$ ensures positive samples are only selected from the same label group. In addition, a Maximum Likelihood Estimation loss $\mathcal{L}_{\text{MLE}}$ [83] over the generated text description t_u is used to enforce semantic coherence in text reasoning. Optimizing $\mathcal{L}_{\text{MLE}} + \mathcal{L}_{\text{CCLoss}}$ improves causal-consistent visual-text semantic learning.

E. Training and Inference Recipe

Training Recipe: Approaching training ADVersa-RAVD undergoes four progressive stages. In stage-❶, we fulfill the pre-training of Abductive-CLIP and CGVP for visual-text and visual-relation semantic alignment. In Stage-❷, we train ADVersa-OAVD driven by the pre-trained Abductive-CLIP, where Abductive-CLIP is also optimized by ADVersa-OAVD (mainly focused in the previous version [9]). In Stage-❸, ADVersa-RAVD is initialized from the ADVersa-OAVD checkpoint (without GA module), which is driven by Abductive-CLIP and CGVP engines, while differently we keep Abductive-CLIP frozen and continue to fine-tune CGVP. Therefore, the triplet-GCN [78] is frozen to retain Gaussian latent priors, while the triplet-GCN decoder $\mathcal{D}$ is fine-tuned in latent space (excluding its MLP head) to adapt to the generative context. Concurrently, we apply LoRA-based fine-tuning to the video diffusion model and jointly train the lightweight ControlNet [82] module for effective structural conditioning. In addition, for Text-reasoning training in ADVersa-RAVD (Stage-❹), $\mathcal{D}$ remains trainable and the video diffusion model is frozen except for the Temporal Attention modules (TA) in ControlNet. QLoRA [88] is applied to the multimodal large language model (MLLM) for refining temporal reasoning and improving the alignment between factual and counterfactual Co-CP samples.

This progressive strategy ensures that each module incrementally refines its representations, culminating in a unified model capable of abductive reasoning, semantic alignment, and counterfactual interpretation across traffic accident scenarios.

‣**Inference Recipe:** The ADVersa-OAVD inference inputs the Co-CPs while the Denoising Diffusion Implicit Model (DDIM) scheduler [85] is taken on the trained 3D U-net conditioned by the text prompt t and $Bbox$ to synthesize a video V_g. Unlike ADVersa-OAVD, beyond the Co-CPs and the text prompt, ADVersa-RAVD is further conditioned on scene graph and structural controls (*e.g.*, semantic maps S and depth maps D) to generate the past or future video V_u, and reason its textual description t_u. The conditioning only uses an incomplete clip input in ADVersa-RAVD. When a full clip is provided, a video V_g can be synthesized.

V. Experiments

A. Tasks and Datasets

For a clear evaluation, to distinguish with the object-centric accident video diffusion ("ADVersa-OAVD") in [9], we adopt "ADVersa-RAVD" to denote the extended scene-graph reconciled accident video diffusion model. ADVersa-RAVD serves three kinds of tasks, *i.e.*, *Rec-CrashPast*, *Pred-Crash*, *Accident-Syn* extensively.

Because of the very short time window of the suddenly appeared accident, we take 16 frames to show the near-crash to crashing (NC-2-C) process concisely. For training ADVersa-OAVD and ADVersa-RAVD, we sample 5,098 clips in 9,727 accident videos of MM-AU with pair-wise accident reason-collision and text descriptions, randomly sampled from the near-crash to crashing frame windows for maintaining diversity 16 frames from [V_r,V_a] in training, and we resize each frame to a size of 224×224.

For ADVersa-OAVD evaluation, we sample 1800 clips in the testing set of the MMAU dataset and input the accident cause and prevention advice description, which helps to check the text-vision alignment of Abductive-CLIP and the critical object learning in the diffusion process.

‣**Rec-CrashPast** recovers the historical near-crash frames and infers the accident cause description under a "6 (crashing frames)-10 (to be recovered near-crash frames)" mode conditioned by (V_a, t_a, $\mathcal{G}_a$).

‣**Pred-Crash** predicts the future crashing frames and reasons the accident category description with the same "6 (near-crash frames)-10 (to be predicted frames)" mode conditioned by (V_r, t_r, $\mathcal{G}_r$). Certainly, we also conduct an ablation analysis on the performance under different numbers of context frames. We sample 1,797 video clips (16 frames of each clip) from the MM-AU dataset as our test set for Rec-CrashPast and Pred-Crash evaluation.

‣**Accident-Syn** sets two subtasks: 1) *normal-to-accident* video diffusion (N2A) and 2) *accident video editing* (AEdit). For the **N2A** task, similar to [27], we adopt the safety-critical dataset BDD-A [89]and sample 2,000 testing clips (16 frames of each clip) as visual prompts. For **AEdit**, 3,000 clips are sampled in the DADA-2000 dataset [13] for editing inference.

B. Metrics and Implementation Setups

This work involves video frame recovery, prediction, synthesis, and textual reasoning tasks. The experimental analysis is scheduled as follows:

‣**Abductive Ability Checking of ADVersa-OAVD**: We adopt Clip-score (CLIPs [9]), Fréchet Video Distance (FVD [90]), and Frames Per Second (FPS) to evaluate the generated frame quality and efficiency of ADVersa-OAVD.

‣**Frame Recovering or Prediction**: Following popular works [91], [92], we evaluate the quality of the generated videos by Structural Similarity (SSIM [93]), Fréchet Video Distance (FVD [90]), Temporal Coherence (TempC [94]), Peak Signal-to-Noise Ratio (PSNR [93]), and Learned Perceptual Image Patch Similarity (LPIPS [95]). Notably, because the generated video frames imply a semantic consistency with accident description, we also employ the Clip-score (CLIPs [9], [96]) to check video-text semantic alignment in generated frames.

‣**Accident Cause and Category Reasoning**: To probe and assess the textual reasoning, we employ an evaluation protocol based on a descriptive Visual Question Answering (VQA) format. This framework uses open-ended questions to prompt the models to generate descriptive and explanatory text based on the video content. The templates are tailored to each task as follows: 1) **Prediction:** `USER: <video> What is the category of this predicted accident video?`" and 2) **Retrospection:** `USER: <video> What is the cause of this retrospective accident video?`". We employ a comprehensive set of standard metrics from the field of natural language processing by AVD2 [29]. These metrics include BLEU [97] (B1 to B4), CIDEr (Cr), SPICE (S) [98], METEOR (M) [99], and ROUGE-L (RL) [100]. All of these metrics prefer higher values for better performance.

‣**Accident Video Synthesis**: Similarly to video recovery, video quality and text-vision semantic alignment are the keys for accident video synthesis. Therefore, CLIPs, FVD, and TempC are also adopted. In particular, because the small region of the reflected visual entities to be edited, such as pedestrians, cyclists, vehicles, etc., we adopt the affordance (**Afd**) metric in [27], which stems from the Add-it image editor [101] and utilizes the gazed regions instead of the object bounding boxes to match a human-preferable visual-entity editing checking. Therefore, the ground-truth gaze maps in DADA-2000 [13] are used in **Afd** in the AEdit task, computing by:

$$\text{Afd} = \frac{1}{N_e}\sum_{e=1}^{N_e}\mathbb{I}(\text{IOU}(B_e^{det}, B_e^{gaze}) > 0), \tag{17}$$

where N_e denotes the number of testing samples, and $\mathbb{I}(\cdot)$ is the indicator function. For each sample e to be edited, B_e^{det} is the bounding box detected by GroundDINO [102] based on the edited frame and a text prompt (reflecting the varied text word, *e.g.*, "a car"), and B_e^{gaze} is the corresponding bounding box derived from the gaze regions (stems from frame-paired gaze maps in DADA-2000 dataset [13]). IOU$(\cdot)$ denotes the standard Intersection over Union function, which measures the overlap between these two boxes. **Notably**, Afd could be used

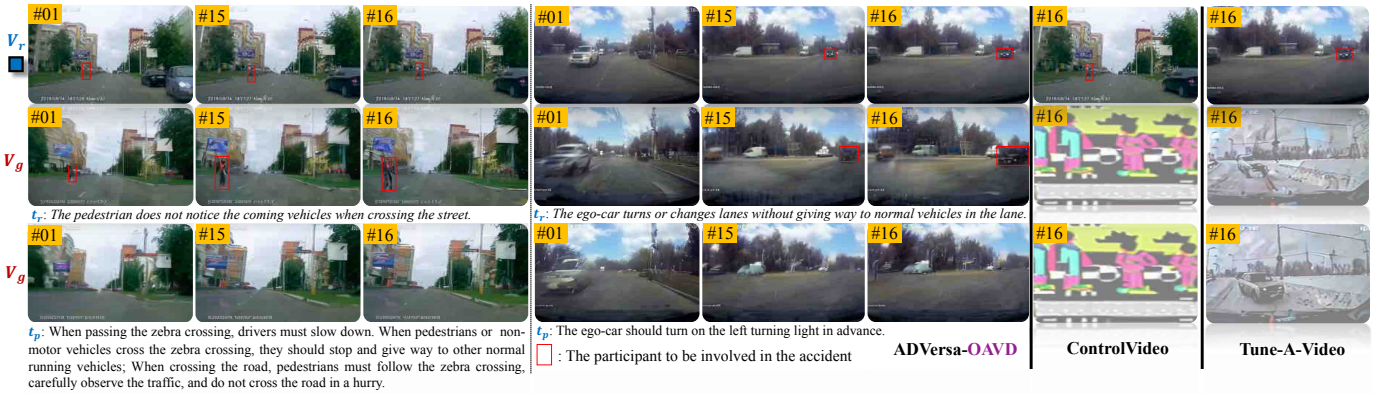


Fig. 7. Some generated results (V_g) prompted by the Co-CPs of (V_r, t_r) or (V_r, t_p). From the generation of ADVersa-OAVD, the participants to be involved in accidents appear in advance when giving the accident reason prompt t_r, while the accident objects disappear when providing the prevention advice prompt t_p. ControlVideo [103] and Tune-A-Video [81] are given the same t_r and t_p prompt with ADVersa-OAVD, respectively. However, the artifacts and unrelated content are generated, and the phenomena of "*appear in advance* " or "*disappear*" do not occur.

in the evaluation when the frame style maintains consistent with original video prompt.

