EQ-TAA: Equivariant Traffic Accident Anticipation via Diffusion-Based Accident Video Synthesis

Jianwu Fang, *Member, IEEE*, Lei-Lei Li, Zhedong Zheng, Hongkai Yu, Jianru Xue, *Member, IEEE*, Zhengguo Li, *Fellow, IEEE*, and Tat-Seng Chua

***Abstract*—Traffic Accident Anticipation (TAA) in traffic scenes is a challenging problem for achieving zero fatalities in the future. Current approaches typically treat TAA as a supervised learning task needing the laborious annotation of accident occurrence duration. However, the inherent long-tailed, uncertain, and fast-evolving nature of traffic scenes has the problem that real causal parts of accidents are difficult to identify and are easily dominated by data bias, resulting in a background confounding issue. Thus, we propose an Attentive Video Diffusion (AVD) model that synthesizes additional accident video clips by generating the causal part in dashcam videos, i.e., from normal clips to accident clips. AVD aims to generate causal video frames based on accident or accident-free text prompts while preserving the style and content of frames for TAA after video generation. This approach can be trained using datasets collected from various driving scenes without any extra annotations. Additionally, AVD facilitates an Equivariant TAA (EQ-TAA) with an equivariant triple loss for an anchor accident-free video clip, along with the generated pair of contrastive *pseudo-normal* and *pseudo-accident* clips. Extensive experiments have been conducted to evaluate the performance of AVD and EQ-TAA, and competitive performance compared to state-of-the-art methods has been obtained.**

***Index Terms*—Traffic accident anticipation, video diffusion model, causal reasoning, equivariant learning, dashcam videos.**

I. Introduction

VAST amount of road accidents have deprived many human lives each year. More than half of the deaths fall on vulnerable road users (*i.e.,* pedestrians, cyclists, and motorbikes), and road accidents are the leading killer for young people aged from 5 to 29 years old [1]. The cause of a traffic accident is significantly influenced by the car to the road network. This shocking fact motivates many researchers and institutes to devote efforts to traffic accident detection or anticipation systems [2].

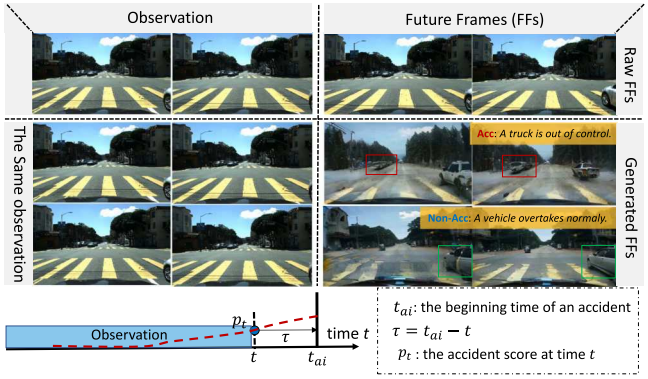


Fig. 1. The generated video frames by the Attentive Video Diffusion model for an accident (red boxes mark accident objects) and an accident-free (marked by green boxes) scenario. The raw videos are sampled in an accident-free HEV-I dataset [5].

In this work, we focus on the Traffic Accident Anticipation (TAA) problem which is different from the Traffic Accident Detection (TAD) [3], [4] with the localization of the accident window. TAA has several historical frames up to time T as the observation and prefers an early accident warning for a time t by maximizing a Time-to-Accident (TTA) duration τ to the starting time of accident t_{ai}, as illustrated in Fig. 1.

No one road user wants to be involved in a traffic accident. This expectation of the whole human world forms the rarity of traffic accidents in terms of time duration and spatial locations. Usually, an accident occurs suddenly and may cover only about 10-30 video frames (0.5-1 seconds), where the road users (pedestrians, cyclists, cars, etc.) take up a small spatial ratio of the whole frame. Therefore, the normal background takes up a large part of video data, which implies a confounding issue [6] for the causal part of accidents. Further, current paradigms typically formulate TAA as a supervised learning framework with the given accident time window and occurrence annotation. This limits their practical generalization as there is limited annotated data. Besides, a huge part of TAA works [2], [7] concentrate on the object-centric (*detection→ tracking*) temporal consistency or spatial interaction framework for feature extraction. However, the complex traffic scenes make these frameworks inadequate for satisfactory TAA performance.

Received 2 November 2023; revised 3 May 2024 and 2 February 2025; accepted 10 March 2025. Date of publication 9 September 2025; date of current version 12 November 2025. This work was supported by NSFC under Grant W2411052 and Grant 62273057. The associate editor coordinating the review of this article and approving it for publication was Dr. Lin Yuanbo Wu. *(Corresponding authors: Jianwu Fang; Zhedong Zheng.)*

Jianwu Fang, Lei-Lei Li, and Jianru Xue are with the National Key Laboratory of Human-Machine Hybrid Augmented Intelligence, Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University, Xi'an 710049, China (e-mail: fangjianwu@mail.xjtu.edu.cn; 670160532lileilei@gmail.com; jrxue@mail.xjtu.edu.cn).

Zhedong Zheng is with the University of Macau, Macao 999078, China (e-mail: zhedongzheng@um.edu.mo).

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

Zhengguo Li is with the Institute for Infocomm Research, Agency for Science, Technology and Research (A*STAR), Singapore 627607 (e-mail: ezgli@i2r.a-star.edu.sg).

Tat-Seng Chua is with the Sea-NExT Joint Center, School of Computing, National University of Singapore, Singapore 119077 (e-mail: dcscts@nus.edu.sg).

Digital Object Identifier 10.1109/TMM.2025.3607808

1520-9210 © 2025 IEEE. All rights reserved, including rights for text and data mining, and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

Authorized licensed use limited to: Universidade de Macau. Downloaded on November 27,2026 at 15:11:58 UTC from IEEE Xplore. Restrictions apply.

Some attempts take the counterfactual sample synthesis to overcome the confounding issue in video understanding tasks, *e.g.*, masking [8], [9], or object/scene-level grounding [10], [11] to cut off the noisy context for task prediction. However, these formulations still rely on the pre-annotated positive/negative labels. Recently, video diffusion models aim to generate the expected video clips with the guidance of words with clear semantic descriptions [12], [13], [14]. This formulation implies a clear semantic alignment to activate the correlation between the key visual concept and given text prompts. By this feature, diffusion models can flexibly create positive and negative samples. Hence, for the TAA problem, we propose an Attentive Video Diffusion (AVD) model for ego-view accident video generation. The first aim of AVD is to tackle the confounding issue by generating and changing the causal part of the accident videos, i.e., interventing the anchor *normal clip* into *accident clip* explicitly. The second aim of AVD is to relax the demand for the annotation of accident video data.

To fulfill the above two aims, the AVD needs two kinds of abilities: **1)** the changing of the causal part of a video conditioned by text prompts with a well-aligned text-video semantic learning, and **2)** preservation of the content details and temporal dependence in the generated frames owning to the problem setting of TAA. To guarantee the first ability, we adopt the stable diffusion model [15] as the backbone and fine-tune our AVD on a large-scale multi-modality accident dataset [16] with over 10 k accident videos and well-aligned text-video annotation. A cross-attention model is designed in AVD to maintain the text-video semantic alignment. For the second ability, we add several frames in the original videos as the condition to the reverse diffusion process of AVD and formulate a spatial-temporal attention model to preserve the content consistency and temporal dependence in the generated video frames. This strategy also facilitates an *equivariant triple loss* to contrast the causal part learning by the triple clip set (***e*-triple$_c$-set**), i.e., *{anchor video clip, pseudo-normal clip, pseudo-accident clip}*. This triple set construction has a similar consideration in the adversarial perturbation of the object Re-ID task [17] or contrastive learning for the online video instance segmentation [18]. However, AVD promotes an Equivariant TAA (EQ-TAA), and the datasets for training EQ-TAA can be collected from arbitrary driving scenes without annotation, avoiding the laborious annotation for positive/negative samples of traditional TAAs. The **contributions** are threefold.

- We propose an attentive video diffusion model that can handle the confounding issue of background frames in TAA and facilitate an *equivariant triple loss* in TAA to enhance the causal part learning on the ***e*-triple$_c$-sets**.[1]
- To the best knowledge, we facilitate an Equivariant TAA (EQ-TAA) model that can be trained by the ego-view traffic videos collected from various driving scenes without any annotation for flexible generalization.
- We conduct extensive experiments on video diffusion and TAA, and show competitive performance to other state-of-the-art TAA methods. We evaluate the video diffusion performance on the HEV-I [5] and BDDA dataset [19] and traffic accident anticipation on DADA-2000 [20], CCD [21], and A3D [22] datasets.

The remainder of this work is as follows. Section II briefly reviews the related works. Section III and Section IV respectively describe the traditional supervised TAA formulation and the causal formulation for our equivariant TAA. The methodology is described in Section V. Section VI presents the experiments and the conclusion is given in Section VII.

II. Related Work

A. Traffic Accident Anticipation

Traffic accidents are a kind of scene anomaly [23] that require a clear definition of the occurrence of a collision. Traditional research pipelines in road accident anticipation primarily focus on trajectory or frame prediction models [24], [25], following a process involving detection, tracking, trajectory prediction, and accident or collision determination. Given the safety-critical nature of accidents, synthetic scenario generation also gathers significant interest in this field [26], [27]. Recently, TAA has entered the deep learning era, and the following two aspects are being featured in most formulations.

Object Feature Interaction in TAA: Some works in TAA explore the interactive relationships within participant groups to determine the temporal consistency and movement risk [2], [28]. Chan et al. [29] contribute the first accident anticipation model using Dynamic-Spatial-Attention (DSA) and Recurrent Neural Network (RNN) to correlate object interactions in driving videos. Inspired by the DSA module, Karim et al. [30] introduce a Dynamic Spatial-Temporal Attention (DSTA) network for traffic accident anticipation. Based on DSA, they model Dynamic Temporal Attention (DTA) to weigh the hidden states of past frames or object regions [7]. Furthermore, they propose a Spatial-Temporal Relational Learning (STRL) model [30] with Gated Recurrent Unit (GRU), which is inspired by the work [21] using the Graph Convolutional Recurrent Network (GCRN) [31] to model STRL. Similarly, Wang et al. [32] also employs the Graph Convolutional Network (GCN) to model object feature interactions in driving scenes. They focus on learning the spatial-temporal continuity for edge weight updating in accident anticipation. Recently, DeepAccident [33] investigated accident anticipation in vehicle-to-vehicle scenarios focusing on trajectory collision. Through these deep learning strategies, TAA research aims to improve early predictions of potential collisions and provide valuable insights for safe decision-making in advanced driving systems.

Frame Feature Consistency in TAA: Compared with modeling object-centric interactions in TAA, frame feature-consistency formulations relax the demand for accurate object detection, which eliminates possible detection errors in severe environments. Driver attention plays a crucial role in TAA since it reflects the driving task indirectly by indicating where drivers want to go or what they focus on. Frame feature consistency explores the hidden state consistency within the encoded frame features and pursues a core semantic learning for accident anticipation.

[1]Each triple clip set contains an anchor video clip, a generated accident clip, and an accident-free clip.

Authorized licensed use limited to: Universidade de Macau. Downloaded on November 27,2026 at 15:11:58 UTC from IEEE Xplore. Restrictions apply.

In this domain, some works take driver attention as an auxiliary clue to assist the causal part learning of accident videos because driver attention in TAA models can help to identify dangerous objects without necessarily relying on object detection. Previous works [20] in this area have explored the relationship between accident prediction and driver attention, yielding promising results. These understandings have inspired further development of TAA models; for example, *e.g.*,, Bao et al. [34] leverage the driver attention to assist the accident anticipation via a deep reinforced learning model. A recent work [35] further explores driver attention and textual description to facilitate TAA to form a cognitive TAA model.

The aforementioned TAA works have shown promising performance, but they all rely on the well-annotated dataset, which limits their practical generalization in different driving scenes. In this work, we explore the video diffusion model for flexible data training without annotations.

B. Video Diffusion Models

Most video diffusion models are inspired by the Denoising Diffusion Probabilistic Models (DDPM) model [36], which has a forward Markov process and a reverse process. *Forward process* corrupts the original sample x_0 by gradually introducing noise to x_k at k step until the sample becomes completely random noise, and the *reverse process* uses a series of Markov chains to remove the predicted Gaussian noise and recover the data progressively. With the diffusion operation, some label-efficient vision learning works [37] can be achieved by generating the pseudo-labels. Recently, video diffusion models have undergone the following two steps.

Condition Control in Diffusion: Based on DDPM, many conditional diffusion models emerge endlessly. The work [38] trains a classifier to control labels with the diffusion guidance of the gradient $\nabla_{x_k} \log p_\phi(y|x_k, k)$ towards any category label y. Visual conditions [39], [40] add images in the sampling process and guide the reconstructed images to be similar to the reference images. Similarly, Kirstain et al. [41] advocate the diffusion of various text-to-image scenarios that may benefit from visual cues, and some influencing works, *e.g.*, DALL-E2 [42] and Stable Diffusion [43], have set an impressive milestone. Moreover, ControlNet [44] presents general support for additional input conditions. However, the above tasks are done in a single-flow framework. Recently, Versatile Diffusion [45] expands the single-flow diffusion into a multi-task multimodal network that handles multiple flows of text-to-image, image-to-text, and their variations in a unified model.

