SCIENCE CHINA
Information Sciences

• RESEARCH PAPER •

November 2026, Vol. 69, Iss. 11, 212101:1–212101:17
https://doi.org/10.1007/s11432-025-5095-1

Learning from the unseen: generative data augmentation for geometric-semantic accident anticipation

Yanchen GUAN[1,2], Haicheng LIAO[1,3], Chengyue WANG[1,2], Xingcheng LIU[1,3], Jiaxun ZHANG[1,2], Keqiang LI[4] & Zhenning LI[1,2,3*]

[1]*State Key Laboratory of Internet of Things for Smart City, University of Macau, Macao 999078, China*
[2]*Department of Civil and Environmental Engineering, University of Macau, Macao 999078, China*
[3]*Department of Computer and Information Science, University of Macau, Macao 999078, China*
[4]*Department of Automotive Engineering, Tsinghua University, Beijing 100084, China*

Received 14 March 2025/Revised 1 December 2025/Accepted 22 January 2026/Published online 26 October 2026

Abstract Anticipating traffic accidents is a critical yet unresolved problem for autonomous driving, hindered by the inherent complexity of modeling interactions between road users and the limited availability of diverse, large-scale datasets. To address these issues, we propose a dual-path framework. On the one hand, we employ a video synthesis pipeline that, guided by structured prompts, derives feature distributions from existing corpora and produces high-fidelity synthetic driving scenes consistent with the statistical patterns of real data. On the other hand, we design a graph neural network enriched with semantic cues, enabling dynamic reasoning over both spatial and semantic relations among participants. To validate the effectiveness of our approach, we release a new benchmark dataset containing standardized, finely annotated video sequences that cover a broad spectrum of regions, weather, and traffic conditions. Evaluations across existing datasets and our new benchmark confirm notable gains in both accuracy and anticipation lead time, highlighting the capacity of the proposed framework to mitigate current data bottlenecks and enhance the reliability of autonomous driving systems.

Keywords autonomous driving, accident anticipation, synthetic videos, vision language model, graph convolutional networks

Citation Guan Y C, Liao H C, Wang C Y, et al. Learning from the unseen: generative data augmentation for geometric-semantic accident anticipation. Sci China Inf Sci, 2026, 69(11): 212101, https://doi.org/10.1007/s11432-025-5095-1

1 Introduction

In recent years, autonomous driving technologies have progressed rapidly, and self-driving vehicles are increasingly regarded as promising solutions for enhancing road safety [1,2]. Anticipating potential accidents is a key component in ensuring timely risk awareness within such scenarios, as it focuses on forecasting potential collisions from dashcam recordings and issuing early alerts. This ability allows drivers or automated systems to take preventive actions in advance, thereby lowering accident likelihood and improving traffic safety [3–5].

Distinct from generic time-series prediction tasks, accident anticipation is characterized by its reliance on anomalous interactions among traffic participants within specific environments, rather than merely long-term sequential dynamics. Existing studies mainly emphasize extending temporal modeling, thereby improving general sequence understanding but overlooking the underlying causes of accidents [6, 7]. This often results in weakly perceived interactions among distant targets, while close interactions suffer from extremely short time-to-accident (ToA), occlusions, or blurred cues—issues that are particularly detrimental in rapidly evolving traffic scenes [8].

Furthermore, the lack of sufficiently large and reliable datasets is also a primary obstacle. Since traffic crashes occur infrequently and in unpredictable circumstances, their randomness and time-critical nature make the collection, curation, and labeling of accident videos particularly costly and labor-intensive [9]. Consequently, accident datasets are limited in size and vary substantially across sources, leading to instability in model performance and a heightened risk of overfitting when evaluated on small, independent datasets [10]. Moreover, cross-dataset discrepancies are pronounced; for example, the DAD dataset [11] reflects subtropical urban scenes without snowfall, while the CCD dataset [12] primarily captures traffic scenarios on temperate open roads. Such differences may cause systematic biases, where specific scenarios or categories are underrepresented.

* Corresponding author (email: zhenningli@um.edu.mo)

© Science China Press 2026

To tackle these challenges, we propose a dual-pronged pipeline. We first integrate a frozen video generation module into the processing pipeline of the original data, leveraging the environmental feature distribution of the raw dataset to construct new traffic scenarios. This approach increases both the volume and diversity of training data while preserving the intrinsic characteristics of the dataset to the greatest extent. Subsequently, to overcome limitations of prior research, we propose a semantic- and geometric-enhanced dynamic graph convolutional network for accident anticipation inference. This architecture captures complex behavioral interactions among participants through joint semantic and spatial-visual fusion, adaptively weighting critical targets to guide the model's focus toward hazardous interactions. Furthermore, instead of treating frames independently, our model processes short video segments, effectively enlarging the receptive field, alleviating transient information loss, and enhancing prediction accuracy.

For empirical validation, we present the multi-source accident anticipation (MAA) dataset, a new and challenging real-world benchmark with richer accident samples and more detailed annotations than prior datasets. We further replicate several representative models based on recurrent neural networks (RNNs), graph convolutional networks (GCNs), Transformers, and self-supervised approaches on the MAA dataset. Results indicate that our framework consistently surpasses prior art across diverse datasets. Moreover, the proposed augmentation strategy effectively mitigates model performance bottlenecks, leading to more stable accident anticipation.

Our main contributions can be summarized as follows.

• We present a traffic video synthesis pipeline for data augmentation in accident anticipation. The method produces diverse driving scenarios while maintaining the statistical characteristics of the original environments, thereby alleviating the challenge of limited training data.

• We develop a semantic- and geometric-enhanced dynamic graph convolutional network with temporal convolutional layers. By integrating the behavioral semantics and visual information of traffic participants with an enhanced temporal reasoning module, our framework achieves more accurate accident anticipation in multiple experiments.

• We provide the MAA dataset, which constitutes the most extensive collection of accident cases with detailed annotations, supporting reproducible evaluation and future exploration.

2 Related work

Accident anticipation aims to forecast collisions before they occur and to provide early warnings, allowing proactive safety measures to be taken. The concept was first explored by Chan et al. [11], who introduced a dynamic spatial attention framework using egocentric visual input, demonstrating the potential of predicting accidents directly from dashcam footage. Subsequent advances extended this idea: Karim et al. [13] incorporated temporal dependencies through a dynamic spatio-temporal attention model, while others introduced graph-based approaches that represent vehicles and pedestrians as nodes, with edges encoding their interactions [14–18]. Such methods often rely on dense connectivity to approximate relational structures [3,10,19]. Parallel efforts have explored incorporating motion cues such as optical flow and depth information to refine object perception, which further improves accident anticipation accuracy [20,21]. Despite these developments, key challenges persist, including the inefficiency of relation modeling and the limited scale of available datasets. Graph structures constructed solely from visual information struggle to capture real-world traffic semantics such as road constraints and vehicle priority relationships, and they often exhibit instability in rare scenarios [22,23]. Moreover, the limited scale of existing public datasets further exacerbates the long-tailed distribution of rare events, resulting in weak model generalization capability [24,25].

Beyond structured interaction modeling, research on scene graph generation has shown that explicitly representing semantic relations between objects can enrich video interpretation [26,27]. In addition, recognizing human actions from video provides a pathway to deeper understanding of ongoing events. Vision-language models (VLMs) have become a promising tool to bridge visual signals with textual representations [28–30]. These pretrained systems bring zero-shot interpretability, enhancing both contextual reasoning and content recognition [31]. In video analysis specifically, VLMs exploit temporal continuity across frames, enabling them to capture evolving actions and mitigate the limitations of single-frame inference, thereby converting continuous visual information into semantic information [32]. Compared with fragmented frame-level visual cues, such continuous semantic information provides a more coherent description of inter-agent interactions and critical environmental influences [33,34]. In addition, the zero-shot reasoning capability of VLMs enables the extraction of diverse semantic attributes for each sample without relying on multiple specialized models [28]. This facilitates a unified characterization of dataset-wide semantic distributions and offers valuable guidance for constructing semantically aligned synthetic data [35,36].