Implementation Details: All experiments are conducted on a platform with 2× NVIDIA RTX A6000 GPUs. In ADVersa-RAVD, the text description and scene graph are modeled by the pre-trained Abductive CLIP and CGVP engines, respectively. The learning rate of Abductive CLIP in ADVersa-OAVD is set to $1e{-}6$ with a batch size of 2, and it is trained for 30,000 iteration steps. The learning rate of the CGVP engine in ADVersa-RAVD is set to $1e{-}5$ with a batch size of 12, and it is trained for 25 epochs. ADVersa-OAVD is trained in the initial learning rate of $5e{-}6$ with the per-device batch size 2, and training runs for 15,000 steps. In the text reasoning task of ADVersa-RAVD, the per-device batch size is 1 and training runs for 10,000 steps. The Adam optimizer is adopted with the default parameters $\beta_1 = 0.9$ and $\beta_2 = 0.999$ for ADVersa-RAVD training.

C. Evaluations on ADVersa-OAVD

Here, we evaluate the ADVersa-OAVD, the Abductive-CLIP and Bboxes. To verify Abductive-CLIP in ADVersa-OAVD, we take two baselines: 1) the original CLIP model [80] and 2) a "Sequential-CLIP (S-CLIP)" that only maintains the positive Co-CPs of the Abductive-CLIP (A-CLIP) (see Fig. 4).

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, ADVersa-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 ADVersa-OAVD catches the dominant object representation for the accident occurrence. Contrarily, ControlVideo and Tune-A-Video generate irrelevant styles with worse performance than ADVersa-OAVD, as listed in Tab. III, which shows that accident knowledge is scarce in this field but rather crucial for the accident video diffusion models. Tab. III shows the results of different diffusion models.

Roles of Different CLIP Models. Tab. III also presents the results of ADVersa-OAVD with varying CLIP models. The

TABLE III
RESULTS WITH TWO POPULAR DIFFUSION MODELS (TAV: TUNE-A-VIDEO [81] AND COVIDEO: CONTROLVIDEO [103]) AND OUR ADVERSA-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	CoVideo	A-OAVD (CLIP)	A-OAVD (S-CLIP)*	A-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; A-OAVD: ADVersa-OAVD.

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

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 semantics within text and videos.

Role of Bboxes From Tab. III 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. Besides, our ADVersa-OAVD can also flexibly generate any accident videos with the input of only object boxes or the accident category descriptions, where more results can be reviewed in the project site. **Notably**, the bounding box information is also involved in the ADVersa-RAVD and serves its scene graph construction.

D. Evaluations on ADVersa-RAVD

In ADVersa-OAVD, the cause prompt is fed to generate videos that accident object "appear in advance" essentially, which encodes an implicit object-level reaction (**Spatially**), and cannot provide a quantitative abductive ability evaluation.

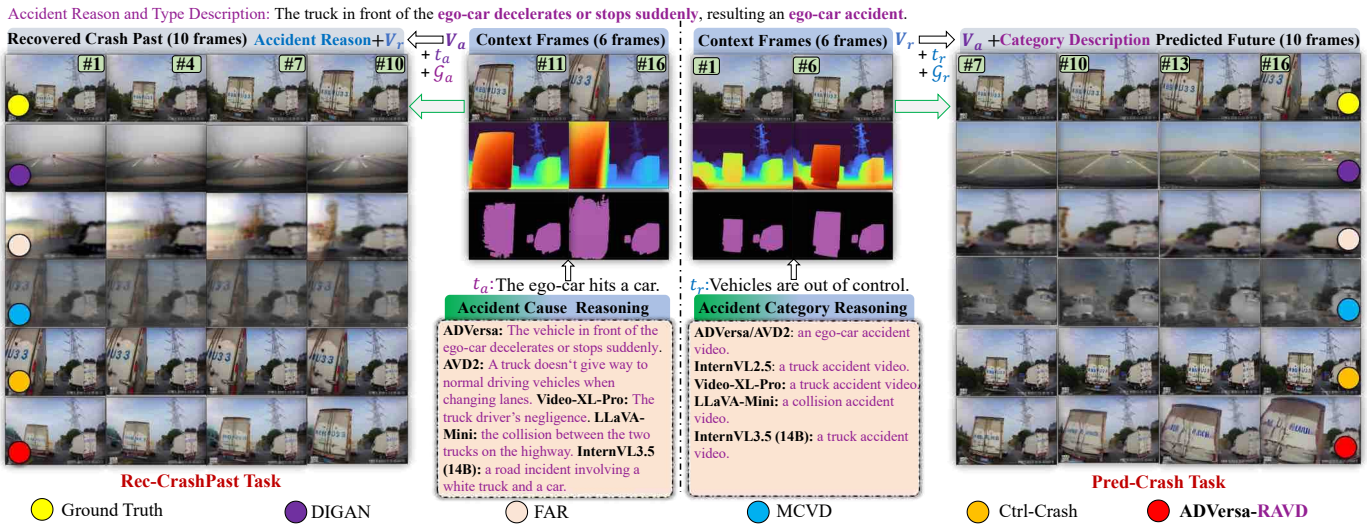


Fig. 9. We visualize the recovering and prediction results of a sample by DIGAN [104], MCVD [91], Ctrl-Crash [28], FAR [92], and our ADVersa-RAVD.

TABLE IV
COMPREHENSIVE PERFORMANCE COMPARISON FOR REC-CRASHPAST (R) AND PRED-CRASH (P) TASKS.

Task	Methods	Designs	CLIPs↑	SSIM↑	FVD↓	PSNR ↑	LPIPS↓
(R)	DIGAN [104]$_{\text{ICLR2022}}$	GAN	24.1	0.501	9830.0	16.422	0.579
	MCVD [91]$_{\text{NeurIPS2022}}$	DF	25.0	0.628	5652.3	19.054	0.322
	RIVER [105]$_{\text{ICCV2023}}$	GAN	21.7	0.178	12591.5	7.764	0.871
	Ctrl-Crash [28]$_{\text{Arxiv2025}}$	DF	27.4	0.626	3635.9	18.739	0.270
	FAR [92]$_{\text{Arxiv2025}}$	AR	24.7	0.428	5866.3	14.066	0.246
	ADVersa-RAVD	DF	**27.8**	**0.637**	**2866.4**	**19.292**	**0.211**
(P)	DIGAN [104]$_{\text{ICLR2022}}$	GAN	24.7	0.500	9744.0	16.373	0.580
	MCVD [91]$_{\text{NeurIPS2022}}$	DF	25.1	**0.660**	4949.4	20.045	0.275
	RIVER [105]$_{\text{ICCV2023}}$	GAN	22.1	0.178	12514.5	7.764	0.870
	Ctrl-Crash [28]$_{\text{Arxiv2025}}$	DF	26.5	0.624	**4343.4**	17.566	0.307
	FAR [92]$_{\text{Arxiv2025}}$	AR	24.8	0.435	5600.4	14.179	0.240
	ADVersa-RAVD	DF	**26.8**	0.626	5069.1	**20.192**	**0.233**

GAN: Generative Adversarial Networks; **DF**: Diffusion; **AR**: Autoaggressive.

TABLE V
THE ACCIDENT CAUSE AND CATEGORY REASONING PERFORMANCE OF SOTA MLLM MODELS. (**BOLD** FONT: THE BEST; **FT**: FINE-TUNED).