Text-to-Video Generation: Text-to-video generation is more challenging than image generation [46] because it requires consistency in structure and spatiotemporal context. The pioneering work [47] presents the first results on video generation using diffusion models, which expands 2D U-net to 3D U-net. Then, Make-A-Video [48] leverages the help of many big diffusion models to synthesize videos conditioned on a given text prompt. Further, long video generation is explored by Yin et al. [49]. However, most diffusion models require iterative computation which is costly for both training and inference. Stable diffusion models [43], on the other hand, model the diffusion processes on the latent space, which can greatly reduce the computational complexity and generate images that can retain image diversity with high-quality visual semantics, which inspires the subsequent video diffusion works [12], [13].

With the video diffusion model conditioned by text prompts, we can generate diverse and large-scale accident or accident-free videos, and the generated videos can be treated as pseudo-positive or negative samples for the equivariant accident anticipation model.

C. Causal Inference in Video/Image Understanding

Causal inference recently is an attractive theory in image/video understanding for improving the model explainability and finding the important correlation between the causal variables to certain tasks, such as Video Question Answering (VQA) [9], [50], [51], [52], image recognition [8], video anomaly detection [53], and object grounding [10], image captioning [54], *etc.* Under the Structural Causal Model (SCM) [55], most video understanding works involved with causal inference theory aims to overcome the linguistic bias or visual content bias, through counterfactually masking the critical objects or words [9], or object/video level grounding [10], [52], i.e., cutting off the backdoor path from the data domain to the videos/images and specific sentences by causal intervention. Then, the causal object or scenes are learned by contrastive learning for the complementary samples (i.e., the original samples, and the intervened ones). However, these formulations still rely on the pre-annotated labels. As for this work, the EQ-TAA prefers a free-label accident anticipation, where the causal intervention is achieved through the Attentive Video Diffusion (AVD) to change the normal video clips to accident clips driven by accident text prompts.

III. Previous Supervised TAA

Here, we briefly recap the supervised TAA prototype [32], [34]. Given the video clip, TAA aims to maximize the Time-to-Accident (TTA) τ, where $\tau = \max\{0, t_{ai} - t\}$, t denotes the first timestamp where the predicted accident probability p_t is larger than a pre-defined threshold (*e.g.*, 0.5 commonly), and t_{ai} is the beginning time of the accident. In the supervised TAA mode, the accident and accident-free samples are paired by video frames and the accident labels y_t (1:accident, 0: accident-free). Consequently, given an observation video clip $V_{1:t} = \{I_1, \ldots, I_t\}$, the objective of TAA is to predict the accident score $\hat{p}_t \in [0, 1]$ and maximize τ:

$$\hat{p}_t = f_y(V_{1:t}), \tag{1}$$

where f_y is the combination of a video content encoder and the accident decoder. The video content encoder extracts the core semantics in the video frames up to time t that may correlate with the occurrence of an accident at time t_{ai}. The accident decoder bridges the encoded video content with the accident occurrence. Notably, a video often has the consistent accident label y_t for each frame. Therefore, the predicted $\hat{p}_t$ is treated as the predicted accident label $\hat{y}_t$ for a clear description.

Authorized licensed use limited to: Universidade de Macau. Downloaded on November 27,2026 at 15:11:58 UTC from IEEE Xplore. Restrictions apply.

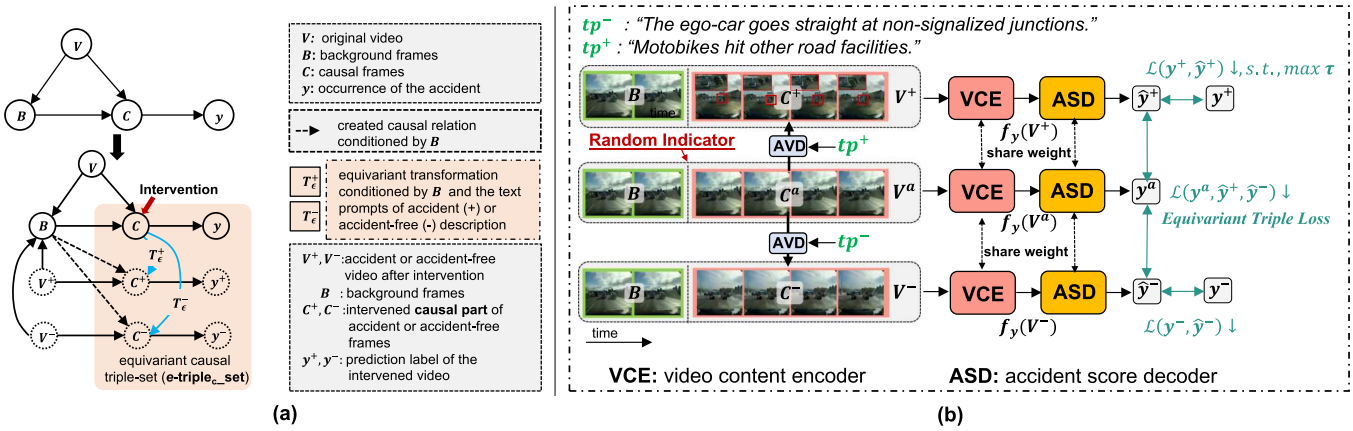


Fig. 2. The causal formulation of our Equivariant TAA, where (a) denotes the causal graph of EQ-TAA, and (b) is the training phase of EQ-TAA. "*tp*+" and "*tp*-" denote the accident and accident-free text prompts, respectively. $\hat{y}^+$ and $\hat{y}^-$ are the predicted accident labels of the generated accident sample and the accident-free sample, and y^a (accident-free commonly) is the accident label of the anchor video clip.

Learning: To optimize the video content encoder and accident decoder, current works commonly adopt the scheme of empirical risk minimization (ERM) [11] between the predicted accident label $\hat{y}_t$ with the ground truth y_t:

$$\min \mathcal{L}_{\text{ERM}}(\hat{y}_t, y_t),\ s.t., \max \tau. \tag{2}$$

Notably, the accident samples have an earliness penalty (maximizing τ) of accident occurrence, commonly modeled by an exponential function [56], while the accident-free samples usually formulate a Cross-Entropy (XE) to measure the label similarity. Essentially, previous supervised TAA works need ground-truth accident labels with laborious annotation, which is impractical in the open-traffic world.

IV. Causal Formulation for EQ-TAA

We argue for research towards the Equivariant TAA work in this field with a more interpretable causal formulation. Specifically, just like the equivariant video representation learning [57], we do not need to annotate the accident label for each video frame. Instead, we create the accident label and advocate an equivariant accident label prediction after we intervene and create the causal part of each video. In light of this, we formulate the TAA with Structural Causal Model (SCM) [55] by checking the causal relation of four variables: input video V, background environment clip B, causal scene clip C, and the created accident label y.

Causal Formulation for TAA: We take a casual look at TAA (i.e., Fig. 2(a)).

- $B \leftarrow V \rightarrow C$: Each accident video V can be divided into the set of background frames $\mathbf{B}$ and the set of accident frames $\mathbf{C}$, partitioned by a *grounding indicator*.
- $B \rightarrow C$: TAA implies the temporal correlation between the historical frames and causal frames. Therefore, B is a backdoor causal relation to C and $B \not\perp C$.
- $C \rightarrow y$: The prediction of accident occurrence y leverages the visual knowledge in causal part C.

The *grounding indicator* aims to separate the background frames and causal frames. Instead of finding the grounding indicator by well-designed models [11], [58], we can take an arbitrary video frame in the normal videos as the starting frame of the accident/accident-free clip, and create the accident/accident-free frames flexibly by the attentive video diffusion model (Section V-A) guided by the accident/accident-free text prompts. Therefore, the grounding indicator here is denoted as *Random Indicator*. Assuming that we can create a pair of an accident clip (with y^+) and an accident-free clip (with y^-) for each anchor video clip V^a, we explore the rationalization for causal-equivariance of TAA.

Rationalization: During training, the ground-truth grounding rationale C is out of reach, while only the annotation-free input (V^a) is observed. Fortunately, we can collect a vast amount of accident-free videos and video-independent text descriptions of various accident types based on commonsense (*e.g.*, positive text prompt: "*An ego-car hits a car*" and negative text prompt: "*A car moves straight on the road*", etc.). Such an "absence" of accident labels motivates TAA to embrace text-guided video generation to create the rationale C. Besides, we also advocate a smooth temporal evolution of video content between B and C. Hence, the text-guided video generation is also conditioned by some frames in the anchor clip V^a. Consequently, we approach a pair of created accident and accident-free causal scenes, i.e., C^+ and C^-, and then generate prediction y^+ and y^- via $y \leftarrow C$. To systematize this, an Attentive Video Diffusion (AVD) process T_ϵ is introduced based on the causal-equivariance principle to create C^+ and C^-, and expects an equivariant change in the response variable (i.e., y^+ and y^-). On top of SCM, we present such notions as:

$$T_\epsilon^+(y^+) = f_{y^+}(T_\epsilon^+(C^+)),\ T_\epsilon^-(y^-) = f_{y^-}(T_\epsilon^-(C^-)). \tag{3}$$

Extending ERM: With the flexible causal part and accident label creation, we are excited that we can extend ERM with an Equivariant Triple Loss (ETL) that can enforce the causal part learning, *i.e.:*

$$\min \mathcal{L}_{\text{ERM}}(\hat{y}_t^+, y_t^+) + \mathcal{L}_{\text{ERM}}(\hat{y}_t^-, y_t^-) + \mathcal{L}_{\text{ETL}}(y_t^a, \hat{y}_t^+, \hat{y}_t^-), \tag{4}$$

where y_t^a is the label of the anchor sample which is set as negative (-) for the easily collected accident-free videos. Besides, we can

Authorized licensed use limited to: Universidade de Macau. Downloaded on November 27,2026 at 15:11:58 UTC from IEEE Xplore. Restrictions apply.

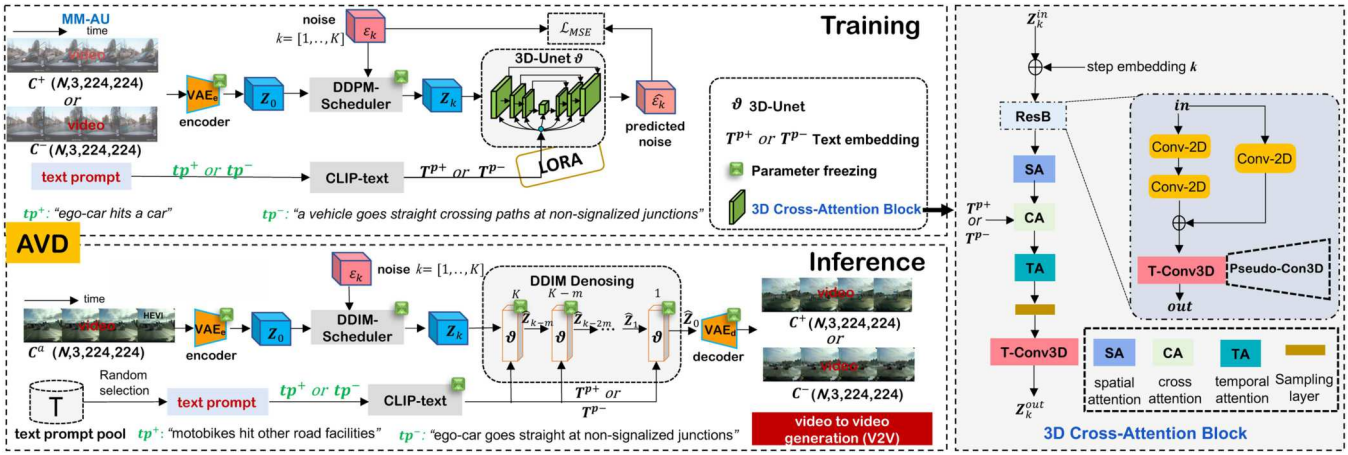


Fig. 3. The training and inference pipelines of AVD, where we utilize a Video-to-Video (V2V) generation framework for guaranteeing the temporal dependence among N generated video frames. In the AVD training phase, we achieve the text-video alignment learning in the **Forward Diffusion** process by inputting accident (C^+) or accident-free (C^-) samples in the MM-AU dataset [16]. In the AVD inference phase, we show the accident or accident-free video generation process through the **Reverse Diffusion** process (stepped by interval m, $m \gg 1$) inputted an anchor video clip (C^a) taken from arbitrary normal driving video datasets, such as HEV-I [5].

guarantee the sample balance for positive and negative samples all the time.

V. Methodology

We model the method flow for Equivariant TAA (EQ-TAA) in Fig. 2(b), where the timestamp of the Random Indicator is denoted as the beginning time of the accident. It contains three main components: 1) The Video Interventer by **AVD** partitioned by a Random Indicator, which creates the causal parts of the accident or accident-free samples, 2) the Video Content Encoding (**VCE**) and 3) the Accident Score Decoding (**ASD**). With this modeling, EQ-TAA is fulfilled with enforced causal part learning by the equivariant triple loss.

A. Video Interventer by AVD

Here, we describe the Attentive Video Diffusion (AVD) model for causal part intervention (i.e., creation). Fig. 3 shows the training and inference flowchart. The AVD model is inspired by the Stable diffusion [15], while we extend it into the video generation field with specially designed video encoding modules. In the training phase, AVD is trained by a large-scale multi-modality accident video dataset MM-AU [16], which contains 11,727 ego-view videos, 58 types of accident description, and 78 types of texts for describing normal traffic behaviors. These text prompts also constitute a text prompt pool $\mathcal{T}$. Each accident video is well-annotated for the accident and accident-free window. The text prompt pool $\mathcal{T}$ and the annotated accident videos are used for AVD Training. In the AVD Inference stage, we can randomly select a video clip in arbitrary normal ego-view driving videos, such as HEV-I [5], as the anchor clip C^a, and change its visual content to accident (C^+) or accident-free (C^-) conditioned by randomly selected accident (tp^+) or accident-free (tp^-) text prompts from $\mathcal{T}$.