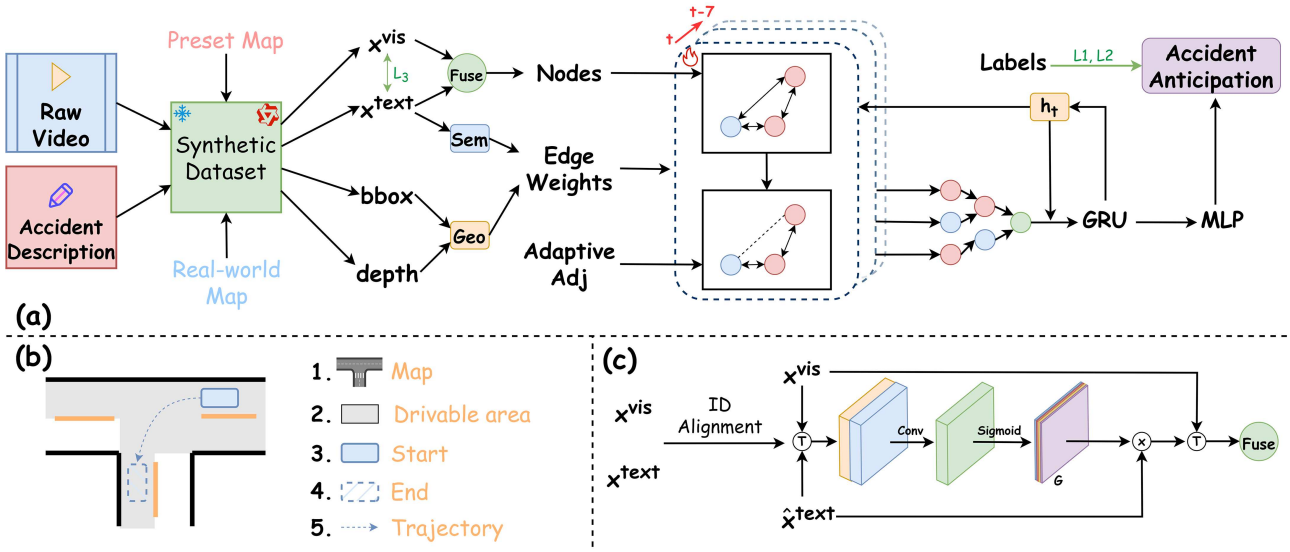


Figure 1 (Color online) The overall framework of our proposed model. (a) The process from raw video to accident anticipation results including traffic video generation and accident anticipation; (b) the important factors that constitute the traffic scene and the order in which they are obtained; (c) the semantic feature and visual feature fusion module.

3 Method

3.1 Problem setup

This work aims to determine, from a stream of dashcam frames, how the risk of a traffic accident evolves over time and to identify the moment at which such risk becomes critical. Let $\mathcal{I} = \{i_1, i_2, \ldots, i_L\}$ denote the input video sequence consisting of L frames. At each frame index m, the proposed model produces a risk estimate, yielding a probabilistic profile $\mathcal{U} = \{u_1, u_2, \ldots, u_L\}$, where u_m indicates the inferred accident likelihood at that instant.

An accident-warning decision is triggered once the predicted score surpasses a decision level δ for the first time at index $\hat{m}$, that is, $u_{\hat{m}} \geqslant \delta$ with $\hat{m} < L$. To characterize how far ahead of the true event the model can raise an alarm, we compute the time-to-accident (TTA) as $\mathrm{TTA} = \lambda - \hat{m}$, where λ marks the ground-truth frame at which the accident begins, and we set $\lambda = 0$ for sequences without accidents. Larger TTA values indicate that the system identifies dangerous situations earlier, thus demonstrating a higher level of anticipatory ability. The overall objective is to learn a reliable risk curve $\mathcal{U}$ while maximizing the advance-detection index $\hat{m}$ for videos in which an accident occurs.

3.2 Traffic video synthesis

Figure 1 shows our entire model framework. We first start with the generation of synthetic data. This module generates driving videos for data augmentation to tackle the data insufficiency.

Given the restricted diversity of existing datasets, we propose a controllable traffic video synthesis framework for synthesizing plausible driving videos. The proposed framework leverages environmental distributions extracted from existing data and reconstructs them into novel scenarios, ensuring that the generated samples preserve the statistical properties of the original dataset. Concretely, we first employ Qwen-VL [37] to obtain the underlying environmental distributions (Figure 2). When the testing samples remain fixed, preserving the original distribution of the dataset helps prevent the training process from being contaminated by environmental conditions absent from the dataset. However, for broader real-world applications, it is essential to address the missing environmental feature distributions, enabling the model to learn from more diverse scenarios. Next, random mapping based on the extracted distributions is applied to controllably determine the environment of new scenarios.

We then adopt two different strategies to assign maps for negative and positive videos. For positive samples, the inherent complexity of designing accident trajectories necessitates pre-defining both the accident type and its spatial location. To address this challenge, we employ preset maps that simplify trajectory design while ensuring coverage of common accident scenarios. According to the National Motor Vehicle Crash Causation Survey (NMVCCS), more than 70% of accidents occur on intersections or straight roads, with intersections alone accounting for nearly 40% [38]. Therefore, as illustrated in Figure 3, our preset maps primarily consist of intersections, T-junctions,

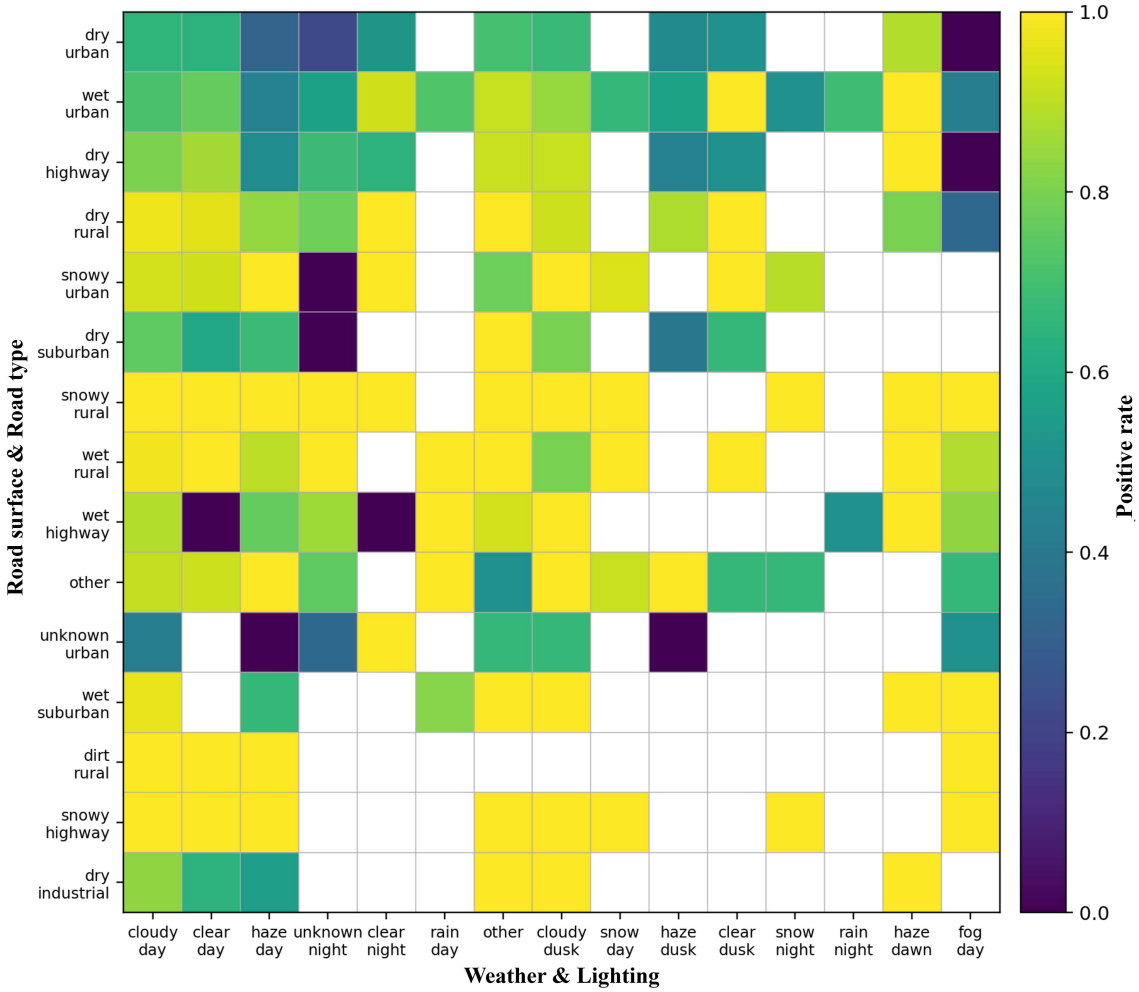


Figure 2 (Color online) Heatmap of positive sample proportions under different environments in the MAA dataset. Within the dataset, accidents' occurrence varies substantially across different conditions. Moreover, despite being the largest accident anticipation dataset to date, certain environments still lack sufficient samples.

straight segments, as well as single-lane and multi-lane configurations.