Task	Method	FT	B1↑	B2↑	B3↑	B4↑	M↑	RL↑	S↑	Cr↑
(R)	InternVL3.5 (4B) [106]		8.50	3.20	1.60	0.60	9.60	11.30	6.00	0.04
	InternVL3 (8B) [107]		7.20	1.10	0.30	0.10	4.10	10.20	3.30	0.04
	Mobile-VideoGPT (1.5B) [108]		10.40	2.70	0.60	0.20	5.80	11.00	4.80	0.05
	Qwen2.5-VL (3B) [109]		5.60	1.10	0.30	0.10	3.30	6.10	0.90	0.01
	InternVL3.5 (14B) [106]		5.60	1.90	0.80	0.20	6.80	6.70	5.80	0.03
	LLaVA-Mini (8B) [110]		10.40	1.90	0.80	0.50	4.60	10.90	5.80	0.03
	AVD2 [29]	✓	27.40	22.40	19.80	17.90	17.20	28.30	**25.40**	**1.61**
	ADVersa-RAVD	✓	**39.30**	**31.70**	**27.20**	**23.50**	**28.90**	**43.10**	22.40	0.50
(P)	InternVL3.5 (4B) [106]		50.70	46.60	44.40	41.80	42.00	82.80	68.60	0.03
	InternVL3 (8B) [107]		50.00	46.00	43.50	41.30	41.70	81.60	72.00	0.20
	Mobile-VideoGPT (1.5B) [108]		50.00	37.90	23.80	9.90	32.10	60.30	46.00	0.07
	Qwen2.5-VL (3B) [109]		73.60	68.40	65.60	63.30	44.30	84.30	71.40	0.01
	InternVL3.5 (14B) [106]		80.60	76.20	73.60	71.40	43.60	84.20	71.60	0.10
	LLaVA-Mini (8B) [110]		80.40	75.60	73.20	70.80	48.40	84.400	73.50	0.78
	AVD2 [29]	✓	73.60	69.00	65.90	63.30	**55.30**	80.00	68.90	0.90
	ADVersa-RAVD	✓	**80.70**	**77.70**	**76.00**	**74.40**	54.80	**88.80**	**75.00**	**1.62**

Here, we further discuss the abductive ability of ADVersa-RAVD explicitly by predicting the future or recovering past scenes (**Temporally**).

1) Result Analysis on Frame Recovering or Prediction: To evaluate ADVersa-RAVD, we compare many SOTA methods for accident video frame recovering (**R**) and prediction (**P**) tasks. They are Dynamics-Aware Implicit Generative Adversarial Networks [104] (DIGAN), Masked Conditional Video Diffusion Model [91] (MCVD), Random Frame Conditioned Flow Integration for Video Prediction Model [105] (RIVER), Frame Autoregressive Model [92] (FAR), and one accident video synthesis work: Controllable Car Crash Video Generation Model [28] (Ctrl-Crash). In particular, these models adopt different designs (GAN, DF-Diffusion, and AR-Autoregressive) for video generation.

Recovering the crash past is not just inpainting, as it requests the model to infer a causal-consistent past with natural scene evolution. Unlike other works, our ADVersa-RAVD is guided by context-aligned textual prompts (*e.g.*, t_a or t_r), further enhanced by conditional scene-graph representations (*e.g.*, $\mathcal{G}_a$ or $\mathcal{G}_r$) for both frame recovering or prediction. This allows the model to better uncover underlying causal cues through enhanced visual-text alignment and generate physically plausible crash past or future. The effectiveness is validated by Tab. IV, and our ADVersa-RAVD achieves the best performance for all evaluation metrics, indicating that ADVersa-RAVD captures the consistency of main identities. Fig. 9 further verifies this effect. FAR [92] and MCVD [91] show visual artifacts such as ghosting and double edges, and DIGAN [104] often loses the scene details and fails to render the main identities in the recovered video. Unlike Ctrl-Crash [28] with almost unchanged visual content, ADVersa-RAVD preserves a natural scene evolution and suppresses temporal drift. The same pattern appears in the frame prediction (right of Fig. 9), we can see that only ADVersa-RAVD predicts the crash from the observed context frames. In particular, the crashing process often results in *motion collapse*, which indicates why ADVersa-RAVD does not achieve the best SSIM or FVD value, but with the highest CLIP_s value, as shown in Tab. IV.

2) Result Analysis on Accident Cause/Category Reasoning: To evaluate the textual abductive ability of ADVersa-RAVD, we analyze the textual accident cause and category reasoning, and the most recent and SOTA AVD2 [29] is

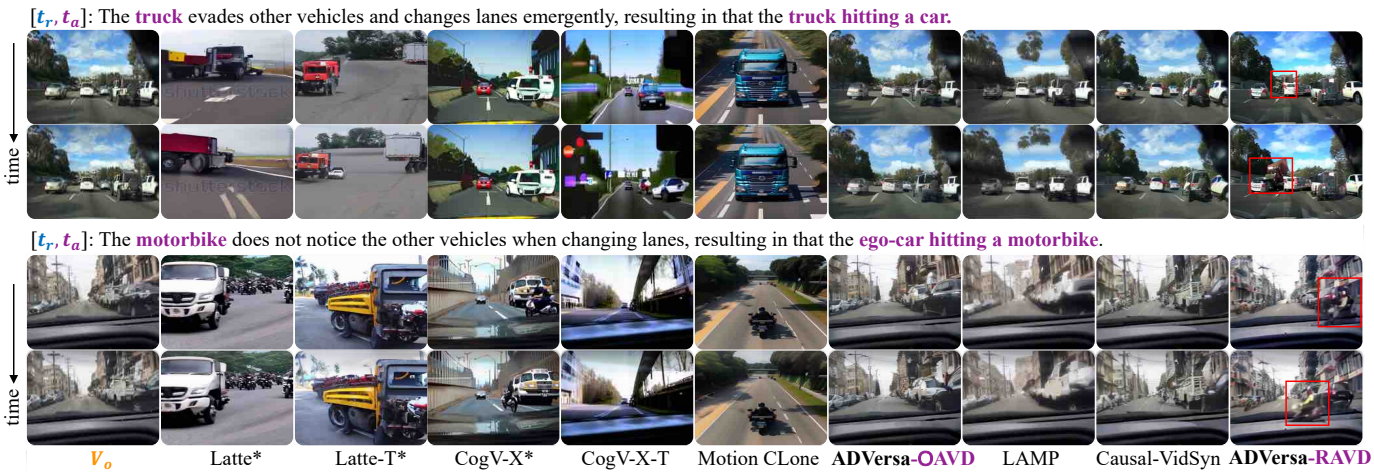


Fig. 10. Sample visualizations of N2A task with accident-free video prompt (V_o) and textual accident cause and collision prompt (t_r,t_a) by Latte* [111], Latte-T [111], CogV-X* [112], CogV-X-T [112], MotionClone [113], ADVersa-OAVD [9], LAMP [114], Causal-VidSyn [27] and ADVersa-RAVD (Best viewed in zoom mode).

selected as a baseline, which undergoes the same training as our ADVersa-RAVD with the MM-AU dataset. Furthermore, we compare several leading open-source Multi-Modal Large-Language Models (MLLMs), including InternVL3(8B) [107], InternVL3.5(4B), InternVL3.5(14B) [106], Mobile-VideoGPT(1.5B) [108], Qwen2.5-VL(3B) [109], and LLaVA-Mini(8B) [110]. Notably, these MMLMs are evaluated directly in a zero-shot setting to assess their general reasoning abilities for textual accident cause and category reasoning.

As shown in Tab. V, the task-specific fine-tuning is essential for accident reasoning. In the Rec-CrashPast task that involves a complex accident cause description, both the fine-tuned AVD2 [29] and our ADVersa-RAVD consistently outperform zero-shot MMLMs. This gap indicates that generic video–language priors are insufficient to capture accident-centric causal semantics without domain adaptation. Fig. 9 presents only AVD2 [29] and our ADVersa-RAVD reasoning close to the ground truth for both *Accident Cause Reasoning* As for *Accident Category Reasoning*, because of the simple text expression, such as "*a car or ego-car accident*", some very new open-source MLLMs present promising reasoning results, such as InternVL3.5 (14B) [106] and LLaVA-Mini (8B) [110]. In summary, domain adaptation of zero-shot MLLMs is needed for textual accident scene reasoning.

3) Result Analysis on Accident Video Synthesis: Accident video synthesis here aims to check the ability of answering "*What the key elements of the accident are?*" by changing or adding the key entity words in the original text prompts. We select several state-of-the-art (SOTA) video diffusion models, divided as training-free and training-based methods for a fair comparison. The training-free methods directly leverage pre-trained models for inference, including ControlVideo [103] (*Abbrev.*, CoVideo), T2V-zero [115], Free-bloom [116], Latte* [111], CogVideoX* (*Abbrev.*, CogV-X* [112]), and MotionClone [113]. Training-based methods are fine-tuned on the same training set as our ADVersa-RAVD, such as ADVersa-OAVD [9]), Tune-A-Video [117] (*Abbrev.*, TAV), LAMP [114], Causal-VidSyn [27], as well as the trained versions Latte-T and CogVideoX-T (*Abbrev.*, CogV-X-T).

TABLE VI
PERFORMANCE ON N2A AND AEDIT TASKS (**BOLD**: THE BEST).