1) AVD Training: We use MM-AU [16] with the well-annotated pairs of videos and text prompts (tp^+ for accident video and tp^- for accident-free video) are adopted to train AVD for text-video content alignment learning. Leveraging the Latent Diffusion Model (LDM) [15] for fine-tuning and text-to-video synthesis, our AVD also uses pre-trained LDM to map the videos into a latent space and model the diffusion process over the latent representation.

AVD consists of three main components: 1) an autoencoder (VAE$_e$) that is used to map a video clip with N frames into a latent representation $\mathbf{Z}_0$; 2) a DDPM-scheduler that aims to forwardly add the noise ε_k to $\mathbf{Z}_0$ and generate noised latent representation $\mathbf{Z}_k$ with k steps for adding noise, and 3) a 3D U-net ϑ which contains six 3D Cross-Attention Blocks (**3D-CAB**) (three encoder layers and three decoder layers) to estimate the noise $\hat{\varepsilon}_k$ and align the temporal text-video content correlation. AVD training aims to minimize the distance between $\hat{\varepsilon}_k$ and ε_k, where $k \in [1, \ldots, K]$ denotes the diffusion step.

LoRA Trick: In the training phase, the text prompt is encoded by a CLIP-text model [59], which is co-trained with the 3D U-net for self-style text-video content alignment learning. To preserve the knowledge in the latent representations and efficiently learn the alignment between videos and texts, we use the Low-Rank Adaptation (LoRA) [60] fine-tuning trick. It takes the entire 3D U-net and CLIP-text model into the LoRA box which freezes the pre-trained LDM and CLIP-text model encoders and greatly improves the training efficiency for the fine-tuning diffusion models on our datasets.

2) AVD Inference: Most modules in the AVD inference phase are similar to those of AVD training. Differently, the DDPM scheduler is switched into the Denosing Diffusion Implict Model (DDIM) scheduler [61] for a faster reverse diffusion process (i.e., changing the diffusion step interval 1 to m and $m = 50$ in this work), which leverages the trained 3D U-net and maintains better semantic consistency to the given video clips than DDPM. Similarly, the DDIM-scheduler also adds the noise ε_k and estimates the noised latent representation $\mathbf{Z}_k$. Differently, $\mathbf{Z}_k$ is denoised to the clean representation $\hat{\mathbf{Z}}_0$ by K/m steps DDIM denoising process, where each step of denoising is fulfilled by the trained 3D U-net ϑ. $\hat{\mathbf{Z}}_0$ is then decoded by VAE$_d$ to generate

Authorized licensed use limited to: Universidade de Macau. Downloaded on November 27,2026 at 15:11:58 UTC from IEEE Xplore. Restrictions apply.

new C^+ or C^- for the accident or accident-free clips. Notably, to preserve the temporal content dependency in the generated frames, we utilize a video-to-video generation pipeline, i.e., all N frames of input video clip C^a are fed into the AVD inference stage and have the same dimension as the generated ones. Here, the text prompt is randomly selected from the text prompt pool $\mathcal{T}$ and encoded by the same CLIP-text model trained in the AVD-training phase.

3) 3D-CAB: It is worth noting that 3D-CAB is involved in the AVD training and inference phases. It is specially designed for spatial-temporal feature learning in our AVD model. A clear structure of 3D-CAB is illustrated in the right part of Fig. 3, which consists of a ResNet Block (ResB), a text-video Cross-Attention (CA) layer, a Spatial Attention (SA) layer, and a Temporal Attention (TA) layer. The core layer is the ResB layer which has a spatial-temporal feature encoding layer (T-Conv3D). In the original 3D U-net [47], 3D convolution is adopted in T-Conv3D. However, it is time-consuming, and we instead introduce four layers of Pseudo-3D convolution [48] to encode the spatial-temporal representation which replaces the $(3 \times 3 \times 3)$ 3D convolution with a $(1 \times 3 \times 3)$ 2D convolution and a $(3 \times 1 \times 1)$ 1D temporal convolution as:

$$\text{Conv-P3D}(\mathbf{Z}_{in}) = \text{Conv-1D}(\text{Conv-2D}(\mathbf{Z}_{in})^T)^T. \quad (5)$$

Based on this strategy, we do not need to modify the 2D U-net shape and fulfill the 3D convolution effect.

SA: Spatial Attention (SA) here aims to learn the self-attention between the spatial tokens in $\mathbf{Z}_{in}^k$, where we denote the input of latent representation as $\mathbf{Z}_{in}^k$ and the output after self-attention as $\mathbf{Z}_{out}^k$ in a consistent denotation.

Given $\mathbf{Z}_{in}^k \in \mathbb{R}^{b\times N\times c\times H\times W}$ with the batch size b, N frames, and c channels of feature maps with size of $H \times W$, we reshape $\mathbf{Z}_{in}^k$ in spatial attention as $\bar{\mathbf{Z}}_{in}^k \in \mathbb{R}^{(b\times N)\times(H\times W)\times c}$, i.e., the token correlation in $b \times N$ frames is modeled over the $(H \times W) \times c$ channels. Then, $\bar{\mathbf{Z}}_{in}^k$ is encoded by a multi-head self-attention (*e.g.*, S heads) as:

$$\mathbf{Z}_{out}^k = \text{MHSA}^{(Spatial)}(\bar{\mathbf{Z}}_{in}^k) = \phi_{\mathbf{W}_s}([h_1; h_2; \dots h_S]), \quad (6)$$

where $\phi_{\mathbf{W}_s}$ is a linear transformation with parameters $\mathbf{W}_s, s \in [1, S]$, and

$$h_s = \text{SAtt}(\phi_{\mathbf{W}_{s\mathbf{q}}}(\mathbf{Z}_{in}^k), \phi_{\mathbf{W}_{s\mathbf{k}}}(\mathbf{Z}_{in}^k), \phi_{\mathbf{W}_{s\mathbf{v}}}(\mathbf{Z}_{in}^k)), \quad (7)$$

where $\phi_{\mathbf{W}_{s\mathbf{q}}}(\mathbf{Z}_{in}^k)$, $\phi_{\mathbf{W}_{s\mathbf{k}}}(\mathbf{Z}_{in}^k)$, and $\phi_{\mathbf{W}_{s\mathbf{v}}}(\mathbf{Z}_{in}^k)$ denote the linear transformations of the query ($\mathbf{k}$), key ($\mathbf{q}$), and value ($\mathbf{v}$) vectors of the s-th self-attention (SAtt) head, respectively.

SAtt is defined as:

$$\text{SAtt}(\mathbf{Z}(\mathbf{q}), \mathbf{Z}(\mathbf{k}), \mathbf{Z}(\mathbf{v})) = \delta(\mathbf{Z}(\mathbf{k})\mathbf{Z}(\mathbf{q})^T/\sqrt{d_k})\mathbf{Z}(\mathbf{v}), \quad (8)$$

where d_k is the head dimension of the query vector and $S = C/d_k$, and δ denotes the *softmax* function.

TA: Temporal Attention (TA) is identical to SA while differs in token correlation which is modeled over the frame dimension. Therefore, we reshape $\mathbf{Z}_{in}^k \in \mathbb{R}^{b\times N\times c\times H\times W}$ into $\bar{\mathbf{Z}}_{in}^k \in \mathbb{R}^{(b\times H\times W)\times N\times c}$, then the $\text{MHSA}^{(temporal)}(\bar{\mathbf{Z}}_{in}^k)$ operations are identical to (6) with the same heads.

CA: Text-Video Cross-Attention (CA) models the cross-modal attention [62] in text tokens $\mathbf{T}^p$ and video tokens $\mathbf{Z}_{in}^k$. To be computable, $\mathbf{T}^p \in \mathbb{R}^{(b\times N)\times l\times c_{text}}$ is followed by a linear layer to transform the feature channels c_{text} to c of $\mathbf{Z}_{in}^k$. Here, l denotes the text prompt length, which is set to 77in this work. The Cross-Attention (CA) is defined as:

$$\text{CA}(\mathbf{Z}(\mathbf{q}), \mathbf{T}^p(\mathbf{k}), \mathbf{T}^p(\mathbf{v})) = \delta(\mathbf{Z}(\mathbf{k})(\mathbf{T}^p(\mathbf{q}))^T/\sqrt{d_k})\mathbf{T}^p(\mathbf{v}). \quad (9)$$

Different from SA and TA, CA receives query ($\mathbf{q}$) from $\mathbf{Z}_k$, while key ($\mathbf{k}$) and value ($\mathbf{v}$) are from $\mathbf{T}^p$. The clear implementation of AVD is illustrated in Algorithm 1, where the blue texts denote the parts with fine-tuning.

Algorithm 1: AVD Algorithm.

Input:
- tp^+/tp^- : text prompt for C^+ or C^-.

Hyperparameters:
- K: The number of diffusion steps;
- β_k:The variance across diffusion steps.

Forward (AVD Training):
▷ Get the text embedding of tp^+ or tp^-.
$\mathbf{T}^{p+}$ or $\mathbf{T}^{p-} = \text{CLIP}_{\text{text}}(tp^{+/-})$
▷ Get the representation of the noise latent $\mathbf{Z}_t$.
-$\mathbf{Z}_0 = \text{VAE}_e(C^{+/-})$;
-$\mathbf{Z}_k$ =DDPM Schedule$(\mathbf{Z}_0, k, \varepsilon_k)$
$\mathbf{Z}_k = \sqrt{\hat{\beta}_k}\mathbf{Z}_0 + \sqrt{1-\hat{\beta}_k}\varepsilon_k$,
where $\forall k = \{1, 2, 3, \dots, K\} \in$ Uniform.
$\hat{\beta}_k = \prod_{i=1}^{k} \alpha_i; \alpha_k = 1 - \beta_k; \varepsilon_k \in \mathcal{N}(0, I)$
▷ **Output**:
$\hat{\varepsilon_k} = \vartheta(\mathbf{Z}_k, k, \mathbf{T}^{p+/-})$
▷ Optimization
$\min \text{VAE}_{\varepsilon\sim\mathcal{N}(0,I)\ k\sim\text{U}(1,K)} \|\hat{\varepsilon_k} - \varepsilon_k\|_2^2$

Reverse (AVD Inference):
Input: The anchor video clip C^a.
▷ Get the representation feature of the accident text or accident-free prompt $\mathbf{T}^{p+}$ or $\mathbf{T}^{p-}$.
$\mathbf{T}^{p+} or \mathbf{T}^{p-} = \text{CLIP}_{text}(tp^{+/-})$,
▷ Denoising the latent token by video-to-video generation.
-$\mathbf{Z}_0 = \text{VAE}_e(C^a)$,
-$\mathbf{Z}_k$ =DDIM Schedule$(\mathbf{Z}_0, k, \varepsilon_k)$,
$\mathbf{Z}_k = \sqrt{\alpha_k}\mathbf{Z}_0 + \sqrt{1-\alpha_k}\varepsilon_k$,
For $k = K, K-m, K-2m, \dots, 1$ **do**
$\varepsilon_k \leftarrow \vartheta(\hat{\mathbf{Z}}_k, k, \mathbf{T}^{p+/-})$, $\hat{\mathbf{Z}}_k = \mathbf{Z}_k$
$\hat{\mathbf{Z}}_{k-1} = \sqrt{\alpha_{k-1}}\frac{\hat{\mathbf{Z}}_k - \sqrt{1-\alpha_k}\varepsilon_k}{\sqrt{\alpha_k}} + \sqrt{1-\alpha_{k-1}}\varepsilon_k$
Output: $C^{+/-} = \text{VAE}_d(\hat{\mathbf{Z}}_0)$

B. Equivariant TAA

As shown in Fig. 2(b), after completing the AVD training phase, we can then freely generate a vast number of accident V^+ or accident-free V^- video clips respectively after grafting N generated frames. Given the triple set of (V^a, V^+, V^-), we have three branches to encode the video content in them and predict the accident probability y_t^a, $\hat{y}_t^+$, and $\hat{y}_t^-$, respectively. To learn the core semantics of object regions and sudden movement for accident anticipation, we consider the uncertainty issue because

Authorized licensed use limited to: Universidade de Macau. Downloaded on November 27,2026 at 15:11:58 UTC from IEEE Xplore. Restrictions apply.

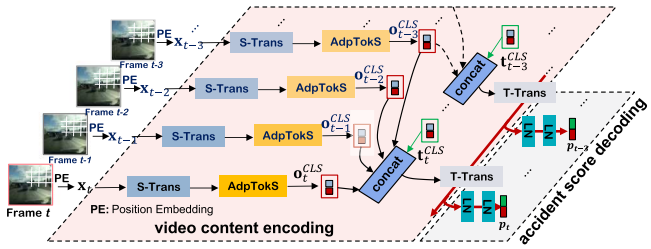


Fig. 4. The video content encoding and accident score decoding modules, where **S-Trans** and **T-Trans** mean the spatial transformer and temporal transformer, and **AdpTokS** denotes an adaptive token sampling module for scene uncertainty consideration.

sudden scene changes caused by critical objects may have direct links to future accidents.