For negative samples, the map assignment is comparatively straightforward. To increase terrain diversity, we randomly select coordinates from publicly available road data provided by the U.S. National Highway Traffic Safety Administration (NHTSA) and download the corresponding high-definition maps.

Figure 1(b) illustrates the key information required for generating driving scenarios. After obtaining the map and defining drivable regions, it is necessary to assign origins and destinations for traffic participants. For participants that do not require fine-grained control, random routes are assigned based on the map and traffic rules, ensuring conflict-free trajectories. Beyond these initial trajectories, each participant's real-time actions are governed by a driving agent.

Specifically, a directed weighted graph is first constructed based on the road network, denoted as

$$G = (V, E, w), \tag{1}$$

where V represents the set of road nodes, $E \subseteq V \times V$ denotes the set of directed edges corresponding to traversable road segments, and w is the temporal weight function that assigns a travel cost to each edge. Each edge $e_{ij} \in E$ is associated with the following attribute:

$$w(e_{ij}) = \frac{L_{ij}}{v_{ij}}, \tag{2}$$

where L_{ij} denotes the road length and v_{ij} represents the speed limit.

The sets of traversable and non-internal edges are defined as the origin and destination edge sets, denoted by

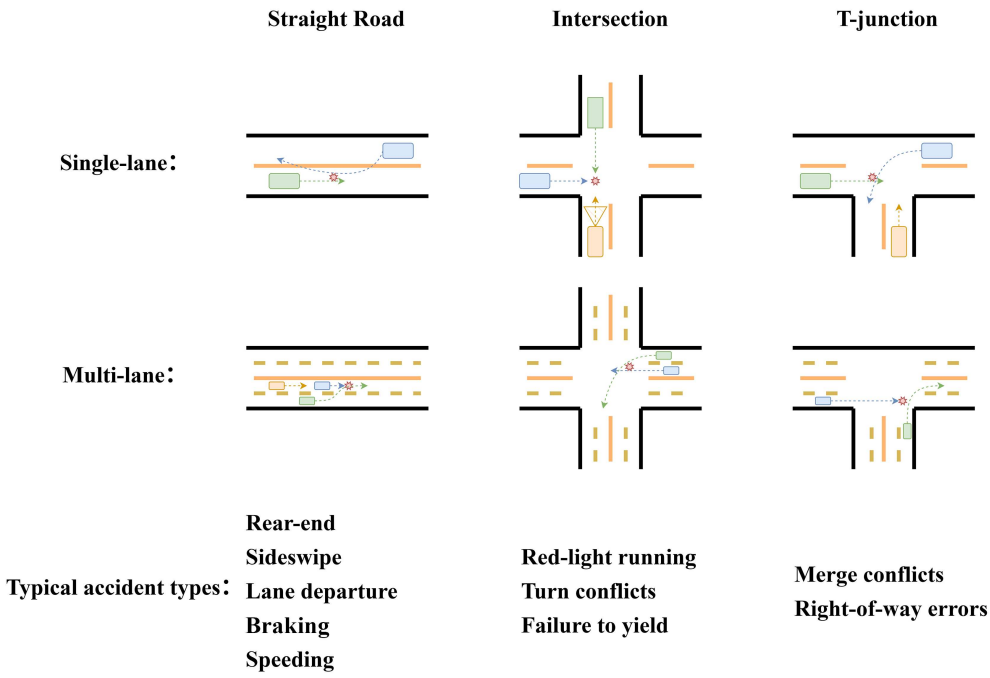


Figure 3 (Color online) The preset maps of accident driving scenarios generation and corresponding accident types. We have designed some preset maps for common scenarios and summarized the corresponding typical accident types to ensure that common accidents can be reconstructed.

E_{src} and E_{dst}, respectively. For each vehicle k, its starting and ending edges are obtained by random sampling:

$$e_s(k) \sim E_{\text{src}}, \qquad e_t(k) \sim E_{\text{dst}}. \tag{3}$$

Vehicle departures are modeled as a Poisson process with rate λ. Let $\{N(t)\}_{t\geqslant 0}$ denote the counting process of vehicle insertions. Then $N(t) \sim \text{Poisson}(\lambda t)$ and the inter-arrival times Δt_k follow an exponential distribution:

$$\Delta t_k \sim \text{Exp}(\lambda), \qquad t(k) = \sum_{i=1}^{k} \Delta t_i, \tag{4}$$

where $t(k)$ is the departure time of vehicle k, $\lambda = \frac{p}{b}$, and p denotes the expected number of vehicle departures within a time window of length b.

For each origin-destination (OD) edge pair $\big(e_s(k), e_t(k)\big)$, the shortest path is computed using the Dijkstra [39] algorithm:

$$\pi(k) = \arg \min_{\pi \in \Pi\left(e_s(k), e_t(k)\right)} \sum_{e_{ij} \in \pi} w(e_{ij}), \tag{5}$$

where $\Pi\big(e_s(k), e_t(k)\big)$ denotes the set of all feasible paths from the origin to the destination. The corresponding path length is given by

$$L(k) = \sum_{e_{ij} \in \pi(k)} L_{ij}. \tag{6}$$

Given a simulation horizon T_{sim}, the expected number of vehicles is

$$\mathbb{E}[N] = \lambda T_{\text{sim}}, \tag{7}$$

and in our implementation, we set $T_{\text{sim}} = 6$ s and discard the first and last 0.5 s when rendering, in order to stabilize the number of visible vehicles and the visual quality.

The entire set of random trips (OD, route, and time for each vehicle) is defined as

$$T = \left\{ \big(e_s(k), e_t(k), t(k), \pi(k)\big) \,\middle|\, k = 1, 2, \ldots, N \right\}. \tag{8}$$

For any trip $\big(e_s(k), e_t(k), t(k), \pi(k)\big)$, the continuous geometric curve along $\pi(k)$ is formed by concatenating the edge-wise centerlines:

$$r^{(k)}(s) = \mathrm{concat}\big(r_{e_1}(s),\, r_{e_2}(s),\, \dots\big), \qquad s \in [0, L(k)]. \tag{9}$$

According to the edge speed v_e, a temporal mapping $t(s)$ is established as

$$t(s) = t(k) + \int_0^s \frac{1}{v(\sigma)}\, \mathrm{d}\sigma. \tag{10}$$

For each temporal sample $t_n^{(k)}$, the spatial position is obtained by inverting the mapping $t(s)$ to find s_n such that $t(s_n) = t_n^{(k)}$, and evaluating the curve $r^{(k)}(s_n)$. The trajectory is sampled with a fixed time step Δt as

$$X^{(k)} = \Big\{ \big(x_n^{(k)},\, y_n^{(k)},\, t_n^{(k)}\big) \Big\}_{n=1}^{N_k}, \qquad t_n^{(k)} = t(k) + n\Delta t, \tag{11}$$

where $N_k = \big\lfloor T^{(k)}/\Delta t \big\rfloor$ and $T^{(k)}$ is the travel time of vehicle k.

All trajectories collectively constitute the output trajectory set:

$$X = \Phi(T) = \Big\{ X^{(1)},\, X^{(2)},\, \dots,\, X^{(N)} \Big\}. \tag{12}$$

After obtaining the initial trajectories, UniAD [40] is employed as the driving agent of each vehicle to establish a closed-loop control system among trajectories, rendered images, and the agent itself. When no surrounding vehicles are present within the ego vehicle's interaction range, the ego vehicle follows its predefined trajectory $X^{(i)}$. When potential interactions with nearby vehicles occur, the driving agent infers the motion relationships between the ego and surrounding vehicles based on real-time rendered images. It then combines this perception with the target route to control the ego vehicle, yielding an updated driving trajectory denoted as $\hat{X}^{(i)}$.