Method	$CLIP_s$↑	FVD↓	TempC↑	Backbone	Afd↑(%)
N2A task (BDD-A [89] (2000))					
T2V-zero [115]$_{CVPR2023}$	<u>26.0</u>	11754.8	0.992	Unet	-
Free-bloom [116]$_{NeurIPS2024}$	25.1	8280.1	0.990	Unet	-
CoVideo [103]$_{ICLR2024}$	24.2	9906.1	0.930	Unet	-
Latte* [111]$_{TMLR2025}$	25.8	11066.2	0.993	DiT	-
CogV-X* [112]$_{ICLR2025}$	25.6	9075.6	0.992	DiT	-
MotionClone [113]$_{ICLR2025}$	24.9	11554.9	0.994	Unet	-
TAV [117]$_{ICCV2023}$	24.1	9305.3	0.902	Unet	-
ADVesrsa-OAVD [9]$_{CVPR2024}$	25.8	6378.9	0.992	Unet	-
LAMP [114]$_{CVPR2024}$	25.4	6208.2	0.991	Unet	-
Latte-T [111]$_{TMLR2025}$	25.3	11196.6	0.993	DiT	-
CogV-X-T [112]$_{ICLR2025}$	24.6	10094.0	0.993	DiT	-
Causal-VidSyn [27]$_{ICCV2025}$	26.5	6192.3	0.994	Unet	-
ADVesrsa-RAVD	**26.6**	**5832.6**	0.982	Unet	-
AEdit task (DADA [13] (3000))					
TAV [117]$_{ICCV2023}$	23.8	10076.2	0.909	Unet	-
Latte-T [111]$_{TMLR2025}$	28.2	12377.3	0.945	DiT	-
CogV-X-T [112]$_{ICLR2025}$	25.3	11420.5	0.945	DiT	-
LAMP [114]$_{CVPR2024}$	26.1	6191.6	**0.971**	Unet	34.7
ADVersa-OAVD [9]$_{CVPR2024}$	26.9	5358.2	0.947	Unet	49.4
Causal-VidSyn [27]$_{ICCV2025}$	28.7	**5352.9**	0.940	Unet	55.4
ADVesrsa-RAVD	**28.9**	5885.3	0.962	Unet	**55.8**

From Tab. VI, Causal-VidSyn [27], our ADVersa-OAVD [9] and ADVersa-RAVD perform better than others for N2A task. However, in the top example of Fig. 10, when the prompt involves a large object (*e.g.*, *truck*), only our ADVersa-RAVD generates correct geometry and scale and preserves the background style. In the bottom sample of Fig. 10, ADVersa-OAVD [9] and Causal-VidSyn [27] tend to generate a tiny and indistinct motorbike, whereas ADVersa-RAVD produces a clear motorbike that aligns with the accident cause and collision prompt [t_r, t_a]. In contrast, although the generated frames of CogV-X* [112] and MotionClone [113] match the text semantics, they disrupt the background style of the visual prompt and generate irrelevant objects, indicating a weak

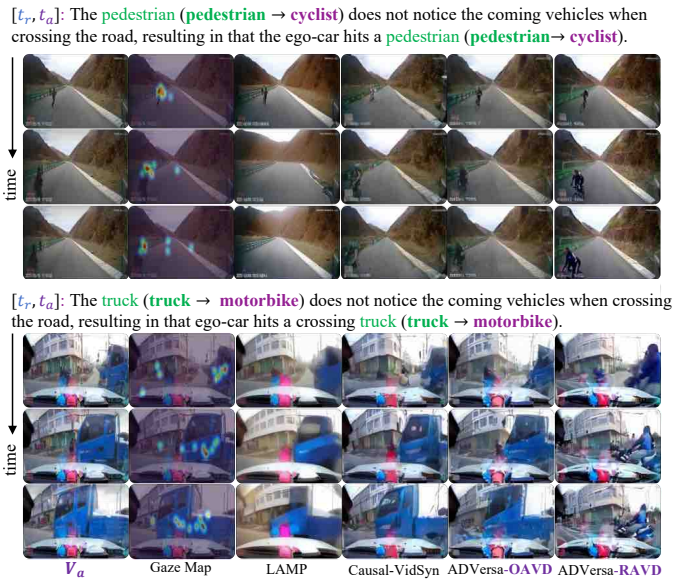


Fig. 11. We visualize two AEdit results prompted by near-crash segment (V_r), the collision segment (V_a) and the varied textual accident cause and collision prompt (t_r,t_a) by LAMP [114], Causal-VidSyn [27], ADVersa-OAVD [9], and ADVersa-RAVD.

coupling between entity generation and scene layout.

From Fig. 10, we can observe that only LAMP, ADVersa-OAVD, Causal-VidSyn and ADVersa-RAVD can maintain the background style. We further show the examples in the AEdit task in Fig. 11. We can see that only ADVersa-RAVD can identify the large object better while maintaining the highlighted regions in the driver gaze maps (*e.g.*, the "truck"→"motorbike" editing sample in Fig. 11(bottom)). Furthermore, when we change "pedestrian" to "cyclist" (Fig. 11 (top)), we can see that Causal-VidSyn [27], ADVersa-OAVD [9], and ADVersa-RAVD show an active response, while the shape of the cyclist is clearer in ADVersa-RAVD, and LAMP [114] generates a failed response in the sample of Fig. 11 (top). To quantitatively measure whether text-reflected video content is synthesized, we introduce the Afd metric. From Tab. VI, we observe that our ADVersa-RAVD obtains the best Afd value to ground the text-reflected visual entity.

E. Diagnostic Analysis

We dismantle main components in ADVersa-RAVD and provide the model versions without Scene-Graph (CGVP) ("w/o SG"), the Style-transfer ("w/o ST"), the Depth & Segment maps ("w/o D&S"), the Depth maps ("w/o Depth maps"), the Segment maps ("w/o Segment maps"), and the Text-Reasoning ("w/o Text-Reasoning"). Then, we analyze the role of each component extensively. Notably, the version "w/o SG" denotes the Depth & Segment maps are all removed in the scene graph, but they are maintained to retain the ControlNet [82]. In contrast, the "w/o D&S", "w/o Depth maps", and "w/o Segment maps" specifically refer to removing these control signals from ControlNet [82]. These versions are all re-trained.

Tab. VII shows the ablation results for frame recovery or frame prediction tasks. Tab. VIII shows the results for the

TABLE VII
THE DIAGNOSTIC EVALUATION OF ADVERSA-RAVD ON FRAME RECOVERY (R) OR FRAME PREDICTION (P) TASKS.

Tasks	Methods	CLIP$_s$↑	SSIM↑	FVD↓	PSNR↑	LPIPS↓
(R)	w/o Scene-Graph (CGVP)	26.4	0.631	3261.9	**21.530**	**0.169**
	w/o Style-transfer	25.7	0.546	5722.1	18.081	0.321
	w/o Depth&Segment maps	26.4	0.604	2936.7	19.070	0.254
	w/o Depth-maps	26.4	0.633	3498.6	19.058	0.282
	w/o Segment-maps	25.3	0.564	3907.7	19.035	0.365
	w/o Text-reasoning	27.2	0.629	2928.3	19.124	0.220
	ADVersa-RAVD [Full]	**27.8**	**0.637**	**2866.4**	19.292	0.211
(P)	w/o Scene-Graph (CGVP)	26.4	0.591	5711.4	19.197	0.278
	w/o Style-transfer	25.6	0.545	6180.8	18.023	0.325
	w/o Depth&Segment maps	26.0	0.602	5329.6	20.173	0.241
	w/o Depth-maps	26.5	0.600	6179.9	19.292	0.283
	w/o Segment-maps	26.4	0.609	5729.4	19.084	0.278
	w/o Text-reasoning	26.7	0.615	5090.6	20.120	0.235
	ADVersa-RAVD [Full]	**26.8**	**0.626**	**5069.1**	**20.192**	**0.233**

textual accident cause and category reasoning task, and Tab. IX shows the diagnostic results for the N2A and AEdit tasks.

‣Roles of Style-transfer (§IV-C1). From Tab. VII and Fig. 12 (left), we find Style-transfer plays an important role, where the substantial metric degradation occurs upon removal, and the qualitative visualizations ("w/o ST") exhibit background style shifts and object inconsistencies, respectively. This indicates that causal-consistent visual generation critically relies on maintaining temporal coherence for recovery or prediction tasks. Consequently, by employing the proposed style-conditioned prior transfer, perceptual abruptness is alleviated, and causal-consistent visual generation is enhanced. Furthermore, as evidenced in Tab. VIII, removing style transfer leads to a notable decline for accident category or cause reasoning tasks. Moreover, as shown in Tab. IX, we validate its importance in the N2A and AEdit tasks, where the performance metrics drop consistently when the style transfer is removed. The visual evidence in Fig. 12 (right) also confirms that without style transfer, the "motorbike" in the generated frames in N2A or AEdit examples exhibits temporal flicker, such as multiple motorcycles at different locations. Contrarily, our style-conditioned prior transfer module mitigates perceptual abruptness while bolstering causal consistency.