1) Uncertainty-Aware Video Content Encoding (VCE): Fig. 4 illustrates the flowchart of VCE, which contains a Spatial Transformer (S-Trans) block, an Adaptive Token Sampling (AdpTokS) block, and the Temporal Transformer (T-Trans) across frames of the token embeddings. Because it has the same structure as VCE in the triple set, we omit the "a", "+" or "-" index and denote $V_{in} = \{v_t\}_{t=1}^{n}$ for the given input with n frames (set as 5in this work) in each encoding process.

S-Trans: We first encode each frame v_t in V_{in} to spatial tokens $\mathbf{x}_t \in \mathbb{R}^{n\times Tok\times d}(Tok = \frac{H}{P}\times\frac{W}{P} = 196$ for 16 patch size P) with a linear projection $((3\times P\times P)\rightarrow d)$, where we set $d = 192$. Then, we add the Position Embedding (PE) $\mathbf{x}_{PE} \in \mathbb{R}^{b\times n\times(Tok+1)\times d}$ and concatenate spatial cls-token $\mathbf{x}_{cls} \in \mathbb{R}^{b\times n\times 1\times d}$ to get new $\mathbf{x}_t \in \mathbb{R}^{b\times n\times(Tok+1)\times d}$. Then we rearrange $\mathbf{x}_t$ to $\hat{\mathbf{x}}_t \in \mathbb{R}^{(b\times n)\times(Tok+1)\times d}$ for recurrently inputting to S-Trans block (H_s-layers). The operation of S-Trans is the same as the Spatial Attention (SA), i.e.,

$$\mathbf{x}_t^o = \text{S-Trans}(\mathbf{x}_t) = \text{MHSA}_{H_s}^{(spatial)}(\mathbf{x}_t). \tag{10}$$

AdpTokS: AdpTokS aims to select the important tokens with core object region semantics representation, which is formulated to consider the uncertainty of token learning. The token selection is based on the token attention ranking steps. Given $\mathbf{x}_t^o$, it is filtered through a linear normalization. Then, the attention ranking leverages the self-attention of queries ($\mathbf{q}$) and keys ($\mathbf{k}$) scaled by d_0 at each attention head as:

$$\mathbf{A}_t = \delta(\mathbf{x}_t^o(\mathbf{q})\mathbf{x}_t^o(\mathbf{k})^T/\sqrt{d_0}). \tag{11}$$

The row summation of attention matrix $\mathbf{A}_t \in \mathbb{R}^{(Tok+1)\times(Tok+1)}$ equals 1, which represents the contribution of all input tokens to the output token. With $\mathbf{A}_t$, the significance score of the ith token in each attention head at time t can be calculated by:

$$a_t^i = \frac{\mathbf{A}_t^{1,i}\times||\mathbf{x}_t^o(\mathbf{v}_i)||}{\sum_{i=2}^{Tok}\mathbf{A}_t^{1,i}\times||\mathbf{x}_t^o(\mathbf{v}_i)||}, \tag{12}$$

where the token weight summation omits the "cls-token" in the first column of $\mathbf{x}_t^o$. The re-sampling of a token is based on its significance score summation over 8 heads, i.e., $e_i = \sum_{j=1}^{8} a_t^{j,i}$, where j is the head index which is fixed as 8 and differs from the dynamic head index s in (6).

Here, we do not adopt the cumulative token significance summation in [63] for a convenient token sampling. To be adaptive for scene learning, similar to [63], we do not discard the tokens with lower e_i, while we only prune the attention weight $\mathbf{A}_t$ to $\mathbf{A}_t'$ with $Tok'(Tok' < Tok)$ tokens. Then, the significant token representation $\mathbf{o}_t$ is obtained by $\mathbf{A}_t'\mathbf{x}_t^o(\mathbf{v})$. In this work, we have 3 layers of AdpTokS to recurrently enforce the important token selection.

T-Trans: To model temporal information, we extract the "cls-token" for each time t and concatenate with the time cls-token $\mathbf{t}_t^{CLS}$ to obtain the concatenated temporal video content representation $\mathbf{O}_t$. $\mathbf{O}_t$ is then fed into T-Trans (H_t-layers) for learning temporal relationships. The operation of T-Trans is the same as the MHSA$^{(temporal)}$. Finally, we use the max-pooling operation to obtain the salient token feature $\hat{\mathbf{o}}_t^{\kappa}$ at time t as:

$$\begin{aligned}\mathbf{O}_t &= \text{Concat}(\mathbf{o}_t^{CLS}, \mathbf{o}_{t-1}^{CLS}, \ldots, \mathbf{o}_{t-n}^{CLS}, \mathbf{t}_t^{CLS}),\\ \hat{\mathbf{o}}_t &= \text{T-Trans}(\mathbf{O}_t) = \text{MHSA}_{H_t}^{(temporal)}(\mathbf{O}_t),\\ \hat{\mathbf{o}}_t^{\kappa} &= \text{maxpooling}(\hat{\mathbf{o}}_t).\end{aligned} \tag{13}$$

After the video content encoding, $\hat{\mathbf{o}}_t^{\kappa}$ is decoded to obtain the accident score at time t.

2) Accident Score Decoding (ASD): The accident score $\hat{p}_t$ is computed by decoding $\hat{\mathbf{o}}_t^{\kappa}$ with two fully connected layers (fc) with the parameters of θ_1 and θ_2 and followed a softmax function, defined as:

$$\hat{p}_t = \delta(fc(fc(\hat{\mathbf{o}}_t^{\kappa};\theta_1);\theta_2)). \tag{14}$$

Notably: To be consistent with 1, the accident score $\hat{p}_t \in [0,1]$ can be denoted as the predicted accident label $\hat{y}_t$.

C. Optimization

We re-formulate (4) for video content learning in EQ-TAA.

ERM Loss: The ERM loss reflects the penalty for the prediction similarity of an accident or accident-free labels at each time. In particular, the accident video clip sample pursues an earliness of the accident. Therefore, the ERM loss is defined as a Cross-Entropy loss (XE) for accident-free samples and an exponential function loss [56] for accident samples:

$$\begin{aligned}\mathcal{L}_{\text{ERM}} &= w\mathcal{L}_{\text{ERM}}(y^-,\hat{y}^-) + \mathcal{L}_{\text{ERM}}(y^+,\hat{y}^+)\\ &= -\left[w\frac{1}{N_I}\sum_{t=1}^{N_I}(1-y^-)\log(1-\hat{y}_t^-)\right.\\ &\quad\left.+\frac{1}{N_I}\sum_{t=1}^{N_I}y^+e^{-\max(0,\frac{\tau}{r})}\log(\hat{y}_t^+)\right],\end{aligned} \tag{15}$$

where w is the weight of negative samples, and we set w to 0.7. N_I is the video sample length, τ is the time interval between current time t to the beginning accident time t_{ai}, and r is the frame rate.

Equivariant Triple Loss: In addition to ERM loss, we are excited that with the AVD model, we can generate a sample pair (an accident sample and an accident-free sample) for the anchor sample. Different from previous contrastive loss [64], [65], we design an Equivariant Triple Loss (ETL) to distinguish the feature representation between $\hat{\mathbf{o}}_t^{\kappa+}$ and $\hat{\mathbf{o}}_t^{\kappa-}$. Different from

Authorized licensed use limited to: Universidade de Macau. Downloaded on November 27,2026 at 15:11:58 UTC from IEEE Xplore. Restrictions apply.

ERM loss, we enforce ETL loss in the causal frames:

$$\mathcal{L}_{\text{ETL}} = \frac{1}{|t_{co} - t_{ai}|} \sum_{t=t_{ai}}^{t_{co}} [||\hat{\mathbf{o}}_t^{\kappa_a}, \hat{\mathbf{o}}_t^{\kappa_-}||^2 - ||\hat{\mathbf{o}}_t^{\kappa_a}, \hat{\mathbf{o}}_t^{\kappa_+}||^2 + \xi], \tag{16}$$

where t_{co} is the collision time of an accident, and the parameter ξ is used to ensure the discriminative difference between $\hat{\mathbf{o}}_t^{\kappa_-}$ and $\hat{\mathbf{o}}_t^{\kappa_+}$ with a pre-defined threshold (set as 0.5 commonly). $\hat{\mathbf{o}}_t^{\kappa_a}$ denotes the token feature of anchor sample in each triple set. Thus, the optimization is achieved by minimizing:

$$\mathcal{L} = \mathcal{L}_{\text{ERM}[1,N_I]} + \lambda \mathcal{L}_{\text{ETL}[t_{ai},t_{co}]}, \tag{17}$$

where λ is a parameter for balancing the weight of two kinds of loss. We find that λ is important with experimentally setting.

VI. Experiments

A. Datasets

We extensively evaluate our Attentive Video Diffusion (AVD) model and the Equivariant Traffic Accident Anticipation (EQ-TAA) to verify the performance.

1) For AVD: As aforementioned, the AVD model is trained on a large-scale multimodality accident dataset (**MM-AU** [16]) in driving scenes, where each video in MM-AU annotates the accident window [t_{ai}, t_{ae}]. To construct the text-video alignment pairs for training AVD, from MM-AU, we randomly select 5,842 accident-free frame clips in [1,t_{ai}] and 7,419 accident frame clips belonging to [t_{ai}, t_{ae}], resulting in a total of 13,271 text-video clips. Each clip consists of $N = 22$ frames with a corresponding text prompt.

After AVD training, we can generate any number of accident or accident-free video clips by the videos in various driving scenes. In this work, we take the **HEV-I** [5] and **BDD-A** [19] datasets as the accident-free database and utilize the trained AVD to conduct a random sampling of 1,800 triple clip sets (**e-triple$_c$-sets**) from them to train the following EQ-TAA [2]. **HEV-I** [5] contains 230 videos, and the length of each video ranges from 10 to 60 seconds. **BDD-A** [19] contains 1,232 videos (=3.5 hours) collected from critical but accident-free scenarios in different occasions and light conditions.

For the training EQ-TAA, each video clip consists of N_I =150 frames (30fps), where the *Random Indicator* (i.e., pseudo t_{ai}) is randomly determined in the frame range of [110,128] of each video clip, which results in pairs of accident and accident-free video clips with 22 synthetic frames starting from the pseudo t_{ai}. Notably, the Random Indicator can also be determined by an arbitrary video frame in HEV-I or BDD-A first, and then the associated video clip can be determined.

2) For EQ-TAA: Our method is tested on the **DADA-2000** [20], **A3D** [22], and **CCD** [28] datasets.

DADA-2000 has 2,000 accident videos, where for a fair comparison, same to DRIVE [34], 1000 ones are selected for training and testing, where we sample 1800 pairs of positive and negative clips from the training set, and all the samples in DADA-2000 have 150 frames (5 seconds).

[2]The code will be available at https://github.com/JWFanggit/Diffusion-TAA.

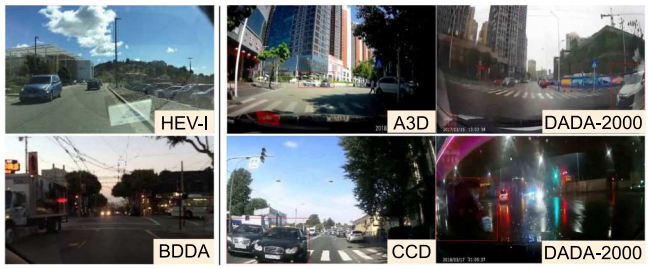


Fig. 5. The frame snapshots of HEV-I [5], BDD-A [19], A3D [22], CCD [28], and DADA-2000 [20].

A3D contains 1,500 dashcam videos. Owing to the different temporal locations for the traffic accident in different videos, we trimmed each A3D video to 150 frames (5 seconds). Similar to the GSC [32] work, we select 30% of the A3D dataset as the testing set.

CCD contains 1,500 dashcam accident videos. Each video is trimmed into 50 frames with 10 fps. Each accident is placed in the last 2 seconds of each accident video. Similar to the setting of CCD, 900 accident videos are chosen for testing.

Notably, as shown in Fig. 5, most videos of A3D and CCD are collected in sunny daytime but DADA-2000 is more challenging with more diversified weather and light conditions. Therefore, we take the training set of DADA-2000 to train models, and A3D and CCD are only taken for testing.

B. Metrics

This work adopts two groups of metrics for evaluation.

1) For AVD, we use the *Fréchet Inception Distance (FID)* [66], [67], *Fréchet Video Distance (FVD)* [68], and the *Inception Score (IS)* [69] as a quality evaluation for the synthetic frames and video clips, respectively. FVD uses a pre-trained I3D Network [70] to calculate the mean vector and covariance matrix of the real video clip C^a (22 frames) and the generated C^+ or C^-. FID uses a pre-trained InceptionV3 Network [71] to calculate the mean vector and covariance matrix of each frame. IS also takes the pre-trained InceptionV3 Network to calculate the exponential value of Kullback–Leibler divergence (KLD) between the generated images and real images. In addition, IS also reflects the distribution diversity of the generated images. Furthermore, we introduce the Contrastive Language-Image Pre-Training score (CLIP$_S$) to measure the alignment degree between text prompts and video frames. Larger CLIP$_S$ values indicate better coherent alignment. The CLIP$_S$ is defined as:

$$\text{CLIP}_S = \frac{1}{N} \sum_{i=1}^{N} \text{sim}(f_\theta(I_i), f_\theta(tp)), \tag{18}$$

where N represents the number of video frames, sim(.,.) represents the cosine similarity function, f_θ denotes the feature extractor function defined by the CLIP model [59]. $I_i \in C^{+/-}$ is the ith video frame, and tp specifies the text prompt.