For accident-related vehicles and the ego vehicle in positive scenarios, trajectories are deterministically generated to ensure that the accident occurs within the ego vehicle's dashcam field of view at a specific time step. Therefore, after obtaining $\Pi\big(e_s(k), e_t(k)\big)$, we feed the textual description of the accident type, the candidate OD pairs of the primary vehicles, and the road network map in XML format into Qwen3-Max [37]. The LLM is instructed to stochastically generate multiple trajectory scenarios corresponding to the same accident type under the following constraints: (i) the origins and destinations of all accident-involved vehicles must satisfy the input OD pairs; (ii) the trajectories of the accident vehicles must contain at least an intersection point to ensure that a collision occurs; (iii) an additional ego vehicle must be present in the scene, and the accident location at the moment of collision must lie within the ego vehicle's forward field of view, specifically within the $\pm 30^\circ$ range of the front-facing camera; and (iv) the generated trajectories for all vehicles must be provided with the accident time explicitly annotated. All generated samples are then manually inspected, and those that exhibit clearly unreasonable behaviors are discarded.

After generating trajectories for all agents, the environmental distributions extracted from the original dataset are used as prompts to guide a pretrained stable diffusion model trained in nuScenes [41] in rendering the scenes. The denoising process is shown in Figure 4. Let the diffusion timestep be denoted as τ, and the noise prediction network as ϵ_θ. The reverse diffusion process is defined as

$$z_{\tau-1} = f_\theta(z_\tau,\, e_{\mathrm{cond}}), \tag{13}$$

where the conditional embedding is given by

$$e_{\mathrm{cond}} = \{e_{\mathrm{text}},\, e_{\mathrm{cam}},\, e_{\mathrm{map}},\, e_{\mathrm{layout}},\, e_{\mathrm{ref}},\, e_{\mathrm{rel}}\}. \tag{14}$$

Here, e_{text} is obtained using a CLIP text encoder. The intrinsic and extrinsic camera parameters e_{cam} are encoded via Fourier embeddings. The BEV grid e_{map} is extracted using a convolutional encoder with positional encoding. All lane and boundary maps, together with vehicle bounding boxes in BEV, are projected onto the image plane of each camera. The concatenation of the lane and vehicle projections is then fed into a conditional encoder to produce e_{layout}. The reference embedding e_{ref} corresponds to a randomly sampled historical frame, introduced to maintain temporal semantic consistency. Let the current ego pose be T_t and the reference pose be T_r; the relative transformation is given by

$$\Delta T = T_t T_r^{-1}, \tag{15}$$

and its Fourier embedding is expressed as

$$e_{\mathrm{rel}} = \gamma(\Delta T), \tag{16}$$

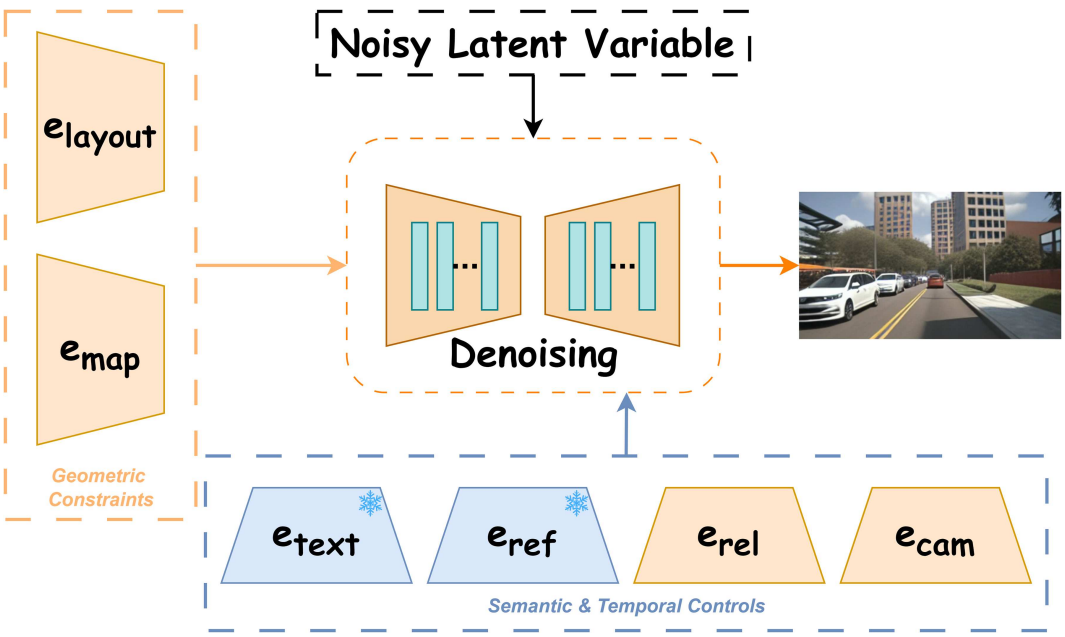


Figure 4 (Color online) The denoising process of scenes' rendering. Geometric constraints use vehicle projection and maps as mandatory constraints, while semantic and temporal controls guide the content of the generated images in terms of semantics and temporal continuity.

which enforces temporal consistency of the background. To ensure geometric coherence among multi-camera views, a cross-view attention mechanism is introduced:

$$h_i' = \mathrm{Attn}\big(h_i, \{h_j\}_{j\neq i}\big), \tag{17}$$

where h_i denotes the latent feature corresponding to the i-th camera.

During inference, an autoregressive strategy is adopted to enhance temporal stability. For the t-th generated frame, the reference embedding is obtained from the previously generated frame $t-1$, and the relative ego-pose embedding is derived from the motion between consecutive poses:

$$e_{\mathrm{ref}}^{(t)} = \mathrm{CLIP}(I_{t-1}), \quad e_{\mathrm{rel}}^{(t)} = \gamma(T_t T_{t-1}^{-1}), \tag{18}$$

leading to the final reverse diffusion-based generation for frame t:

$$I_t = g_\theta\big(z_\tau^{(t)}, e_{\mathrm{text}}, e_{\mathrm{cam}}, e_{\mathrm{map}}, e_{\mathrm{layout}}, e_{\mathrm{ref}}^{(t)}, e_{\mathrm{rel}}^{(t)}\big), \tag{19}$$

where $z_\tau^{(t)}$ denotes the noisy latent variable at diffusion step τ for frame t.

Figure 5 shows some examples of actual synthetic traffic scenes, including HD maps and corresponding multi-view images of both negative and positive scenes, rendering examples in different weather conditions, and some synthetic dashcam videos.

To evaluate the quality of the synthetic traffic scene videos, we conduct a comparative study using the DAD dataset as a reference and benchmark against another synthetic dataset, EMM-AU [42]. The evaluation metrics include Fréchet Inception Distance (FID), Inception Score (IS), Fréchet Video Distance (FVD), Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index (SSIM), and CLIP similarity [43–47].

The experimental results in Table 1 demonstrate that, compared with EMM-AU, our generated videos are substantially closer to the real distribution, achieving superior fidelity and clarity. However, in terms of temporal consistency and discriminability, our method is slightly weaker than EMM-AU. Moreover, under the same benchmarks, both approaches exhibit comparable levels of semantic alignment. From a visual standpoint, our generated videos display certain imperfections in lane line consistency and distortions in the continuous shapes of vehicles and buildings. Nevertheless, compared with EMM-AU, since our generative model is employed solely for rendering, the synthesized videos avoid unrealistic vehicle dynamics and remain kinematically closer to real-world behavior.



Figure 5 (Color online) Examples of the generated videos. (a) Multi-view videos generated for accident and non-accident scenarios along with corresponding high-definition maps and traffic flows; (b) video examples of different weather and road conditions; (c) examples of primary perspective videos in chronological order.

Table 1 Visual quality comparison of synthetic videos. **Bold** numbers indicate superior performance for the corresponding metric.