‣Roles of Scene-Graph (CGVP). ADVersa-RAVD is driven by the pre-trained CGVP that the latent representation of the scene graph can be attentively injected into the video diffusion process to maintain enhanced visual–text alignment. As shown in Tab. VII, removing CGVP ("w/o SG") causes a drop in CLIP$_s$ for recovery and prediction tasks. Moreover, in Fig. 12 (left), the model even predicts incorrect semantic relations (*e.g.*, misclassifying a cyclist as a car). In contrast, since causal-consistent performance has already been largely preserved by the style-transfer, text reasoning tasks are less affected. This suggests that the text-reasoning task is more dependent on temporal–causal consistency, where CGVP contributes primarily to semantic relation alignment. In addition, the scene graph is crucial for encoding the structural information of critical objects. As shown in Fig. 12 (right), removing CGVP results in objects not preserving the correct positions. These observations highlight the CGVP in enhancing semantic relation

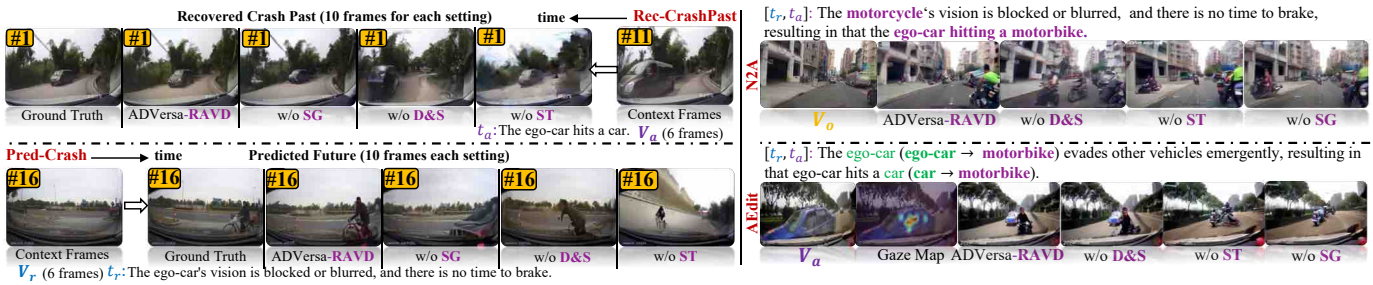


Fig. 12. Visualizations of one Rec-CrashPast, Pred-Crash, N2A and AEdit example with the ablation checking of ADVersa-RAVD.

TABLE VIII
THE DIAGNOSTIC EVALUATIONS OF ADVERSA-RAVD ON TEXTUAL ACCIDENT CAUSE AND CATEGORY REASONING TASK FINE-TUNED ON VIDEO-LLAVA-7B [83].

Task	Methods	B1↑	B2↑	B3↑	B4↑	M↑	RL↑	S↑	Cr↑
(R)	w/o Depth&Segment maps	39.0	31.5	27.0	23.3	28.8	42.9	22.5	0.49
	w/o Scene-Graph (CGVP)	38.9	30.6	25.0	20.1	22.6	42.7	22.8	0.46
	w/o Style-transfer	37.9	30.2	24.6	20.3	22.3	42.4	22.5	0.43
	w/o Depth-maps	37.3	29.6	25.0	21.1	23.6	42.1	**26.5**	0.42
	w/o Segment-maps	37.7	29.5	26.3	22.8	22.6	**44.5**	26.2	0.41
	ADVesrsa-RAVD[Full]	**39.3**	**31.7**	**27.2**	**23.5**	**28.9**	43.1	22.4	**0.50**
(P)	w/o Depth&Segment maps	79.8	76.3	74.3	72.5	53.3	87.8	75.0	1.27
	w/o Scene-Graph (CGVP)	80.1	76.8	74.9	73.2	53.9	88.1	75.0	1.38
	w/o Style-transfer	80.0	76.5	74.6	72.8	53.7	87.9	75.0	1.32
	w/o Depth-maps	79.5	75.8	73.7	71.9	52.6	87.4	75.0	1.18
	w/o Segment-maps	79.8	75.6	74.5	72.9	52.8	87.2	75.0	1.18
	ADVesrsa-RAVD [Full]	**80.7**	**77.7**	**76.0**	**74.4**	**54.8**	**88.8**	75.0	**1.62**

TABLE IX
THE DIAGNOSTIC EVALUATIONS ON **N2A** AND **AEDIT** TASKS (**BOLD** FONT: THE BEST).

Method	CLIP_s↑	FVD↓	TempC↑	Afd↑(%)	W-Afd↑
N2A task					
w/o Depth&Segment maps	26.4	5840.3	0.975	-	-
w/o Scene-Graph (CGVP)	26.5	6013.7	0.973	-	-
w/o Style-transfer	26.4	6207.2	0.977	-	-
ADVesrsa-RAVD [Full]	**26.6**	**5832.6**	**0.982**	-	-
AEdit task					
LAMP [114]$_{\text{CVPR2024}}$	26.1	6191.6	**0.971**	34.7	0.57
ADVersa-OAVD [9]$_{\text{CVPR2024}}$	26.9	5358.2	0.947	49.4	0.76
Causal-VidSyn [27]$_{\text{ICCV2025}}$	28.7	**5352.9**	0.940	55.4	0.86
w/o Depth&Segment maps	28.1	5916.3	0.968	59.7	1.24
w/o Scene-Graph (CGVP)	28.1	7941.6	0.967	59.6	1.23
w/o Style-transfer	27.7	7078.5	**0.971**	**62.7**	1.15
ADVesrsa-RAVD [Full]	**28.9**	5885.3	0.962	55.8	**1.31**

alignment and structural fidelity in frame generation.

Roles of Depth & Segment Maps in ControlNet. Depth & Segment maps are the key conditional signals in ControlNet. As shown in Tab. VII, removing Depth & Segment maps entirely results in a performance drop. Further analysis of individual signals shows that segment maps are particularly important because removing segment maps leads to a significant degradation, which indicates that instance-level segmentation provides strong constraints for recovery and prediction tasks. Interestingly, as presented in Tab. VIII, removing depth or segment maps yields slight increases in ROUGE-L (RL) and SPICE (S) metrics. This is likely due to the limited sensitivity of these text-reasoning metrics, which capture high-level semantic alignment rather than structural fidelity. Overall, depth and segment maps help the model focus on critical objects more effectively. This effect is also verified by Fig. 12, *i.e.*, when depth and segment maps are removed, the generated results exhibit additional distracting objects with blurred shapes. Without the depth and segment maps, the model may lose strict control over object boundaries and geometry.

Roles of Text-Reasoning. From Tab. VII, text-reasoning exerts a modest positive influence on the diffusion process. This indicates that the model primarily relies on key conditional signals such as Depth&Segment maps and Scene Graph to learn causal-consistent visual generation and visual-text alignment, while text-reasoning provides only a marginal supplementary effect. Therefore, we only evaluate text-reasoning on frame recovery and prediction tasks.

Affordance (Afd) Ablation in the AEdit Task. From Tab. IX, we find that the evaluation by Afd can effectively validate that what is edited aligns with human preference, but overlooks the quality and difficulty of the edits involving a *large scale difference* issue (*e.g.*, synthesizing a large vehicle from a small motorbike shown in Fig. 12), leading to an inflated Afd score, such as ADVersa-RAVD "w/o Style-transfer" with the best Afd (62.7) but poorest visual quality (FVD: 7078.5). To address this, we propose the Weighted-Afd (**W-Afd**) by weighting each successful edit with the area ratio between the gazed region and the generated entity area, computed by:

$$\text{W-Afd} = \frac{1}{N_e}\sum_{e=1}^{N_e}\mathbb{I}(\text{IOU}(B_e^{det}, B_e^{gaze}) > 0)\cdot\frac{A(B_e^{gaze})}{A(B_e^{det})}, \quad (18)$$

where $A(\cdot)$ returns the pixel area of a given bounding box. As mentioned earlier, Afd quantifies the human perceptual judgment of generated videos, while W-Afd provides a stricter criterion that can restrict the influence of "*large object scale difference*" in the editing process, and reflects how well-grounded objects are aligned with human attention in terms of both location and scale.

From Tab. IX and Fig. 12, the W-Afd can provide a more trustworthy evaluation of text-reflected visual editing.