2) For TAA, we employ the same metrics used in previous works [21], [34], i.e., Average Precision (AP) [21], and Time-to-Accident (TTA) [34], which includes the variants of TTA$_a$ and mTTA. AP evaluates the average accuracy with

Authorized licensed use limited to: Universidade de Macau. Downloaded on November 27,2026 at 15:11:58 UTC from IEEE Xplore. Restrictions apply.

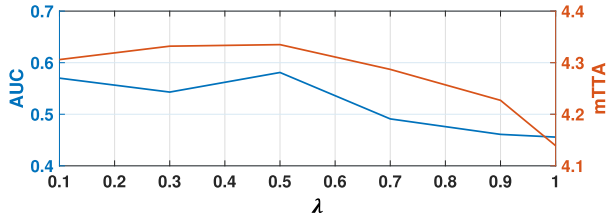


Fig. 6. The AUC and mTTA values, *w.r.t.*, λ, on the testing set of DADA-2000.

different thresholds of precision and recall rate. TTA_a is computed by setting the accident score threshold in each video as a to evaluate the *earliness* of a positive prediction, which is determined by the time interval between *the first warning frame* (with the accident score larger than a) and the starting time of the accident (t_{ai}). mTTA denotes the mean TTA, which is computed by the Expectation of $\text{TTA}_a, a \in [0, 1]$. In this work, we also evaluate the methods based on video-level Area under ROC curve (AUC) [34] for DADA-2000. AUC is a metric for evaluating the performance of accident or accident-free determinations. These metrics pursue a larger value.

In addition, TAA is a temporal forecasting problem and owns a prediction uncertainty issue. In this work, we further introduce a prediction Variance (**Var**) when the accident scores of frames are larger than a pre-defined threshold (set as 0.5 in this work), and stable anticipation (low Var values) is preferred for reducing false alarms in driving.

C. Implementation Details

In this work, the ETL loss in (17) is the key component for the TAA equivariance of this work, where the parameter λ acts as a trade-off between the ETL loss and ERM loss. Therefore, we experimentally check the role of λ by setting it as [0.1,0.3,0.5,0.7,0.9]. Here, we evaluate the performance difference on the testing set of the DADA-2000 with the AUC and mTTA metrics. Fig. 6 shows that $\lambda = 0.5$ obtains the best values and is adopted for more consideration of ETL loss.

Configuations for AVD: For training AVD, we set the learning rate as $5e^{-6}$, and the batch size is set as 2. Xformers [72] is utilized to optimize hardware memory efficiency. The number of training steps K is 8000. During the AVD inference, we generate videos with the same configuration, while the inference step interval m is set as 50.

Configuations for EQ-TAA: The layers of S-Trans (H_s) and T-Trans (H_t) are set as 2 and 3, respectively. The dimensions of two fully connected layers (fc) are transformed as 192→ 64 →2. The ETL loss parameter ξ is set to 0.5. We use the AdamW optimizer with a learning rate of $1e^{-4}$, β_1 of 0.9, β_2 of 0.999, and weight decay of 0.01. Because of the large-scale samples, all experiments in EQ-TAA are trained for one epoch with a batch size of 1.

All experiments are performed on a platform with one Nvidia RTX 3090Ti GPU and 64 G RAM.

D. Evaluation for AVD

In this work, we select four state-of-the-art video diffusion models including Tune-A-Video (*abbrev.* TAV) [15], Ho et al. [47], ControlVideo (Cvideo) [73], and Text2Video-Zero (T2V-Zero) [74] to compare with our AVD. To be fair, we also take the anchor video frames as the condition for these diffusion models. TAV takes the DDIM inversion to generate video frames. For each video diffusion model, we generate 1,800 clips, and each contains 22 frames. Cvideo and T2V-Zero are the training-free video diffusion methods. In addition, we also evaluate the roles of different self-attention models in our AVD, which results in two versions of AVD, namely "*AVD w/o SA*" (without $\text{MHSA}^{(Spatial)}$ defined in (6)) and "*AVD w/o TA*" (without $\text{MHSA}^{(temporal)}$).

TABLE I
PERFORMANCE COMPARISON FOR VIDEO DIFFUSION MODELS WITH 1800 GENERATED CLIPS SAMPLED FROM **HEV-I** [5] AND **BDDA** [19]

Method	CLIP_S↑	FID↓	FVD↓	IS↑	FPS↑
Ho et al. [47]	21.77	2098.65	9202.25	3.75	1.0
TAV [15]	24.80	854.05	8810.35	3.02	2.7
CVideo [73]	22.06	928.12	10551.49	4.11	0.5
T2V-Zero [74]	26.22	**585.73**	9842.69	**4.33**	1.1
AVD w/o TA	28.95	806.88	9651.40	2.01	4.4
AVD w/o SA	21.82	917.55	10648.71	3.13	4.4
AVD	**29.63**	673.35	**7924.75**	2.36	4.4

Table I presents the quantitative metric results, from which we can observe that our AVD can obtain significant improvement from generation quality (FVD), text-video alignment, and computation efficiency. We can see that the $\text{MHSA}^{(Spatial)}$ offers stronger supports to AVD model. We present some snapshots of generated video frames in Fig. 7. It is clear that the $\text{MHSA}^{(Spatial)}$ can enforce the textual details of generated frames. Without $\text{MHSA}^{(Spatial)}$, the generated frames show an unclear appearance. Without $\text{MHSA}^{(temporal)}$, many artifacts appear. For instance, after adding the accident text prompt (tp^+) of "*The ego-car hits a pedestrian*"), The version of "AVD w/o SA" is unresponsive for the crossing pedestrian. "AVD w/o TA" cannot generate a vehicle when adding the accident-free text prompt (tp^-) "*The ego-car turns on a road with the same direction to another vehicle*."

Notably, the IS values of our AVD are smaller than other models. However, in the visual comparison in Fig. 7, our AVD is better at preserving the content style of the anchor video frames than the Cvideo, TAV, T2V-Zero, and Ho et al.'s approach. AVD also enforces a responsive change of the specific object reflected by the largest CLIP_S value. Although the anchor frames are taken as the input condition, the methods of Ho et al., T2V-Zero, and Cvideo tend to output irrelevant frame content while TAV generates many artifacts. Although T2V-Zero obtains the best FID value, it has no relation to the anchor video frames and outputs irrelevant visual content, as shown in Fig. 7. Therefore, the IS metric is easily influenced by the noisy video content.

To further check the visual comparison of different video diffusion models, we compare Tune-A-Video (TAV), and Text2Video-Zero (T2V-Zero) in Fig. 8. The results show that T2V-Zero generates the video frame with the highest quality but it does not preserve the anchor video frame style and the text prompts are not responsive in the generated frames. TAV again

Authorized licensed use limited to: Universidade de Macau. Downloaded on November 27,2026 at 15:11:58 UTC from IEEE Xplore. Restrictions apply.

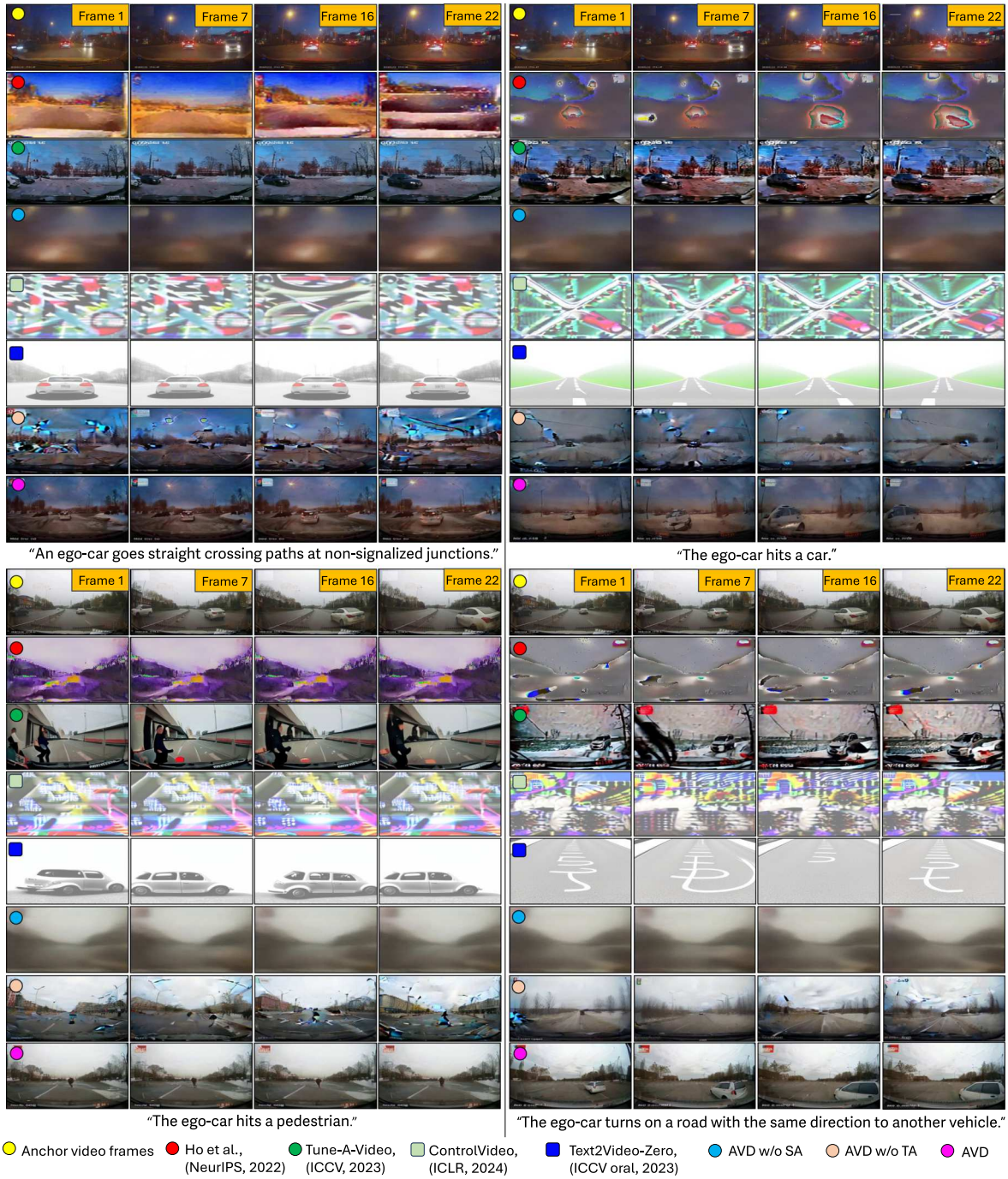


Fig. 7. The snapshots of different video diffusion models on the two videos of BDD-A dataset [19], where the results in each block are obtained by the method of Ho et al. [47], Tune-A-Video [15], ControlVideo [73], Text2Video-Zero [74], our AVD and its variants (Best viewed in color mode).

shows many artifacts and the generated video content has a large distance to the anchor video frames.

E. Overall Evaluation for EQ-TAA

1) Competitors: Since this work contributes an Equivariant TAA (EQ-TAA) method, we compare its performance with several state-of-the-art supervised and Equivariant TAA (EQ-TAA) methods driven by Tune-A-Video (TAV) [15], Ho et al. [47], ControlVideo (Cvideo) [73], and Text2Video-Zero (T2V-Zero) [74]. Based on the main formulation, we categorize the supervised approaches to **Object-centric TAAs (Ob.TAAs)**, and **Frame-centric TAAs (Fr. TAAs)**. Object-centric TAAs are as follows. To be fair, we try my best to balance the training samples in supervised approaches with real positive and negative video clips and our EQ-TAA driven by generated video clips (∼1800 pairs of negative and positive clips for training all methods).

DSA [29] is the first deep learning-based TAA work in dash-cam videos, which model a Dynamic Spatial Attention (DSA)

Authorized licensed use limited to: Universidade de Macau. Downloaded on November 27,2026 at 15:11:58 UTC from IEEE Xplore. Restrictions apply.