Dataset	FID	IS	FVD	PSNR	SSIM	CLIP
EMM-AU [42]	175.58	**2.71**	**30.87**	9.38	0.39	0.84
Ours	**91.28**	2.17	34.46	**10.18**	**0.40**	**0.84**

3.3 Accident anticipation framework

3.3.1 *Video preprocessing*

In the initial stage of video processing, as shown in Figure 6, it is essential to rapidly and accurately detect vehicles, pedestrians, and other traffic-related entities in each frame. We first apply YOLOv8 [48], a lightweight and high-performance object detection model widely adopted in industrial applications, to perform object detection on the input videos. Subsequently, the number of detected objects per frame is capped at 19, and the ByteTracker [49] algorithm is utilized to assign a consistent ID across frames for each target. Visual features are extracted from

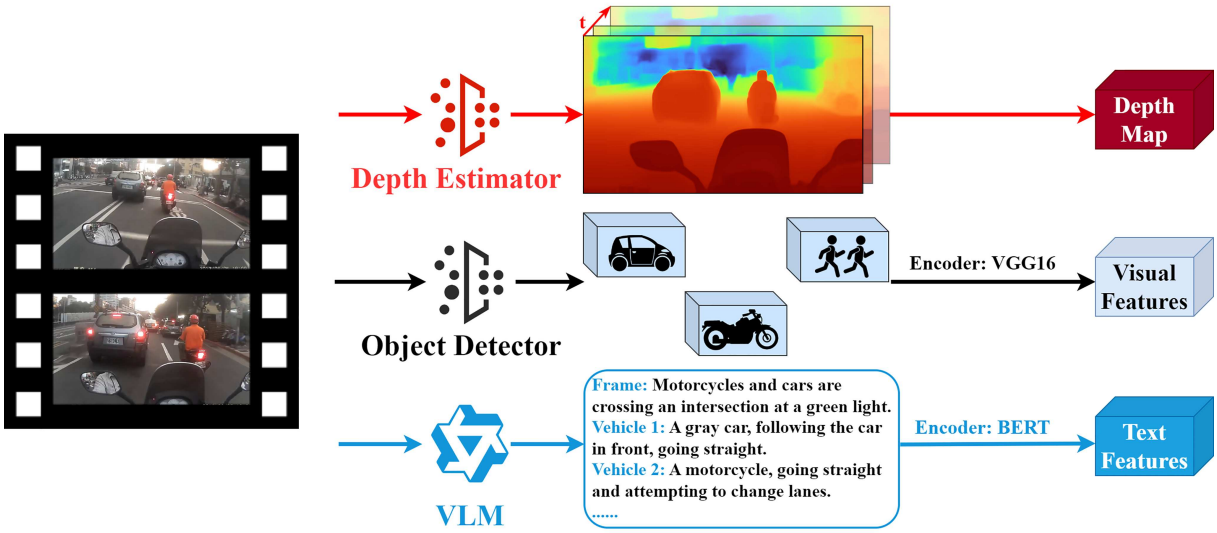


Figure 6 (Color online) Video preprocessing framework. We preprocess each video to obtain the key information required for inference, including object- and frame-level visual features, semantic features, and geometric cues.

both the entire frame and the detected objects using a VGG16 backbone [50], while depth information for every pixel is estimated with the ZOE Depth model [51]. In addition, Qwen-VL [37] is employed to generate behavioral descriptions of each target and a scene-level summary, which are then matched with their corresponding visual IDs. Finally, textual information is encoded with BERT [52] and aligned with the visual feature dimensions, thereby completing the preprocessing stage. The data structure after the above data processing is as follows:

$$\begin{aligned} &F_{vis} \in \mathbb{R}^{B\times T\times N\times F}, \\ &Det \in \mathbb{R}^{B\times T\times (N-1)\times F}, \\ &F_{text} \in \mathbb{R}^{B\times T\times N\times F}, \\ &Depth \in \mathbb{R}^{B\times T\times W\times H}, \end{aligned} \tag{20}$$

where F_{vis}, Det, F_{text}, and $Depth$ refer to the visual feature representation, the object detection output, the text-derived feature tensor, and the depth information. The symbols B, T, N, and F indicate the batch count, the temporal length of the sequence, the total number of aggregated features (covering both frame-level and object-level cues), and the dimensionality of each feature vector. For the depth modality $Depth$, the parameters W and H specify the spatial resolution of each frame.

3.3.2 *Graph construction with geometric and semantic guided connectivity*

To capture spatial interactions among traffic participants, we model each frame as a graph where objects are nodes and pairwise relations form edges. The adjacency matrix is adaptively learned and further refined by both geometric and semantic cues.

Given detected objects per frame $O = N - 1$, we parameterize the adjacency matrix with low-rank factorization:

$$A = \phi_{softmax}(UV) \in \mathbb{R}^{O\times O}, \quad \tilde{A} = D^{-\frac{1}{2}}(A + I)D^{-\frac{1}{2}}, \tag{21}$$

where $U \in \mathbb{R}^{O\times O}$ and $V \in \mathbb{R}^{O\times O}$ are learnable embeddings, I is the identity matrix, and D is the diagonal degree matrix. The role of this adaptive adjacency matrix is to dynamically model the relationship between objects in a learnable way, and to ensure stable and reasonable feature aggregation by determining which objects will interact with each other and the strength of the interaction.

Traffic participants located near one another exhibit an increased risk of potential crashes, particularly when their speed differences are large. In contrast, entities separating from each other are less likely to be engaged in hazardous interactions. Therefore, to enrich connectivity with physical constraints, we compute the 3D geometric distances and relative velocities between objects. For each video, let $c_{t,i}$ denote the 2D center of object i at frame t, and $z_{t,i}$ denote its depth from $Depth$. The combined distance is defined as

$$d_{t,ij} = s^2\|c_{t,i} - c_{t,j}\|_2^2 + |z_{t,i} - z_{t,j}|^2, \tag{22}$$

where s is a normalization constant. The relative velocity is estimated by frame-wise differences:

$$v_{t,ij} = d_{t,ij} - d_{t-1,ij}. \tag{23}$$

After normalization, geometry-driven edge weights are given by

$$W^{geo}_{t,ij} = \alpha \exp(-\bar{d}_{t,ij}) + (1-\alpha)\,\bar{v}_{t,ij}, \quad \alpha = \frac{a}{a+1} \tag{24}$$

with the initial balance parameter $a = 1$.

In addition to visual-spatial relationships, we explicitly model semantic object interactions in the construction of the graph structure, thereby simultaneously considering the spatial and semantic dependencies between objects. This significantly improves the ability to model long-term dependencies and causal reasoning, making the model's understanding of complex scenes more comprehensive and accurate. For semantics, object-level textual embeddings aligned with IDs are normalized, and cosine similarities are converted into text-driven edge weights:

$$W^{text}_{t,ij} = \frac{\exp\left(\langle \hat{x}^{text}_{t,i}, \hat{x}^{text}_{t,j} \rangle / \tau_{text}\right)}{\sum_{j'} \exp\left(\langle \hat{x}^{text}_{t,i}, \hat{x}^{text}_{t,j'} \rangle / \tau_{text}\right)}, \tag{25}$$

where $\hat{x}^{text}_{t,i}$ denotes normalized text features and τ_{text} is a temperature parameter.

The overall edge weight matrix combines geometry- and text-driven terms via a learnable gate:

$$W_t = (1-\lambda) W^{geo}_t + \lambda W^{text}_t, \quad \lambda = \sigma(\beta), \tag{26}$$

where $W_t \in \mathbb{R}^{O \times O}$ encodes both geometry-aware and text-guided connectivity, and λ is dynamically controlled by β, which allows our model to learn the distribution of semantic and geometric weights in different scenarios during training.

Once the structural graph is constructed, we perform multi-granularity multimodal fusion to ensure semantic consistency across modalities. Specifically, at the object level, we align each detected entity's visual embedding with its corresponding textual embedding and apply a gating mechanism to adaptively emphasize the more reliable modality. This design allows the model to preserve fine-grained semantic cues while remaining robust to noise or missing information in either modality. At the frame level, we further integrate global visual scene embeddings with sentence-level textual descriptions to provide a coherent context that guides object interactions. Such hierarchical fusion guarantees that both local interactions and global scene semantics are consistently represented.