Roles of the Frame Mask Regulator in §IV-C. Frame mask regulator provides a flexible visual frame prompt for Rec-CrashPast and Pred-Crash tasks. We evaluate it by setting different numbers of input context frames, denoted as $L \in \{1, 3, 4, 6\}$. When L=1, it is the most extreme situation, and only one frame is observed for predicting its future or recovering its past. Unsurprisingly, from Fig. 13, this setting

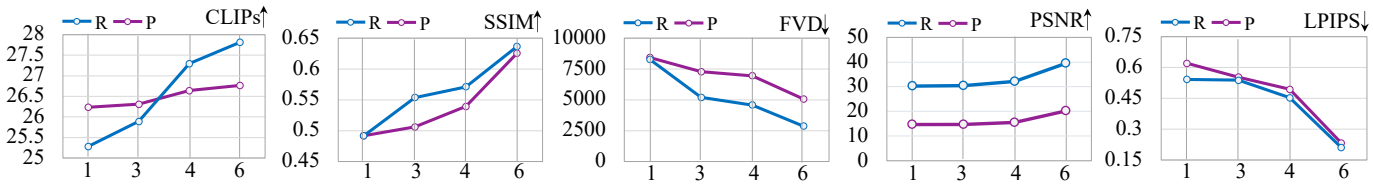


Fig. 13. The ablation studies for different lengths of context frames for frame recovering and frame prediction by our ADVersa-RAVD. Markers correspond to $L = 1, 3, 4, 6$ from left to right on each curve.

obtains the worst results for the Rec-CrashPast and Pred-Crash tasks. Certainly, we can expect that more observed frames (*i.e.*, $L > 1$) will obtain better performance. Under this assumption, when we gradually add more observed frames, we obtain clear performance improvements, as presented in Fig. 13. However, we find the Rec-CrashPast task consistently demonstrates a more pronounced improvement than the Pred-Crash task. This is evidenced by the steeper **R** curves compared to the **P** curves in metrics such as CLIP$_s$ and SSIM. This phenomenon implies that the crashing frames with significant motion distortion are harder to synthesize than the stable motion frames. Because the prediction task implies more uncertain motion variation, longer context frames are required to approach the critical event.

Interestingly, from Tab. V and Tab. VIII, we find that the textual reasoning about the cause and category of accident exhibits a different phenomenon, where reasoning about accident categories (*e.g.*, "car accident") is often simpler than reasoning complex cause descriptions. This highlights a difference between visual reasoning and textual reasoning.

VI. Conclusion and Future Insights

This work presents ADVersa, a unified framework that addresses three subtasks: 1) visual past recovery of near-crash scenes, 2) visual prediction of near-crash scenes, and 3) accident video synthesis for an abductive accident video understanding framework. To support ADVersa, a large-scale ego-view multi-modal accident dataset (MM-AU) is constructed, along with a suite of benchmarks on accident video understanding.

In ADVersa, we propose two methods: ~~the~~ Object-centric Accident Video Diffusion (ADVersa-OAVD) method is driven by an Abductive-CLIP model and performs implicit abductive reasoning, and an explicit abductive learning method Relation-Centric Accident Video Diffusion model (ADVersa-RAVD), which is driven by the Abductive-CLIP model and the Contrastive Graph Video Pre-Training (CGVP) model together that enhances the semantic learning and alignment of key regions and text-video co-occurrence. Extensive experiments verify that ADVersa-OAVD and ADVersa-RAVD show promising abilities for abductive accident video understanding, outperforming superior video diffusion performance over many state-of-the-art diffusion models. In particular, our proposed Weighted-Affordance (W-Afd) metric offers a more trustworthy evaluation of text-reflected visual editing and provides a strong metric for exploring the visual grounding task in accident video understanding. In the future, we will investigate video diffusion for some downstream tasks, such as accident anticipation and decision-making.

References

[1] Editorial, "Safe driving cars," *Nat. Mach. Intell.*, vol. 4, pp. 95–96, 2022.
[2] M. Godbole, X. Gao, and Z. Tu, "Drama-x: A fine-grained intent prediction and risk reasoning benchmark for driving," *arXiv preprint arXiv:2506.17590*, 2025.
[3] H. Yu, X. Zhang, Y. Wang, Q. Huang, and B. Yin, "Fine-grained accident detection: database and algorithm," *IEEE Trans. Image Process.*, vol. 33, pp. 1059–1069, 2024.
[4] C. Liang, W. Wang, T. Zhou, and Y. Yang, "Visual abductive reasoning," in *CVPR*, 2022, pp. 15 544–15 554.
[5] H. Yang, Y. Zhou, J. Wu, H. Liu, L. Yang, and C. Lv, "Human-guided continual learning for personalized decision-making of autonomous driving," *IEEE Trans. Intell. Transp. Syst.*, vol. 26, no. 4, pp. 5435–5447, 2025.
[6] B. Liao, S. Chen, H. Yin, B. Jiang, C. Wang, S. Yan, X. Zhang, X. Li, Y. Zhang, Q. Zhang *et al.*, "Diffusiondrive: Truncated diffusion model for end-to-end autonomous driving," in *CVPR*, 2025, pp. 12 037–12 047.
[7] J. Zhang, Y. Guan, C. Wang, H. Liao, G. Zhang, and Z. Li, "LATTE: A real-time lightweight attention-based traffic accident anticipation engine," *Inf. Fusion*, vol. 122, p. 103173, 2025.
[8] Y. Kumamoto, K. Ohtani, D. Suzuki, M. Yamataka, and K. Takeda, "AAT-DA: accident anticipation transformer with driver attention," in *WACV*, 2025, pp. 1052–1061.
[9] J. Fang, L. Li, J. Zhou, J. Xiao, H. Yu, C. Lv, J. Xue, and T. Chua, "Abductive ego-view accident video understanding for safe driving perception," in *CVPR*, 2024, pp. 22 030–22 040.
[10] Y. Yao, M. Xu, Y. Wang, D. J. Crandall, and E. M. Atkins, "Unsupervised traffic accident detection in first-person videos," in *IROS*, 2019, pp. 273–280.
[11] Y. Yao, X. Wang, M. Xu, Z. Pu, Y. Wang, E. M. Atkins, and D. J. Crandall, "Dota: Unsupervised detection of traffic anomaly in driving videos," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 1, pp. 444–459, 2023.
[12] M. M. Karim, Z. Yin, and R. Qin, "An attention-guided multistream feature fusion network for early localization of risky traffic agents in driving videos," *IEEE Trans. Intell. Veh.*, vol. 9, no. 1, pp. 1792 – 1803, 2024.
[13] J. Fang, D. Yan, J. Qiao, J. Xue, and H. Yu, "DADA: driver attention prediction in driving accident scenarios," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 6, pp. 4959–4971, 2022.
[14] M. Kang, W. Lee, K. Hwang, and Y. Yoon, "Vision transformer for detecting critical situations and extracting functional scenario for automated vehicle safety assessment," *Sustainability*, vol. 14, no. 15, p. 9680, 2022.
[15] H. Luo and F. Wang, "A simulation-based framework for urban traffic accident detection," in *ICASSP*, 2023, pp. 1–5.
[16] J. Fang, J. Qiao, J. Xue, and Z. Li, "Vision-based traffic accident detection and anticipation: A survey," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, no. 4, pp. 1983–1999, 2024.
[17] T. Ji, N. Chakraborty, A. Schreiber, and K. Driggs-Campbell, "An expert ensemble for detecting anomalous scenes, interactions, and behaviors in autonomous driving," *Int. J. Robotics Res.*, vol. 44, no. 6, pp. 1055–1077, 2025.
[18] Z. Xing, H. Chen, B. Xie, J. Xu, Z. Guo, X. Xu, J. Hao, C.-W. Fu, X. Hu, and P.-A. Heng, "Echotraffic: Enhancing traffic anomaly understanding with audio-visual insights," in *CVPR*, 2025, pp. 19 098–19 108.
[19] W. Bao, Q. Yu, and Y. Kong, "Uncertainty-based traffic accident anticipation with spatio-temporal relational learning," in *ACM MM*, 2020, pp. 2682–2690.
[20] M. M. Karim, Y. Li, R. Qin, and Z. Yin, "A dynamic spatial-temporal attention network for early anticipation of traffic accidents," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 7, pp. 9590–9600, 2022.

[21] A. V. Malawade, S. Yu, B. Hsu, D. Muthirayan, P. P. Khargonekar, and M. A. A. Faruque, "Spatiotemporal scene-graph embedding for autonomous vehicle collision prediction," *IEEE Internet Things J.*, vol. 9, no. 12, pp. 9379–9388, 2022.

[22] A. Patel, B. Li, M. S. Rasooli, N. Constant, C. Raffel, and C. Callison-Burch, "Bidirectional language models are also few-shot learners," in *ICLR*, 2023.

[23] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," *NeurIPS*, vol. 36, pp. 34 892–34 916, 2023.

[24] J. Chen, D. Zhu, X. Shen, X. Li, Z. Liu, P. Zhang, R. Krishnamoorthi, V. Chandra, Y. Xiong, and M. Elhoseiny, "Minigpt-v2: large language model as a unified interface for vision-language multi-task learning," *arXiv preprint arXiv:2310.09478*, 2023.

[25] Y. Xu, "Traffic incident detection based on hmm," in *ICIST*, 2013, pp. 942–945.

[26] R. Zhang, B. Wang, J. Zhang, Z. Bian, C. Feng, and K. Ozbay, "When language and vision meet road safety: leveraging multimodal large language models for video-based traffic accident analysis," *arXiv preprint arXiv:2501.10604*, 2025.