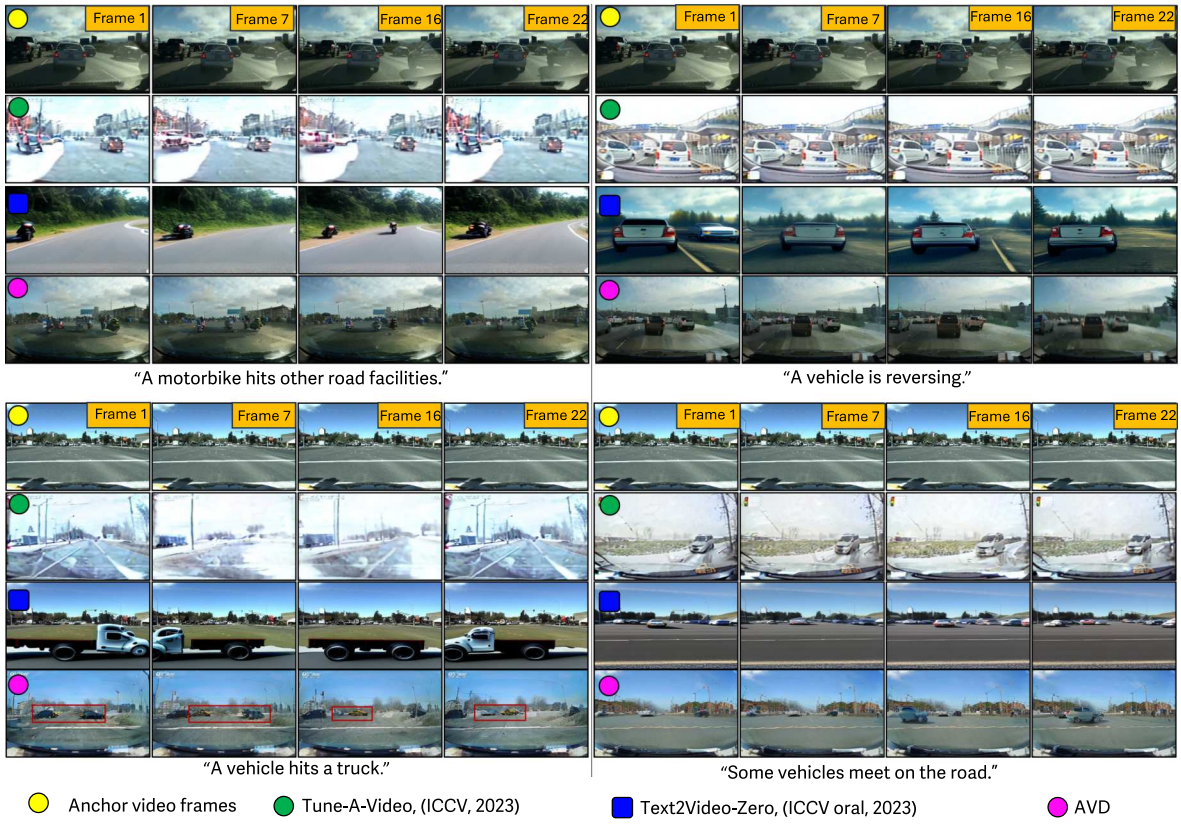


Fig. 8. The snapshots of different video diffusion models on the two videos of BDD-A dataset [19], where the results in each block are obtained by the method of Tune-A-Video [15], Text2Video-Zero [74], and our AVD (Best viewed in color mode).

incorporating the Recurrent Neural Network (RNN) to correlate the spatial-temporal correlation of objects in frames.

UncertaintyTA [21] models a graph-based spatial-temporal relation learning model, and considers the uncertainty estimation for the predicted accident scores.

DSTA [30] formulates a dynamic spatial-temporal attention model to infer the correlation of object interaction between video frames, and **GSC** [32] models a spatio-temporal graph continuity to infer the temporal consistency of video frames.

In addition to object-centric TAAs, three frame-centric TAAs, i.e., **DRIVE** [34], **XAI-AN** [75], and **Cog-TAA** [35], are selected. Different from the object-centric approaches which need to pre-detect the object accurately, frame-centric TAAs formulate the accident anticipation by spatial-temporal frame feature encoding and accident score decoding frameworks. DRIVE [34] involves the driver fixation map specially to assist the core semantic region learning, and Cog-TAA [35] introduces the text description as input and takes the fixation attention to guide critical object learning for cognitive TAA.

2) Result Analysis: For the DADA-2000 dataset, the metrics of AP, AUC, $\text{TTA}_{0.5}$, and mTTA are all used, while the AUC metric is omitted for the CCD and A3D datasets because the testing samples in them are all positive accident videos. Because we adopt the video clips with synthetic video frames to train the model, we make an equivalent training data scale to the training sets of the A3D, CCD, and DADA-2000 datasets. We use the results reported in other works on A3D and CCD datasets for comparison. For the DADA-2000 dataset, the results of other methods are reported in DRIVE [34] and Cog-TAA [35] with the equivalent training samples.

TABLE II
THE PERFORMANCE EVALUATION FOR DIFFERENT TAA CATEGORIES (TAA. CAT) IN THE TESTING SET OF THE A3D [22] AND CCD [28] DATASETS, WHERE THE BEST VALUE OF EACH METRIC IS MARKED IN **BOLD** FONT

TAA. Cat.	Method	A3D [22]		CCD [28]	
		AP↑	mTTA↑	AP↑	mTTA↑
Ob.TAAs	UncertaintyTA [21]	0.944	**4.920**	0.995	4.740
	DSA [29]	0.934	4.410	0.996	4.520
	DSTA [75]	-	-	0.996	**4.870**
	GSC [32]	-	-	0.949	2.620
Fr. TAAs	XAI-AN [76]	-	-	0.940	4.570
	DRIVE [34]	0.918	4.045	0.992	4.490
	Cog-TAA [35]	0.920	4.128	0.997	4.630
EQ-TAA	Driven by Ho *et al.* [47]	0.907	3.865	0.985	4.108
	Driven by TAV [15]	0.905	4.005	0.986	4.301
	Driven by CVideo [73]	0.903	3.626	**0.999**	2.814
	Driven by T2V-Zero [74]	0.906	3.536	**0.999**	2.685
	Driven by **AVD**	**0.953**	4.321	0.989	4.407

As shown in Tables II and III, compared with the supervised ones, the equivariant methods still have a distance towards achieving the best performance obtained by Cog-TAA,

Authorized licensed use limited to: Universidade de Macau. Downloaded on November 27,2026 at 15:11:58 UTC from IEEE Xplore. Restrictions apply.

TABLE III
THE PERFORMANCE EVALUATION FOR DIFFERENT TAA CATEGORIES (TAA. CAT) IN THE TESTING SET OF THE DADA-2000 [20] DATASET, WHERE THE BEST VALUE IS MARKED IN **BOLD** FONT

TAA. Cat.	Method	DADA-2000 [20]			
		AP↑	AUC↑	$TTA_{0.5}$ ↑	mTTA↑
Ob.TAAs	UncertaintyTA [21]	-	0.600	3.849	-
	AdaLEA [56]	-	0.550	3.890	-
	DSA [29]	-	0.470	3.095	-
Fr. TAAs	DRIVE [34]	0.690	0.727	3.657	4.259
	Cog-TAA [35]	0.744	**0.833**	**4.020**	**4.497**
EQ-TAAs	Driven by Ho *et al.* [47]	0.722	0.500	3.724	4.159
	Driven by TAV [15]	**0.752**	0.532	3.668	4.286
	Driven by CVideo [73]	0.716	0.448	1.462	3.546
	Driven by T2V-Zero [74]	0.698	0.421	1.023	3.481
	Driven by **AVD**	0.725	0.580	3.894	4.258

especially for the AUC metric as shown in Table III. It indicates that because of the synthetic nature, the generated accident videos and accident-free videos still need to be enhanced from their inter-class difference with more well-designed text-video coherent alignment models.

EQ-TAA (AVD) vs. other EQ-TAAs: We compare the roles of different video diffusion models for TAA here. Actually, for the TAA task, the label of each frame is influenced by the content difference between the historical frame observation and the accident frames. From Fig. 7, we can see that although TAV [15] generates many artifacts, it shows a significant difference between the generated ones and the anchor video frames. Therefore, it obtains the best AP value and mTTA value for the DADA-2000 dataset, as shown in Table III. Besides, the generated video frames need to reflect the causal objects well. Therefore, although CVideo [73] and T2V-Zero [74] generate fully different video content to the anchor video frames, their performance for mTTA values is not satisfactory. With the consideration of temporal evolution, EQ-TAA (AVD) has superior performance in AUC and $TTA_{0.5}$ metrics to other EQ-TAAs. For the sunny daytime (A3D and CCD datasets), EQ-TAA (AVD) is more promising than other EQ-TAAs and achieves the best AUC value of 0.580 for DADA-2000.

EQ-TAA vs. Supervised TAAs: From the model architecture in Fig. 4, it can be seen that EQ-TAAs belong to the Fr. TAAs. At the same time, we differentiate them with different denotations in Tables II and III for a clearer presentation. As for A3D and CCD datasets, EQ-TAAs seem not better than other supervised Ob.TAAs, which is because Ob.TAAs focus on targeted objects rather than frame-based EQ-TAAs. However, EQ-TAA (AVD) shows competitive performance with other object-centric ones in the DADA-2000 dataset even with a superior $TTA_{0.5}$ value to Ob. TAAs. DADA-2000 is more challenging than A3D and CCD because of severe weather and light conditions (as shown in Fig. 5). Therefore, object detection errors in DADA-2000 can be eliminated in the frame-centric TAAs, such as DRIVE [34] and Cog-TAA [35]. Because of the frame background influence in the generated video frames, the video quality still needs to be enhanced compared with the real accident videos. Therefore, some supervised TAAs show better performance than EQ-TAAs.

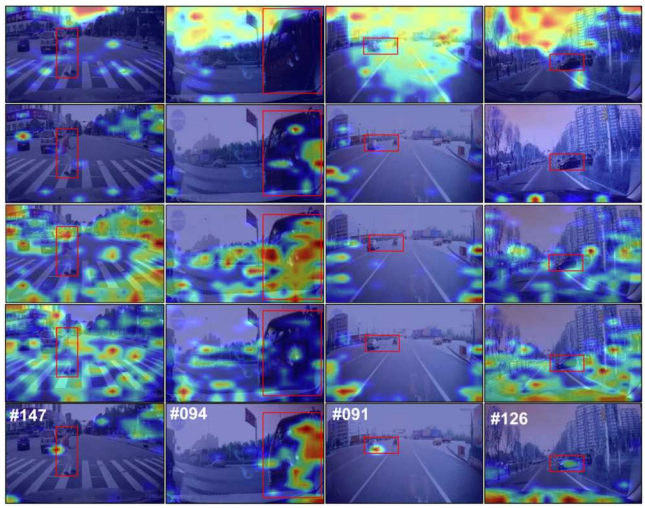


Fig. 9. The activated maps of the Layer Norm (LN) of "S-Trans" by Grad-CAM++ [76], where the EQ-TAA is trained by the training set obtained by Ho et al. [47] (the 1st row), Tune-A-Video [15] (the 2nd row), Text2Video-Zero [74] (the 3nd row), ControlVideo [73] (the 4nd row), and our AVD (the 3rd row), respectively. The crashing objects are marked by red boxes.

Notably, the AUC values of EQ-TAAs indicate that EQ-TAA (AVD) still needs space to be improved while EQ-TAA formulation is promising and does not need the laborious annotation of accident windows. From Figs. 7 and 8, our accident video diffusion can be improved with the object-centric or causality inference.

We also visualize some snapshots in Fig. 9, where the activated map of the Layer Norm (LN) of S-Trans is generated by the Grad-CAM++ method [76]. From the results, it becomes clear that the activated maps by our AVD can localize the crashing object very well compared with other ones. The method of Ho et al. [47] always focuses on the sky regions. Tune-A-Video [15] usually misses the crashing object. Because of the irrelevant video content in the generation, Text2Video-Zero [74] and ControlVideo [73] focus on more background in TAA optimization, which provides the evidence to the low TTA values in Tables II and III.

In summary, EQ-TAA (AVD) can outperform the other diffusion models on the A3D and CCD datasets, and obtain better AUC and $TTA_{0.5}$ values in the DADA-2000 dataset. For the supervised TAAs, the object-centric approach may be a good choice for clear environment conditions, while the frame-centric ones are better for severe conditions without detection error issues. Next, we evaluate the role of different components of EQ-TAA models.

F. Ablations for EQ-TAA

1) EQ-TAA (AVD) vs. Other Supervised TAAs (AVD): To verify the performance of our EQ-TAA (described in Section V-B), we take the generated accident and accident-free video clips (i.e., 1800 pairs of clips) to train two supervised TAAs, i.e., XAI-AN [75] and Cog-TAA [35], and test the performance with

Authorized licensed use limited to: Universidade de Macau. Downloaded on November 27,2026 at 15:11:58 UTC from IEEE Xplore. Restrictions apply.

TABLE IV
THE PERFORMANCE EVALUATION FOR DIFFERENT TAA METHODS *DRIVEN BY AVD*, WHERE THE BEST VALUE OF EACH METRIC IS MARKED IN **BOLD** FONT

Method	A3D [22]		CCD [28]		DADA-2000 [20]	
	AP↑	mTTA↑	AP↑	mTTA↑	AP↑	mTTA↑
XAI-AN [76]	0.788	2.387	0.992	2.960	0.756	2.260
Cog-TAA [35]	0.905	4.128	**0.997**	4.041	**0.781**	4.063
EQ-TAA	**0.953**	**4.321**	0.989	**4.407**	0.725	**4.258**

TABLE V
THE ROLES OF $\mathcal{L}_{ETL}$ AND "ADATOKS" IN EQ-TAA (AVD)

Baselines	Datasets	AP↑	AUC↑	$TTA_{0.5}$↑	mTTA↑	Var↓
EQ-TAA (AVD)	DADA-2000 [20]	**0.725**	**0.580**	**4.338**	**4.498**	**0.049**
-$\mathcal{L}_{\text{ETL}}$		0.702	0.523	3.894	4.258	0.109
-$\mathcal{L}_{\text{ETL}}$-AdaToks		0.724	0.438	3.933	4.378	0.106
EQ-TAA (AVD)	A3D [22]	**0.952**	-	**3.996**	**4.438**	**0.084**
-$\mathcal{L}_{\text{ETL}}$		0.905	-	3.995	4.437	0.110
-$\mathcal{L}_{\text{ETL}}$-AdaToks		0.905	-	3.596	4.321	0.107
EQ-TAA (AVD)	CCD [28]	**0.989**	-	**4.402**	**4.875**	**0.098**
-$\mathcal{L}_{\text{ETL}}$		0.982	-	4.401	4.845	0.151
-$\mathcal{L}_{\text{ETL}}$-AdaToks		0.982	-	4.401	4.407	0.154

aforementioned datasets. Notably, Cog-TAA [35] involves the scene description and driver attention to form a cognitive TAA. Because there is no driver attention and scene description in the generated video clips, we utilize the Cog-TAA version only with video frame input. The training details follow the official setting of XAI-AN [75] and Cog-TAA [35]. From the results in Table IV, Cog-TAA performs better in positive and negative classification than other methods for CCD and DADA-2000 datasets, but our EQ-TAA shows better performance on the earliness warning of accidents (larger mTTA values).