For each object i at frame t, let $x^{vis}_{t,i} \in \mathbb{R}^F$ denote its visual embedding extracted from F_{vis}, and $\hat{x}^{text}_{t,i} \in \mathbb{R}^F$ denote the aligned textual embedding. We employ a gating mechanism to adaptively combine the two representations:

$$x^{fuse}_{t,i} = g\big([x^{vis}_{t,i}; \hat{x}^{text}_{t,i}]\big) \odot x^{vis}_{t,i} + (1 - g(\cdot)) \odot \hat{x}^{text}_{t,i}, \tag{27}$$

where

$$g(\cdot) = \sigma\big(W_g [x^{vis}_{t,i}; \hat{x}^{text}_{t,i}] + b_g\big), \tag{28}$$

$\sigma(\cdot)$ is the sigmoid activation, and $[\cdot;\cdot]$ denotes feature concatenation.

Similarly, we fuse frame-level features. Let f^{vis}_t be the global visual embedding from F_{vis}, and $\hat{f}^{text}_t$ be the sentence-level embedding from F_{text}. Their fusion is defined as

$$f^{fuse}_t = g\Big([f^{vis}_t; \hat{f}^{text}_t]\Big) \odot f^{vis}_t + (1 - g(\cdot)) \odot \hat{f}^{text}_t. \tag{29}$$

3.3.3 *Temporal modeling*

To jointly capture spatial and temporal dependencies, we propose a dynamic weighted graph convolutional network followed by multi-scale temporal encoders.

Given fused object-level features $X^{fuse}_t = [x^{fuse}_{t,1}, \ldots, x^{fuse}_{t,O}]$, the GCN layer updates node embeddings as

$$H^{(l+1)}_t = \sigma\Big(W_t \odot \tilde{A}_t H^{(l)}_t \Psi^{(l)}\Big), \tag{30}$$

where $H^{(0)}_t = X^{fuse}_t$, $\tilde{A}_t$ is the normalized adjacency, and $\Psi^{(l)}$ are learnable weights. After $L = 2$ layers, node embeddings are pooled:

$$u_t = \frac{1}{O} \sum_{i=1}^{O} H^{(L)}_{t,i}. \tag{31}$$

The pooled representation u_t is concatenated with f_t^{fuse} to form z_t. To enlarge temporal receptive fields, we apply a temporal convolutional network

$$\tilde{z}_{1:T} = \mathrm{TCN}(z_{1:T}), \tag{32}$$

followed by a GRU for sequential modeling

$$h_{1:T} = \mathrm{GRU}(\tilde{z}_{1:T}). \tag{33}$$

Frame-level predictions are computed as

$$\ell_t = W_2\,\phi(W_1 h_t + b_1) + b_2, \quad p_t = \mathrm{softmax}(\ell_t), \tag{34}$$

where $\ell_t \in \mathbb{R}^2$ are accident anticipation logits.

3.3.4 *Loss function*

The model is supervised with three complementary objectives.

(i) Frame-level anticipation loss. Positive samples are reweighted according to their distance from the time-of-accident τ, which encourages models to make predictions as early as possible and with greater confidence closer to the moment of an accident:

$$w_t^+ = \exp(-\max(0, \tau - t - 1/fps)), \tag{35}$$

and the frame-level anticipation loss is

$$L_1 = \frac{1}{BT}\sum_{b=1}^{B}\sum_{t=1}^{T}\Big(y\,w_t^+\,\mathrm{CE}(\ell_t, y) + (1-y)\,\mathrm{CE}(\ell_t, y)\Big). \tag{36}$$

(ii) Video-level pooling loss. From the perspective of the entire positive scene video, at least one frame in the first T_{pool} frames should be sufficient to determine the occurrence of an accident. Therefore, video-level classification loss is applied to improve the accuracy of the model's accident prediction:

$$L_2 = \frac{1}{B}\sum_{b=1}^{B}\mathrm{CE}\Big(\max_{t\leqslant T_{pool}}\ell_t, y\Big). \tag{37}$$

(iii) Semantic alignment loss. To enforce cross-modal semantic consistency, we impose a contrastive constraint between visual and textual representations. For each time step t, we take the global visual embedding f_t^{vis} and the corresponding sentence-level textual embedding $\hat{f}_t^{text}$, and apply a temperature-scaled, bidirectional information noise-contrastive estimation (InfoNCE) objective to pull the cross-modal positive pair at the same time step closer while pushing other samples/time steps in the mini-batch apart as negatives:

$$\begin{aligned}
L_{v\to t}(t) &= -\log\frac{\exp\big(\langle \bar{f}_t^{vis},\, \bar{f}_t^{text}\rangle/\tau_c\big)}{\sum_{(b',t')\in\mathcal{B}}\mathbf{m}_{(t),(b',t')}\exp\big(\langle \bar{f}_t^{vis},\, \bar{f}_{t'}^{text,(b')}\rangle/\tau_c\big)},\\
L_{t\to v}(t) &= -\log\frac{\exp\big(\langle \bar{f}_t^{text},\, \bar{f}_t^{vis}\rangle/\tau_c\big)}{\sum_{(b',t')\in\mathcal{B}}\mathbf{m}_{(t),(b',t')}\exp\big(\langle \bar{f}_t^{text},\, \bar{f}_{t'}^{vis,(b')}\rangle/\tau_c\big)},\\
L_3 &= \frac{1}{|T|}\sum_{t\in T}\frac{1}{2}\Big(L_{v\to t}(t) + L_{t\to v}(t)\Big).
\end{aligned} \tag{38}$$

For numerical stability, we perform ℓ_2 normalization before projection and use a lightweight MLP as the projection head; we also adopt a temporal-neighborhood mask to avoid treating adjacent frames from the same video as negatives. Serving as an orthogonal regularizer to the main task, this loss strengthens cross-modal alignment, encourages the object/frame-level fusion gates to learn more reliable evidence weighting, and stabilizes the text-driven edge weights in the graph, thereby improving the modeling of long-range dependencies and causal cues.

(iv) Final objective. The overall training loss is

$$L = L_1 + \gamma L_2 + L_3, \quad \gamma = T, \tag{39}$$

where L_1 is a frame-level loss, and L_2 is a video-level loss. The frame number T is used as an amplification factor to prevent L_2 from being overwhelmed by L_1. L_3 serves as an auxiliary loss to stabilize cross-modal alignment, achieving the complementary advantages of frame-level fine supervision, video-level early anticipation constraints, and cross-modal consistency.

Table 2 The detailed information of MAA and other existing datasets. mToA represents the mean start time of the accident.

Dataset	Videos	Positives	Negative	Total length (h)	FPS	mToA
DAD [11]	1750	620	1130	2.43	20	4.50
DADA-2000 [54]	2000	2000	–	6.09	30	–
CCD [12]	4500	1500	3000	6.25	10	3.82
CADP [55]	1416	1416	–	5.2	28	3.69
A3D [53]	1500	1500	–	2.08	20	3.37
MAA (Ours)	6000	4000	2000	8.33	10	3.16

Table 3 Comparison of models on DAD, A3D, CCD, MAA and Synthetic DAD (sDAD). **Bold** and underlined values represent the best and second-best performance. Instances where values are not available are marked with a dash ("–"). a) As shown in Table 2, the mToA of the CCD dataset is 3.82, so results larger than this value cannot be compared on the same scale.

Model	DAD		A3D		CCD		MAA		sDAD	
	AP (%)	mTTA (s)	AP (%)	mTTA (s)	AP (%)	mTTA (s)	AP (%)	mTTA (s)	AP (%)	mTTA (s)
DSA [11]	48.1	1.34	92.3	2.95	98.7	3.08	–	–	–	–
VideoMAE [59]	48.1	3.93	–	–	–	–	76.5	1.90	55.1	**4.02**
AdaLEA [60]	52.3	3.43	92.9	3.16	99.2	3.45	–	–	–	–
UString [12]	53.7	3.53	93.2	**3.24**	99.5	3.74	69.9	1.61	58.2	3.10
DSTA [13]	56.1	3.66	93.5	2.87	99.3	3.87[a)]	74.5	1.85	76.5	2.21
GSC [10]	60.4	2.55	94.9	2.62	99.4	3.68	–	–	–	–
AccNet [21]	60.8	3.58	94.9	2.62	99.5	**3.78**	73.1	1.69	73.8	2.55
VideoCLIP [47]	68.5	3.75	–	–	–	–	74.7	1.72	72.4	3.73
THAT-NET [20]	77.8	3.64	–	–	99.5	4.59[a)]	–	–	–	–
ViViT-FE [61]	78.8	4.03	–	–	–	–	75.8	1.97	82.1	3.86
Ours	**85.8**	**4.11**	**95.5**	3.20	**99.7**	3.68	**80.8**	**1.99**	**90.1**	3.71

4 Experiments

We assess the effectiveness of our pipeline on four real-world datasets: the Dashcam Accidents Dataset (DAD) [11], the AnAn Accident Detection (A3D) dataset [53], the Car Crash Dataset (CCD) [12], and the Multi-source Accident Anticipation (MAA) dataset introduced in this work. The detailed information of the existing datasets is shown in Table 2 [11, 12, 53–55].