[27] L.-l. Li, J. Fang, J. Xiao, S. Pang, H. Yu, C. Lv, J. Xue, and T.-S. Chua, "Causal-entity reflected egocentric traffic accident video synthesis," in *ICCV*, 2025.

[28] A. Gosselin, G. Y. Luo, L. Lara, F. Golemo, D. Nowrouzezahrai, L. Paull, A. Jolicoeur-Martineau, and C. Pal, "Ctrl-crash: Controllable diffusion for realistic car crashes," *arXiv preprint arXiv:2506.00227*, 2025.

[29] C. Li, K. Zhou, T. Liu, Y. Wang, M. Zhuang, H.-a. Gao, B. Jin, and H. Zhao, "Avd2: Accident video diffusion for accident video description," in *ICRA*, 2025.

[30] L. Xu, H. Huang, and J. Liu, "SUTD-TrafficQA: A question answering benchmark and an efficient network for video reasoning over traffic events," in *CVPR*, 2021, pp. 9878–9888.

[31] Y. Liu, G. Li, and L. Lin, "Cross-modal causal relational reasoning for event-level visual question answering," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 10, pp. 11 624–11 641, 2023.

[32] C. Zang, H. Wang, M. Pei, and W. Liang, "Discovering the real association: Multimodal causal reasoning in video question answering," in *CVPR*, 2023, pp. 19 027–19 036.

[33] J. Xiao, P. Zhou, A. Yao, Y. Li, R. Hong, S. Yan, and T. Chua, "Contrastive video question answering via video graph transformer," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 11, pp. 13 265–13 280, 2023.

[34] M. Shoman, D. Wang, A. Aboah, and M. Abdel-Aty, "Enhancing traffic safety with parallel dense video captioning for end-to-end event analysis," in *CVPR*, 2024, pp. 7125–7133.

[35] Q. M. Dinh, M. K. Ho, A. Q. Dang, and H. P. Tran, "Trafficvlm: A controllable visual language model for traffic video captioning," in *CVPR*, 2024, pp. 7134–7143.

[36] C. Liu *et al.*, "Traffic scenario understanding and video captioning via guidance attention captioning network," *IEEE Trans. Intell. Transp. Syst.*, vol. 25, no. 5, pp. 3615–3627, 2023.

[37] S. Malla, C. Choi, I. Dwivedi, J. H. Choi, and J. Li, "Drama: Joint risk localization and captioning in driving," in *WACV*, 2023, pp. 1043–1052.

[38] Z. Zhou, C. Cui, Y. Yang, B. Wang, X. Tang, N. S. Alghamdi, Z. Li, and F. Huang, "Content-guided video captioning with cross-modal transformer for driving scenario understanding," *IEEE Trans. Intell. Transp. Syst., DOI: 10.1109/tits.2025.3572682*, 2025.

[39] K. T. Xuan, K. N. Nguyen, B. H. Ngo, V. D. Xuan, M.-H. An, and Q.-V. Dinh, "Divide and conquer boosting for enhanced traffic safety description and analysis with large vision language model," in *CVPR*, 2024, pp. 7046–7055.

[40] A. Dosovitskiy, G. Ros, F. Codevilla, A. M. López, and V. Koltun, "CARLA: an open urban driving simulator," in *CoRL*, vol. 78, 2017, pp. 1–16.

[41] R. F. Benekohal and J. Treiterer, "Carsim: Car-following model for simulation of traffic in normal and stop-and-go conditions," *Transportation Research Record*, vol. 1194, pp. 99–111, 1988.

[42] C. Xu, A. Petiushko, D. Zhao, and B. Li, "Diffscene: Diffusion-based safety-critical scenario generation for autonomous vehicles," in *AAAI*, 2025, pp. 8797–8805.

[43] J. Wang, A. Pun, J. Tu, S. Manivasagam, A. Sadat, S. Casas, M. Ren, and R. Urtasun, "Advsim: Generating safety-critical scenarios for self-driving vehicles," in *CVPR*, 2021, pp. 9909–9918.

[44] W. Ding, B. Chen, B. Li, K. J. Eun, and D. Zhao, "Multimodal safety-critical scenarios generation for decision-making algorithms evaluation," *IEEE Robotics Autom. Lett.*, vol. 6, no. 2, pp. 1551–1558, 2021.

[45] Z. Guo, Y. Zhou, and C. Gou, "Drivinggen: Efficient safety-critical driving video generation with latent diffusion models," in *ICME*, 2024, pp. 1–6.

[46] L. Chen, X. Wei, J. Li, X. Dong, P. Zhang, Y. Zang, Z. Chen, H. Duan, Z. Tang, L. Yuan *et al.*, "Sharegpt4video: Improving video understanding and generation with better captions," *NeurIPS*, vol. 37, pp. 19 472–19 495, 2024.

[47] J. Xie, Z. Yang, and M. Z. Shou, "Show-o2: Improved native unified multimodal models," *arXiv preprint arXiv:2506.15564*, 2025.

[48] J. Xie, W. Mao, Z. Bai, D. J. Zhang, W. Wang, K. Q. Lin, Y. Gu, Z. Chen, Z. Yang, and M. Z. Shou, "Show-o: One single transformer to unify multimodal understanding and generation," in *ICLR*, 2025.

[49] C. Zhou, L. Yu, A. Babu, K. Tirumala, M. Yasunaga, L. Shamis, J. Kahn, X. Ma, L. Zettlemoyer, and O. Levy, "Transfusion: Predict the next token and diffuse images with one multi-modal model," in *ICLR*, 2025.

[50] F. Chan, Y. Chen, Y. Xiang, and M. Sun, "Anticipating accidents in dashcam videos," in *ACCV*, vol. 10114, 2016, pp. 136–153.

[51] J. Fang, J. Qiao, J. Bai, H. Yu, and J. Xue, "Traffic accident detection via self-supervised consistency learning in driving scenarios," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 7, pp. 9601–9614, 2022.

[52] H. Kim, K. Lee, G. Hwang, and C. Suh, "Crash to not crash: Learn to identify dangerous vehicles using a simulator," in *AAAI*, 2019, pp. 978–985.

[53] M. S. Aliakbarian, F. S. Saleh, M. Salzmann, B. Fernando, L. Petersson, and L. Andersson, "VIENA: A driving anticipation dataset," in *ACCV*, 2019, pp. 449–466.

[54] T. Wang, S. Kim, W. Ji, E. Xie, C. Ge, J. Chen, Z. Li, and P. Luo, "Deepaccident: A motion and accident prediction benchmark for V2X autonomous driving," *CoRR*, vol. abs/2304.01168, 2023.

[55] T. You and B. Han, "Traffic accident benchmark for causality recognition," in *ECCV*, vol. 12352, 2020, pp. 540–556.

[56] C. Parikh, D. Rawat, T. Ghosh, R. K. Sarvadevabhatla *et al.*, "Roadsocial: A diverse videoqa dataset and benchmark for road event understanding from social video narratives," in *CVPR*, 2025, pp. 19 002–19 011.

[57] C. Liu, Z. Li, F. Chang, S. Li, and J. Xie, "Temporal shift and spatial attention-based two-stream network for traffic risk assessment," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 8, pp. 12 518–12 530, 2022.

[58] A. K. AlShami, A. Kalita, R. Rabinowitz, K. Lam, R. Bezbarua, T. Boult, and J. Kalita, "Coool: Challenge of out-of-label a novel benchmark for autonomous driving," *arXiv preprint arXiv:2412.05462*, 2024.

[59] N. P. Desai, A. Etemad, and M. A. Greenspan, "Cyclecrash: A dataset of bicycle collision videos for collision prediction and analysis," in *WACV*, 2025, pp. 6688–6698.

[60] J. Hessel *et al.*, "The abduction of sherlock holmes: A dataset for visual abductive reasoning," in *ECCV*, vol. 13696, 2022, pp. 558–575.

[61] H. Zhang, Y. K. Ee, and B. Fernando, "Rca: Region conditioned adaptation for visual abductive reasoning," in *ACM MM*, 2024, pp. 9455–9464.

[62] K. Tan, Z. Qi, J. Zhong, Y. Xu, and W. Zhang, "KN-VLM: knowledge-guided vision-and-language model for visual abductive reasoning," *Multim. Syst.*, vol. 31, no. 2, p. 146, 2025.

[63] T. Chen *et al.*, "Mecd: Unlocking multi-event causal discovery in video reasoning," *NeurIPS*, vol. 37, pp. 92 554–92 580, 2024.

[64] T. M. Le, V. Le, K. Do, S. Gupta, S. Venkatesh, and T. Tran, "Finding the trigger: Causal abductive reasoning on video events," *arXiv preprint arXiv:2501.09304*, 2025.

[65] M. Li *et al.*, "Multi-modal action chain abductive reasoning," in *ACL*, 2023, pp. 4617–4628.