2) Roles of "ETL-Loss" and AdaToks: In the EQ-TAA method, the special modules are the "ETL-Loss" and the Adaptive Token Sampling (AdaToks).

To validate the effectiveness of ETL loss or AdaToks, we remove the ETL loss or AdaToks and compare it with the full EQ-TAA (AVD) model. From Table V, we observe a significant decrease in AP values for both the DADA-2000 and A3D datasets when $\mathcal{L}_{ETL}$ or AdaToks is absent. There is a significant degradation in $TTA_{0.5}$ or mTTA value when $\mathcal{L}_{ETL}$ or the AdaToks is removed. Larger TTA values mean earlier accident anticipation. The highest AUC value of our full EQ-TAA (AVD) in the DADA-2000 dataset seems to suggest that the anticipation with $\mathcal{L}_{ETL}$ or AdaToks may be more promising. In addition, we visualize some anticipated accident occurrence curves for real and synthetic videos in Fig. 10. We find that the anticipated accident score curves by "EQ-TAA-$\mathcal{L}_{ETL}$" and "EQ-TAA-$\mathcal{L}_{ETL}$-AdaToks" show frequent fluctuation. The $TTA_{0.5}$ metric utilized is precautious because it is computed by the time interval between *the first warning time* (the first frame with the accident score larger than 0.5) with the starting time of the accident. Usually, the accident score may decrease to less than 0.5 after the first warning frame. However, to make a precautious warning, many works aforementioned still use it as a main metric. Therefore, we introduce a Variance (**Var**) metric computed by the

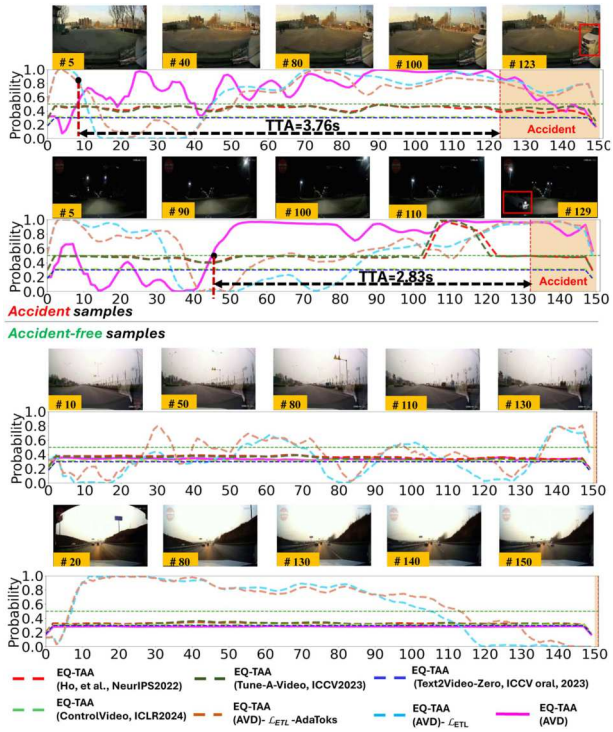


Fig. 10. The anticipated accident occurrence curves of two pairs of positive (top two rows) and negative (bottom two rows) accident videos, where one pair of real samples are sampled in the testing set of DADA-2000 [20].

frame-level variance of the accident scores (> 0.5) to provide a prediction stability evaluation. The results of Table V show that the full EQ-TAA (AVD) generates a more stable prediction.

3) Anticipation Stability of EQ-TAAs Driven by Different Diffusion Models: In addition, we also check the prediction stability of full EQ-TAA driven by other video diffusion models. From the curves in Fig. 10, the full EQ-TAA model driven by AVD is more stable than "EQ-TAA-$\mathcal{L}_{ETL}$" and "EQ-TAA-$\mathcal{L}_{ETL}$-AdaToks". It indicates that the ETL loss and AdaToks modules are indispensable for stable anticipation. Besides, EQ-TAA (AVD) is better than EQ-TAA (Tune-A-Video), EQ-TAA (Ho et al.), EQ-TAA (ControlVideo) and EQ-TAA (Text2Video-Zero). EQ-TAA (Ho et al.), EQ-TAA (ControlVideo), and EQ-TAA (Text2Video-Zero) are powerless to determine the accident and accident-free videos because of the indiscriminative and irrelevant frame content aforementioned in Figs. 7 and 8. Notably, EQ-TAA (ControlVideo), and EQ-TAA (Text2Video-Zero) demonstrate failure for all the samples in Fig. 10, which makes the **Var** value evaluation of them meaningless. As shown in Fig. 11, our EQ-TAA (AVD) is more stable for positive accident anticipation.

4) Sample Scale vs. TAA Accuracy: As for video generation, it is necessary to investigate the sample scale of synthetic accident videos corresponding to the TAA performance. We check the AUC and AP values for different numbers of the generated samples by AVD in the TAA training phase. Fig. 12(a) shows that the AP and AUC values do not show a monotonic increase while demonstrating a decrease when the number of samples is more than 3000. The reason is that our EQ-TAA takes the

Authorized licensed use limited to: Universidade de Macau. Downloaded on November 27,2026 at 15:11:58 UTC from IEEE Xplore. Restrictions apply.

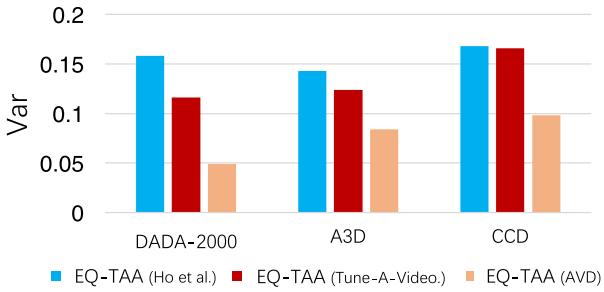


Fig. 11. The comparison of Var values for EQ-TAA framework driven by Ho et al. [47], Tune-A-Video [15], and our AVD.

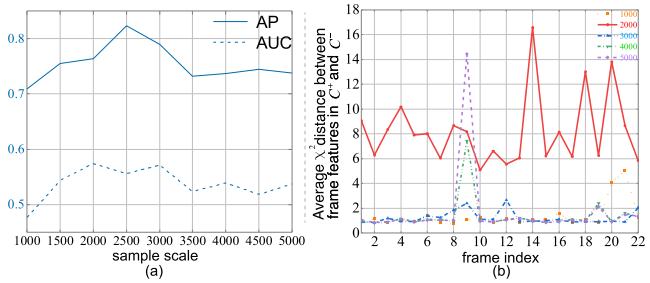


Fig. 12. Sample Scale vs. TAA Accuracy. (a) is accident anticipation accuracy (AP and AUC) concerning the sample scale, and (b) denotes the frame-wise feature difference for the generated positive and negative clips with 22 frames.

Video-to-Video (V2V) generation which leverages the anchor video clip C^a to restrain the style of the generated ones (i.e., C^+ and C^-). When the sample scale increases, it may involve many similar video clips. We plot the average feature difference (Fig. 12(b)) of C^+ and C^- with different sample scales by computing the χ^2 distance of frame-wise feature vectors (22 frames) extracted by ResNet50. 2000 samples own the most manifest feature difference, which positively responds to the observation in Fig. 12(a), i.e., enlarging the feature difference of C^+ and C^- is important.

VII. Conclusions and Discussions

This work for the first time considers the text-video coherent alignment for accident scenarios in driving scenes, resulting in an Attentive Video Diffusion model (AVD) for handling the confounding issue of background frames for Traffic Accident Anticipation (TAA). The AVD generates smoother, style-consistent frames with anchor ones compared with other state-of-the-art video diffusion models. In addition, the AVD model facilitates the Equivariant TAA that can be trained by the video data collected from various driving scenes. EQ-TAA is verified to be comparable to state-of-the-art supervised traffic anticipation methods through extensive experiments.

Discussions: Towards better accident video synthesis for the TAA task, compared with the overemphasized frame quality, the manifestation of the causal object is more important with the restriction of background influence and responsive change of text-guided video content. This is proved by the better positive and negative accident clip classification ability of models trained on real accident datasets. Therefore, the Inception Score (IS) metric is not good because it is easily influenced by the noisy context of the background even though the quality of generated images is good. As a video-based formulation of TAA tasks, the FVD metric is promising because of the temporal correlation. In addition, the metrics for evaluating the riskiness of generated video frames may be more appropriate and will be explored in the future.

Besides, more synthesized samples do not mean better TAA performance, which has two main factors: 1) the feature difference between the synthesized positive and negative samples, and 2) the diversity of the training samples. Better TAA performance prefers larger positive/negative feature differences and more diverse training samples. Therefore, if we sample the normal video clips as the anchor samples, it is better to enlarge the sampling interval of frames in one video or even from different videos.

References

[1] W. H. Organization, "Global health observatory: Number of road traffic deaths," 2023. [Online]. Available: https://www.who.int/gho/road-safety/mortality/traffic-deaths-number/en/
[2] W. M. D. Chia, S. L. Keoh, C. Goh, and C. Johnson, "Risk assessment methodologies for autonomous driving: A survey," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 10, pp. 16923–16939, Oct. 2022.
[3] 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, Jul. 2022.
[4] 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, Apr. 2024.
[5] Y. Yao et al., "Egocentric vision-based future vehicle localization for intelligent driving assistance systems," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2019, pp. 9711–9717.
[6] Z. Chen, K. -Y. Lin, and W. -S. Zheng, "Consistent intra-video contrastive learning with asynchronous long-term memory bank," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 3, pp. 1168–1180, Mar. 2023.
[7] 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, Jan. 2024.
[8] W. Qin, H. Zhang, R. Hong, E. -P. Lim, and Q. Sun, "Causal interventional training for image recognition," *IEEE Trans. Multimedia*, vol. 25, pp. 1033–1044, 2023.
[9] L. Chen, Y. Zheng, Y. Niu, H. Zhang, and J. Xiao, "Counterfactual samples synthesizing and training for robust visual question answering," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 11, pp. 13218–13234, Nov. 2023.
[10] W. Wang, J. Gao, and C. Xu, "Weakly-supervised video object grounding via causal intervention," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 3, pp. 3933–3948, Mar. 2023.
[11] Y. Li, X. Wang, J. Xiao, W. Ji, and T. -S. Chua, "Invariant grounding for video question answering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 2928–2937.
[12] D. Ceylan, C. P. Huang, and N. J. Mitra, "Pix2video: Video editing using image diffusion," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2023, pp. 23149–23160.
[13] S. Liu, Y. Zhang, W. Li, Z. Lin, and J. Jia, "Video-p2p: Video editing with cross-attention control," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 8599–8608.
[14] R. Li, N. Wang, F. Feng, G. Zhang, and X. Wang, "Exploring global and local linguistic representations for text-to-image synthesis," *IEEE Trans. Multimedia*, vol. 22, no. 12, pp. 3075–3087, Dec. 2020.
[15] J. Z. Wu et al., "Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2023, pp. 7589–7599.
[16] J. Fang et al., "Abductive ego-view accident video understanding for safe driving perception," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 22030–22040.

Authorized licensed use limited to: Universidade de Macau. Downloaded on November 27,2026 at 15:11:58 UTC from IEEE Xplore. Restrictions apply.

[17] D. Liu et al., "Generative metric learning for adversarially robust open-world person re-identification," *ACM Trans. Multim. Comput. Commun. Appl.*, vol. 19, no. 1, pp. 20:1–20:19, 2023.

[18] K. Ying et al., "CTVIS: Consistent training for online video instance segmentation," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2023, pp. 899–908.

[19] Y. Xia et al., "Predicting driver attention in critical situations," in *Proc. Asian Conf. Comput. Vis.*, 2018, pp. 658–674.

[20] 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, Jun. 2022.

[21] W. Bao, Q. Yu, and Y. Kong, "Uncertainty-based traffic accident anticipation with spatio-temporal relational learning," in *Proc. ACM Int. Conf. Multimedia*, 2020, pp. 2682–2690.

[22] Y. Yao, M. Xu, Y. Wang, D. J. Crandall, and E. M. Atkins, "Unsupervised traffic accident detection in first-person videos," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2019, pp. 273–280.

[23] Y. Lu, C. Cao, Y. Zhang, and Y. Zhang, "Learnable locality-sensitive hashing for video anomaly detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 2, pp. 963–976, Feb. 2023.

[24] N. Kim and J. -W. Kang, "Dynamic motion estimation and evolution video prediction network," *IEEE Trans. Multimedia*, vol. 23, pp. 3986–3998, 2021.

[25] Z. Chang, X. Zhang, S. Wang, S. Ma, and W. Gao, "STAM: A spatiotemporal attention based memory for video prediction," *IEEE Trans. Multimedia*, vol. 25, pp. 2354–2367, 2023.

[26] M. S. Aliakbarian et al., "Viena: A driving anticipation dataset," in *Proc. Asian Conf. Comput. Vis.*, 2019, pp. 449–466.

[27] D. Xiao, W. G. Geiger, H. Y. Yatbaz, M. Dianati, and R. Woodman, "Detecting hazardous events: A framework for automated vehicle safety systems," in *Proc. IEEE 25th Int. Conf. Intell. Transp. Syst.*, 2022, pp. 641–646.