The MAA dataset is introduced as a benchmark dedicated to accident anticipation, constructed by integrating multiple publicly available video resources such as D2-City, BDD100K, and DOTA [56–58]. MAA consists of 6000 clips, of which two-thirds are accident-positive and the rest are negative samples. To ensure fairness, we re-processed all clips to maintain consistent resolution, frame rate, and duration, and re-annotated the accident start time within a specific frame interval. The dataset is partitioned into 4500 samples for training and 1500 for testing. A detailed comparison between MAA and existing datasets is provided in Table 2. Compared to other datasets, MAA not only contains more videos but also offers greater coverage. Because MAA draws from multiple sources, its scenes span diverse geographic regions, including Asia, the Americas, and internet videos (while DAD is limited to a single region). Furthermore, MAA encompasses a wider range of climate and traffic conditions, along with richer annotations, providing a more comprehensive resource for accident prediction research. We have deployed several representative baseline models on MAA to provide performance references for future research.

4.1 Evaluation results

Table 3 [10–13, 20, 21, 47, 59–61] summarizes comparative results across DAD, A3D, CCD, and MAA using average precision (AP) and mean time-to-accident (mTTA) as the main evaluation indicators. On the DAD dataset, our method surpasses existing approaches with gains of 10.3% in AP and 12.3% in mTTA. For A3D and CCD, where AP values are close to saturation, our model still shows noticeable improvements.

For the MAA benchmark, we replicate multiple representative baselines, including advanced accident anticipation models based on RNNs and graph neural networks, as well as temporal reasoning frameworks based on Transformers, self-supervised learning, and multimodal models. Evaluation on MAA shows our method outperforming the runner-up by 6.5% AP and 1% mTTA, validating the advantage of the semantic-geometric graph design and positioning it as a new benchmark for accident anticipation.

We further investigate the impact of integrating synthetic data into the DAD benchmark. When 40% of the

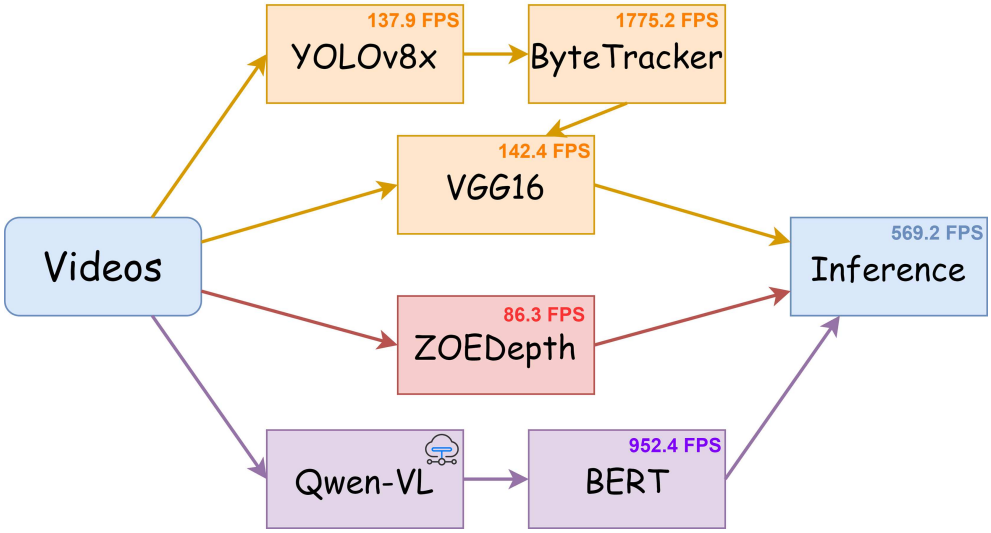


Figure 7 (Color online) The computational efficiency of the key modules within the overall framework. The efficiency of each module is quantified by the average number of video frames it can process per second.

training set is expanded with generated videos, most baselines show certain gains. For models with advanced performance, this improvement ranges from 4.2% to 5.0%. Additional trials include supplementing the dataset with 40% real traffic clips and, alternatively, substituting 40% of the original training data with generated samples. The reported results are AP = 90.9, mTTA = 4.02 for the former and AP = 83.8, mTTA = 3.40 for the latter. These findings indicate that, while synthetic augmentation contributes positively, its effectiveness remains inferior to augmenting with genuine traffic footage. Besides, when generated data directly replaces part of the original training set, performance deteriorates, indicating the presence of a measurable domain gap, which is expected when synthetic samples replace real data. Based on the experimental results, the performance difference between the two data sources ranges from 0.9% to 2.3%, with this gap further diminishing as the dataset size increases. Moreover, the performance discrepancy between synthetic and real data is substantially smaller than the performance gains achieved through the use of synthetic data. It is worth noting that when additional data are incorporated into the training set, the mTTA metric shows a moderate decline. We observe that the ratio of mTTA to mToA in the DAD dataset reaches 91.3%, whereas in the larger-scale MAA dataset this ratio drops to 63.0%. This result aligns with expectations: the expansion of data volume mitigates overfitting observed in smaller datasets, requiring the model to rely on richer contextual information for accident prediction rather than making rapid forecasts solely based on recurring visual patterns.

Considering practical deployment scenarios, we analyze the computational efficiency of the proposed framework. Under the experimental settings, the runtime of each module is shown in Figure 7. Given that Qwen-VL contains approximately 3.4 B parameters, on-device deployment is generally infeasible in real-world applications. Therefore, in our efficiency analysis, we assume an edge-cloud collaborative architecture, in which Qwen-VL is deployed in the cloud to accommodate its substantial computational demands. Moreover, because scene semantics evolve slowly across consecutive frames, performing VLM-based scene analysis once every 5 or 10 frames (i.e., every 0.1–0.2 s) is sufficient.

In the ideal steady-state pipeline, the throughput is determined by the slowest module (the depth estimation module), yielding a pipeline frame rate of $FPS_{\text{pipeline}} = 86.3$. The per-frame output latency corresponds to the cumulative time of the longest sequence of serial operations, which is approximately 16.6 ms. Overall, the depth estimation and VLM modules constitute the primary bottlenecks in the computational pipeline; however, these limitations can be effectively mitigated through cloud-based execution and do not pose significant obstacles to real-world deployment.

4.2 Ablation studies

Table 4 summarizes the ablation outcomes on the DAD dataset, highlighting how individual components such as the input representations, the GRU unit, the temporal convolutional module and the GCN layers contribute to overall performance. The results show that removing or replacing critical components consistently degrades inference accuracy. Among these modules, visual information, GCN, and GRU contribute most significantly to

Table 4 Ablation studies on the DAD dataset. AAM, TCN represent the adaptive adjacency matrix and temporal convolutional network.

Model	Input			GCN layers			Fuse		AAM	TCN	GRU	Input	
	Vision	Frame text	Object text	1	2	3	Gated	Concat				AP (%)	mTTA (s)
A	–	✓	✓	–	✓	–	–	–	✓	✓	✓	44.3	3.12
B	✓	–	✓	–	✓	–	✓	–	✓	✓	✓	84.4	3.51
C	✓	✓	–	–	✓	–	✓	–	✓	✓	✓	77.4	3.44
D	✓	–	–	–	✓	–	–	–	✓	✓	✓	76.1	3.72
E	✓	✓	✓	✓	–	–	✓	–	✓	✓	✓	83.2	3.68
F	✓	✓	✓	–	–	✓	✓	–	✓	✓	✓	84.3	3.61
G	✓	✓	✓	–	✓	–	–	✓	✓	✓	✓	81.9	3.98
H	✓	✓	✓	–	✓	–	✓	–	–	✓	✓	77.7	3.51
I	✓	✓	✓	–	✓	–	✓	–	✓	–	✓	78.4	3.85
J	✓	✓	✓	–	✓	–	✓	–	✓	✓	–	50.1	3.55
K	✓	✓	✓	–	–	–	✓	–	–	✓	✓	46.9	3.90
Final	✓	✓	✓	–	✓	–	✓	–	✓	✓	✓	85.8	4.11

performance, corresponding to primary information input, scene interaction modeling, and temporal reasoning, respectively. The optimal number of GCN layers is found to be 2, while gated fusion provides the most effective approach for information integration.