[66] C. Tan, C. K. Yeo, C. Tan, and B. Fernando, "Abductive action inference," *CoRR*, vol. abs/2210.13984, 2022.

[67] ——, "Inferring past human actions in homes with abductive reasoning," in *WACV*, 2025, pp. 8249–8258.

[68] Y. K. Ee, H. Zhang, A. Matyasko, and B. Fernando, "Deduce and select evidences with language models for training-free video goal inference," in *WACV*, 2025, pp. 5937–5947.

[69] Z. Ge, S. Liu, F. Wang, Z. Li, and J. Sun, "YOLOX: exceeding YOLO series in 2021," *CoRR*, vol. abs/2107.08430, 2021.

[70] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *ECCV*, 2014, pp. 740–755.

[71] S. Chen, P. Sun, Y. Song, and P. Luo, "Diffusiondet: Diffusion model for object detection," in *ICCV*, 2023, pp. 19 830–19 843.

[72] F. Yu, H. Chen, X. Wang, W. Xian, Y. Chen, F. Liu, V. Madhavan, and T. Darrell, "BDD100K: A diverse driving dataset for heterogeneous multitask learning," in *CVPR*, 2020, pp. 2633–2642.

[73] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo *et al.*, "Segment anything," in *ICCV*, 2023, pp. 4015–4026.

[74] L. Yang, B. Kang, Z. Huang, X. Xu, J. Feng, and H. Zhao, "Depth anything: Unleashing the power of large-scale unlabeled data," in *CVPR*, 2024.

[75] J. Zhu *et al.*, "Tracking anything in high quality," *CoRR*, vol. abs/2307.13974, 2023.

[76] J. Fang, D. Yan, J. Qiao, J. Xue, H. Wang, and S. Li, "DADA-2000: can driving accident be predicted by driver attention? Analyzed by A benchmark," in *ITSC*, 2019, pp. 4303–4309.

[77] B. Ni, H. Peng, M. Chen, S. Zhang, G. Meng, J. Fu, S. Xiang, and H. Ling, "Expanding language-image pretrained models for general video recognition," in *ECCV*, 2022, pp. 1–18.

[78] O. Ashual and L. Wolf, "Specifying object attributes and relations in interactive scene generation," in *ICCV*, 2019, pp. 4560–4568.

[79] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *CVPR*, 2022, pp. 10 684–10 695.

[80] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *ICML*, 2021, pp. 8748–8763.

[81] J. Z. Wu, Y. Ge, X. Wang, S. W. Lei, Y. Gu, Y. Shi, W. Hsu, Y. Shan, X. Qie, and M. Z. Shou, "Tune-a-Video: One-shot tuning of image diffusion models for text-to-video generation," in *ICCV*, 2023, pp. 7623–7633.

[82] L. Zhang, A. Rao, and M. Agrawala, "Adding conditional control to text-to-image diffusion models," in *ICCV*, 2023, pp. 3836–3847.

[83] B. Lin, Y. Ye, B. Zhu, J. Cui, M. Ning, P. Jin, and L. Yuan, "Video-llava: Learning united visual representation by alignment before projection," in *EMNLP*, 2024, pp. 5971–5984.

[84] Y. Li, H. Liu, Q. Wu, F. Mu, J. Yang, J. Gao, C. Li, and Y. J. Lee, "Gligen: Open-set grounded text-to-image generation," in *CVPR*, 2023, pp. 22 511–22 521.

[85] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *NeurIPS*, vol. 33, pp. 6840–6851, 2020.

[86] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "Nerf: Representing scenes as neural radiance fields for view synthesis," *Communications of the ACM*, vol. 65, no. 1, pp. 99–106, 2021.

[87] Y. Wang, Z. Li, W. Zhang, Z. Zhang, B. Xie, X. Liu, W. Zeng, and X. Jin, "Scene graph disentanglement and composition for generalizable complex image generation," in *NeurIPS*, 2024.

[88] T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer, "Qlora: Efficient finetuning of quantized llms," in *NeurIPS*, 2023.

[89] Y. Xia, D. Zhang, J. Kim, K. Nakayama, K. Zipser, and D. Whitney, "Predicting driver attention in critical situations," in *ACCV*, vol. 11365, 2018, pp. 658–674.

[90] T. Unterthiner, S. van Steenkiste, K. Kurach, R. Marinier, M. Michalski, and S. Gelly, "Towards accurate generative models of video: A new metric & challenges," *CoRR*, vol. abs/1812.01717, 2018.

[91] V. Voleti, A. Jolicoeur-Martineau, and C. Pal, "Mcvd-masked conditional video diffusion for prediction, generation, and interpolation," *NeurIPS*, 2022.

[92] Y. Gu, w. Mao, and M. Z. Shou, "Long-context autoregressive video modeling with next-frame prediction," *arXiv preprint arXiv:2503.19325*, 2025.

[93] Z. Wang, "Image quality assessment: Form error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 604–606, 2004.

[94] C. Qi, X. Cun, Y. Zhang, C. Lei, X. Wang, Y. Shan, and Q. Chen, "Fatezero: Fusing attentions for zero-shot text-based video editing," in *ICCV*, 2023, pp. 15 886–15 896.

[95] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *CVPR*, 2018, pp. 586–595.

[96] Z. Xiao, Y. Zhou, S. Yang, and X. Pan, "Video diffusion models are training-free motion interpreter and controller," 2024.

[97] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *ACL*, 2002, pp. 311–318.

[98] R. Vedantam, C. Lawrence Zitnick, and D. Parikh, "Cider: Consensus-based image description evaluation," in *CVPR*, 2015, pp. 4566–4575.

[99] S. Banerjee and A. Lavie, "Meteor: An automatic metric for mt evaluation with improved correlation with human judgments," in *IEEvaluation@ACL*, 2005, pp. 65–72.

[100] C.-Y. Lin and F. J. Och, "Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics," in *ACL*, 2004, pp. 605–612.

[101] Y. Tewel, R. Gal, D. Samuel, Y. Atzmon, L. Wolf, and G. Chechik, "Add-it: Training-free object insertion in images with pretrained diffusion models," in *ICLR*, 2025.

[102] S. Liu, Z. Zeng, T. Ren, F. Li, H. Zhang, J. Yang, Q. Jiang, C. Li, J. Yang, H. Su *et al.*, "Grounding dino: Marrying dino with grounded pre-training for open-set object detection," in *ECCV*, 2024, pp. 38–55.

[103] Y. Zhang, Y. Wei, D. Jiang, X. Zhang, W. Zuo, and Q. Tian, "Controlvideo: Training-free controllable text-to-video generation," in *ICLR*, 2024.

[104] S. Yu, J. Tack, S. Mo, H. Kim, J. Kim, J.-W. Ha, and J. Shin, "Generating videos with dynamics-aware implicit generative adversarial networks," in *ICLR*, 2022.

[105] A. Davtyan, S. Sameni, and P. Favaro, "Efficient video prediction via sparsely conditioned flow matching," in *ICCV*, 2023, pp. 23 206–23 217.

[106] W. Wang *et al.*, "Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency," *arXiv preprint arXiv:2508.18265*, 2025.

[107] J. Zhu *et al.*, "Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models," *arXiv preprint arXiv:2504.10479*, 2025.

[108] A. Shaker *et al.*, "Mobile-videogpt: Fast and accurate video understanding language model," *arXiv preprint arXiv:2503.21782*, 2025.

[109] S. Bai *et al.*, "Qwen2.5-vl technical report," *arXiv preprint arXiv:2502.13923*, 2025.

[110] S. Zhang, Q. Fang, Z. Yang, and Y. Feng, "Llava-mini: Efficient image and video large multimodal models with one vision token," in *ICLR*, 2025.

[111] X. Ma, Y. Wang, G. Jia, X. Chen, Z. Liu, Y. Li, C. Chen, and Y. Qiao, "Latte: Latent diffusion transformer for video generation," *Trans. Mach. Learn. Res.*, 2025.

[112] Z. Yang, J. Teng, W. Zheng, M. Ding, S. Huang, J. Xu, Y. Yang, W. Hong, X. Zhang, G. Feng *et al.*, "Cogvideox: Text-to-video diffusion models with an expert transformer," in *ICLR*, 2025.

[113] P. Ling *et al.*, "Motionclone: Training-free motion cloning for controllable video generation," in *ICLR*, 2025.

[114] R. Wu, L. Chen, T. Yang, C. Guo, C. Li, and X. Zhang, "Lamp: Learn a motion pattern for few-shot video generation," in *CVPR*, 2024, pp. 7089–7098.

[115] L. Khachatryan *et al.*, "Text2video-zero: Text-to-image diffusion models are zero-shot video generators," in *ICCV*, 2023, pp. 15 908–15 918.

[116] H. Huang, Y. Feng, C. Shi, L. Xu, J. Yu, and S. Yang, "Free-bloom: Zero-shot text-to-video generator with llm director and ldm animator," *NeurIPS*, vol. 36, 2024.

[117] J. Z. Wu *et al.*, "Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation," in *ICCV*, 2023, pp. 7589–7599.