[28] S. Malla et al., "DRAMA: Joint risk localization and uncertainty-based traffic accident anticipation with spatio-temporal relational learningtioning in driving," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, 2023, pp. 1043–1052.

[29] F. Chan, Y. Chen, Y. Xiang, and M. Sun, "Anticipating accidents in dashcam videos," in *Proc. Asian Conf. Comput. Vis.*, 2016, pp. 136–153.

[30] 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, Jul. 2022.

[31] Y. Seo, M. Defferrard, P. Vandergheynst, and X. Bresson, "Structured sequence modeling with graph convolutional recurrent networks," in *Proc. 25th Int. Conf. Neural Inf. Process.*, 2018, pp. 362–373.

[32] T. Wang et al., "GSC: A graph and spatio-temporal continuity based framework for accident anticipation," *IEEE Trans. Intell. Veh.*, vol. 9, no. 1, pp. 2249–2261, Jan. 2024.

[33] T. Wang et al., "Deepaccident: A motion and accident prediction benchmark for V2X autonomous driving," in *Proc. AAAI Conf. Artif. Intell.*, 2024, pp. 5599–5606.

[34] W. Bao, Q. Yu, and Y. Kong, "DRIVE: Deep reinforced accident anticipation with visual explanation," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2021, pp. 7599–7608.

[35] L. -L. Li, J. Fang, and J. Xue, "Cognitive traffic accident anticipation," *IEEE Intell. Transp. Syst. Mag.*, vol. 16, no. 5, pp. 17–32, Sep./Oct. 2024.

[36] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2020, pp. 6840–6851.

[37] L. Liao et al., "Unsupervised foggy scene understanding via self spatial-temporal label diffusion," *IEEE Trans. Image Process.*, vol. 31, pp. 3525–3540, 2022.

[38] P. Dhariwal and A. Q. Nichol, "Diffusion models beat GANs on image synthesis," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2021, pp. 8780–8794.

[39] J. Choi, S. Kim, Y. Jeong, Y. Gwon, and S. Yoon, "ILVR: Conditioning method for denoising diffusion probabilistic models," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2021, pp. 14347–14356.

[40] A. Lugmayr et al., "Inpainting using denoising diffusion probabilistic models," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 11451–11461.

[41] Y. Kirstain, O. Levy, and A. Polyak, "X&fuse: Fusing visual information in text-to-image generation," 2023, *arXiv:2303.01000*.

[42] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, "Hierarchical text-conditional image generation with CLIP latents," 2022, *arXiv:2204.06125*.

[43] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 10674–10685.

[44] L. Zhang, A. Rao, and M. Agrawala, "Adding conditional control to text-to-image diffusion models," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2023, pp. 3813–3824.

[45] X. Xu, Z. Wang, E. J. Zhang, K. Wang, and H. Shi, "Versatile diffusion: Text, images and variations all in one diffusion model," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2023, pp. 7720–7731.

[46] D. Wang, Z. Hu, Y. Zhou, R. Hong, and M. Wang, "A text-guided generation and refinement model for image captioning," *IEEE Trans. Multim.*, vol. 25, pp. 2966–2977, 2023.

[47] J. Ho et al., "Video diffusion models," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2022.

[48] U. Singer et al., "Make-a-video: Text-to-video generation without text-video data," in *Proc. Int. Conf. Learn. Representations*, 2023.

[49] S. Yin et al., "NUWA-XL: Diffusion over diffusion for extremely long video generation," in *Proc. Annu. Meeting Assoc. Comput. Linguistics*, 2023, pp. 1309–1320.

[50] D. Ko, J. S. Lee, W. Kang, B. Roh, and H. Kim, "Large language models are temporal and causal reasoners for video question answering," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2023, pp. 4300–4316.

[51] Y. Wei, Y. Liu, H. Yan, G. Li, and L. Lin, "Visual causal scene refinement for video question answering," in *Proc. ACM Int. Conf. Multimedia*, 2023, pp. 377–386.

[52] C. Zang, H. Wang, M. Pei, and W. Liang, "Discovering the real association: Multimodal causal reasoning in video question answering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 19027–19036.

[53] Y. Liu et al., "Learning causality-inspired representation consistency for video anomaly detection," in *Proc. ACM Int. Conf. Multimedia*, 2023, pp. 203–212.

[54] X. Yang, H. Zhang, and J. Cai, "Deconfounded image captioning: A causal retrospect," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 11, pp. 12996–13010, Nov. 2023.

[55] M. Glymour, J. Pearl, and N. P. Jewell, *Causal Inference in Statistics: A Primer*. Hoboken, NJ, USA: Wiley, 2016.

[56] T. Suzuki, H. Kataoka, Y. Aoki, and Y. Satoh, "Anticipating traffic accidents with adaptive loss and large-scale incident DB," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 3521–3529.

[57] J. Huang, Y. Huang, Q. Wang, W. Yang, and H. Meng, "Self-supervised representation learning for videos by segmenting via sampling rate order prediction," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 6, pp. 3475–3489, Jun. 2022.

[58] Y. Li, X. Wang, J. Xiao, and T. -S. Chua, "Equivariant and invariant grounding for video question answering," in *Proc. ACM Int. Conf. Multimedia*, 2022, pp. 4714–4722.

[59] A. Radford et al., "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 8748–8763.

[60] E. J. Hu et al., "Lora: Low-rank adaptation of large language models," in *Proc. Int. Conf. Learn. Representations*, 2022.

[61] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," in *Proc. Int. Conf. Learn. Representations*, 2021.

[62] C. Liu et al., "Focus your attention: A focal attention for multimodal learning," *IEEE Trans. Multimedia*, vol. 24, pp. 103–115, 2022.

[63] M. Fayyaz et al., "Adaptive token sampling for efficient vision transformers," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 396–414.

[64] B. Xu, X. Shu, J. Zhang, G. Dai, and Y. Song, "Spatiotemporal decouple-and-squeeze contrastive learning for semisupervised skeleton-based action recognition," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 8, pp. 11035–11048, Aug. 2024.

[65] B. Xu and X. Shu, "Pyramid self-attention polymerization learning for semi-supervised skeleton-based action recognition," 2023, *arXiv:2302.02327*.

[66] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2017, pp. 6626–6637.

[67] W. Yu, M. Zhu, N. Wang, X. Wang, and X. Gao, "An efficient transformer based on global and local self-attention for face photo-sketch synthesis," *IEEE Trans. Image Process.*, vol. 32, pp. 483–495, 2023.

[68] T. Unterthiner et al., "Towards accurate generative models of video: A new metric & challenges," 2018, *arXiv:1812.01717*.

[69] T. Salimans et al., "Improved techniques for training gans," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2016, pp. 2226–2234.

Authorized licensed use limited to: Universidade de Macau. Downloaded on November 27,2026 at 15:11:58 UTC from IEEE Xplore. Restrictions apply.

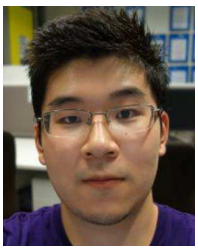
[70] J. Carreira and A. Zisserman, "Quo vadis, action recognition? A new model and the kinetics dataset," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 6299–6308.
[71] C. Szegedy et al., "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 2818–2826.
[72] B. Lefaudeux et al., "xformers: A modular and hackable transformer modeling library," 2022. [Online]. Available: https://github.com/facebookresearch/xformers
[73] Y. Zhang et al., "Controlvideo: Training-free controllable text-to-video generation," in *Proc. Int. Conf. Learn. Representations*, 2024.
[74] L. Khachatryan et al., "Text2video-zero: Text-to-image diffusion models are zero-shot video generators," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2023, pp. 15908–15918.
[75] M. M. Karim, Y. Li, and R. Qin, "Towards explainable artificial intelligence for early anticipation of traffic accidents," *Transp. Res. Rec.*, vol. 2676, no. 6, pp. 743–755, 2021.
[76] A. Chattopadhay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, "Grad-CAM++: Generalized gradient-based visual explanations for deep convolutional networks," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, 2018, pp. 839–847.



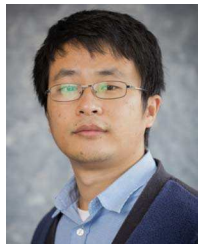
Jianwu Fang (Member, IEEE) received the Ph.D. degree in SIP (signal and information processing) from the University of Chinese Academy of Sciences, Beijing, China, in 2015. He was a Visiting Scholar with NExT++ Research Centre, School of Computing, National University of Singapore, Singapore. He is currently an Associate Professor with the Institute of Artificial Intelligence and Robotics, Xian Jiaotong University, Xian, China. His research interests include computer vision, multimedia analysis, and their applications in intelligent transportation. He chaired many special sessions of ITSC, IV, and ACM MM. He was the recipient of the First Prize of Natural Science in Shaanxi Province and Best Paper Nominees Award in ITSC 2023. He is also an Associate Editor for IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS.



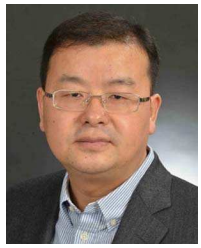
Lei-Lei Li received the bachelor's and master's degrees in transportation from Chang'an University, Xi'an, China, in 2021 and 2024, respectively. He is currently working toward the Ph.D. degree with Xi'an Jiaotong University, Xi'an. His research interests include traffic video processing and accident anticipation.



Zhedong Zheng received the B.S. degree from Fudan University, Shanghai, China, in 2016, and the Ph.D. degree from the University of Technology Sydney, Ultimo, NSW, Australia, in 2021. He was a Postdoctoral Research Fellow with the School of Computing, National University of Singapore, Singapore. He is currently an Assistant Professor with the University of Macau, Macao, China. His research interests include robust learning for image retrieval, generative learning for data augmentation, and unsupervised domain adaptation. He was the recipient of IEEE Circuits and Systems Society Outstanding Young Author Award of 2021. He was the Senior PC of IJCAI and AAAI and Area Chair of ACM MM'24, ACM MM'25, and ICASSP'25.



Hongkai Yu received the Ph.D. degree in computer science and engineering from the University of South Carolina, Columbia, SC, USA. He is currently an Assistant Professor with the Department of Electrical and Computer Engineering, Cleveland State University, Cleveland, OH, USA. His research interests include computer vision, machine learning, intelligent transportation, and intelligent vehicles. He is also the Area Chair of IEEE MIPR 2022 and ACM MM 2022-2024 conferences. He is an Associate Editor for IEEE TRANSACTIONS ON INTELLIGENT VEHICLES.



Jianru Xue (Member, IEEE) received the B.S. degree from the Xi'an University of Technology, Xi'an, China, in 1994, and the M.S. and Ph.D. degrees from Xi'an Jiaotong University, Xi'an, in 1999 and 2003, respectively. From 2002 to 2003, he was with FujiXerox, Tokyo, Japan and visited the University of California, Los Angeles, CA, USA, from 2008 to 2009. In 1999, he joined the Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University, where he currently is a Full Professor. His research interests include computer vision, visual localization and navigation, and video coding based on analysis. He and his team were winners of IEEE ITS Institute Lead Award in 2014. He and his students won the Best Application Paper award at the Asian Conference on Computer Vision 2012.



Zhengguo Li (Fellow, IEEE) received the B.Sc. degree in applied mathematics, and the M.Eng. degree in automatic control from Northeastern University, Shenyang, China, in 1992 and 1995, respectively, and the Ph.D. degree in automatic control from Nanyang Technological University, Singapore, in 2001. He is currently with the Agency for Science, Technology and Research, Singapore. Dr. Li has been actively involved in the development of H.264/AVC and HEVC since 2002. His current research interests include computational photography, mobile imaging, video processing, and delivery, and switched and impulsive control. He was the TPC Chair of IEEE IECON in 2023 and 2020, respectively, Special Session Chair of IEEE ICASSP 2022, Technical Brief Co-Founder of SIGGRAPH Asia, and leading Workshop Chair of IEEE ICME in 2013. He is an Associate Editor for IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY from 2020 to 2024, Senior Area Editor of IEEE TRANSACTIONS ON IMAGE PROCESSING from 2020 to 2024, and Editor of MDPI *Sensors* from 2022 to 2024 and *APSIPA Transactions on Signal and Information Processing* from 2022 to 2025. He is awarded Asia-Pacific Artificial Intelligence Association Fellow.



Tat-Seng Chua received the Ph.D. from the University of Leeds, Leeds, U.K. He was the Acting and Founding Dean with the School from 1998 to 2000. He is currently the KITHCT Chair Professor with the School of Computing, National University of Singapore, Singapore. He is the Co-Director with NExT, a joint center between NUS and Tsinghua University to develop technologies for live social media search and Co-Founder of two technology startup companies in Singapore. His main research interests include multimedia information retrieval, social media analytics, and the extraction, retrieval, and question-answering (QA) of text and rich media arising from the web and multiple social networks. Dr. Chua is the 2015 winner of the prestigious ACM SIGMM award for Outstanding Technical Contributions to Multimedia Computing, Communications, and Applications. He is the Chair of the steering committee of ACM International Conference on Multimedia Retrieval and Multimedia Modeling conference series. Dr. Chua is also the General Co-Chair of ACM Multimedia 2005, ACM CIVR (now ACM ICMR) 2005, ACM SIGIR 2008, and ACM Web Science 2015. He was on the editorial boards of four international journals.

Authorized licensed use limited to: Universidade de Macau. Downloaded on November 27,2026 at 15:11:58 UTC from IEEE Xplore. Restrictions apply.