4.3 Qualitative analysis

To further evaluate the effectiveness of this pipeline, we visualize several traffic scenes from the MAA dataset and compare its performance against the second-best baseline. Figures 8(a)–(d) illustrate four challenging scenarios, where Figures 8(a)–(c) are positive accident scenes and Figure 8(d) is a negative case.

In Figure 8(a), the blue car in front initiates a rightward lane change while a white car drives straight at high speed along the right lane, ultimately colliding with the blue car without deceleration. From a geometric perspective, the relative speed is high and the distance between the two vehicles is short. From a semantic perspective, a yielding conflict arises between lane changing and straight-line driving. Our model rapidly increases accident confidence and correctly anticipates the accident, with confidence declining after the white car exits the view. By contrast, the baseline model, which relies more heavily on visual features, fails to predict the accident due to the white car's brief appearance in the camera's field of view.

In Figure 8(b), a heavy truck traveling near the right lane edge runs over a raised roadside curb with its front-right wheel, causing a minor accident. Since no vehicle-to-vehicle interaction occurs and the curb feature is visually obscured due to snow coverage, neither our model nor the baseline successfully anticipates the accident beforehand. Both models only trigger warnings after detecting changes in the truck's pose.

Figure 8(c) depicts a rear-end collision while the ego vehicle is waiting at a red light. Because the colliding vehicle is not visible, both models lack access to critical cues and fail to predict the accident in advance. Notably, in both Figures 8(b) and (c), abnormal post-accident vehicle pose changes cause a rise in accident confidence, demonstrating the models' sensitivity to traffic anomalies.

Finally, in Figure 8(d), after the traffic light turns green, the white car begins moving forward while the car ahead remains stationary. The distance decreases, relative speed increases, and semantically there appears to be a rear-end risk. In reality, however, the distance remains safe but is obscured by partial occlusion. Our model mistakenly raises an accident warning, while the baseline remains at a moderate-to-low confidence level in the multi-vehicle scenario and is not misled.

5 Conclusion

This research presents an accident anticipation pipeline that leverages a dynamic graph convolutional network enriched with geometric and semantic cues. To support systematic evaluation, we also introduce the multi-source accident anticipation dataset, which contains diverse accident scenarios with comprehensive annotations. In addition, we design a world model-based video augmentation pipeline that generates synthetic driving scenes and verify its benefit in enhancing anticipation accuracy. Extensive experiments have proven that our method is superior to previous methods on several benchmarks, setting new performance standards. Importantly, our analysis highlights that synthetic data, although still imperfect, provide measurable benefits for targeted tasks. Looking ahead, future

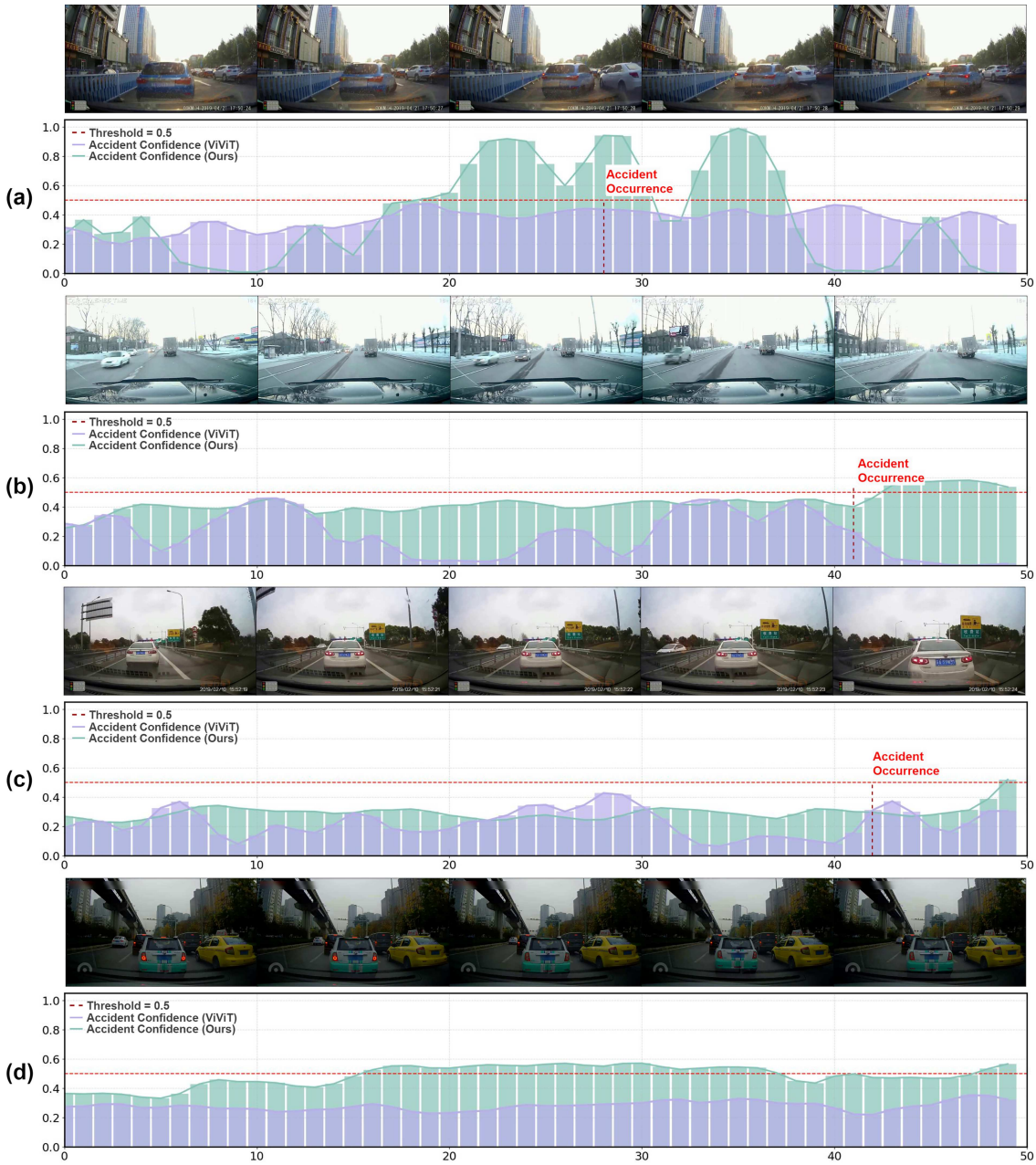


Figure 8 (Color online) Visualization of model performance in some scenes. (a) Successful anticipation of a positive scene; (b) failed anticipation of a positive scene; (c) failed anticipation of another positive scene; (d) failed anticipation of a negative scene.

work should emphasize improving the realism of generated content, particularly in terms of object consistency, noise suppression, and scene fidelity.

Acknowledgements This work was supported by Science and Technology Development Fund of Macao (Grant Nos. 0007/2025/RIC, 0122/2024/RIB2, 0215/2024/AGJ, 0074/2025/AMJ, 001/2024/SKL), Research Services and Knowledge Transfer Office, University of Macau (Grant Nos. SRG2023-00037-IOTSC, MYRG-GRG2024-00284-IOTSC), Shenzhen-Hong Kong-Macao Science and Technology Program Category C (Grant No. SGDX20230821095159012), Science and Technology Planning Project of Guangdong (Grant No. 2025A0505010016), National Natural Science Foundation of China (Grant No. 52572354), State Key Lab of Intelligent Transportation System (Grant No. 2024-B001), and Jiangsu Provincial Science and Technology Program (Grant No. BZ2024055).

Data availability The data used in this work are publicly released at https://github.com/humanlabmembers/Multi-source-Accident-Anticipation.

References

1 Bagloee S A, Tavana M, Asadi M, et al. Autonomous vehicles: challenges, opportunities, and future implications for transportation policies. J Mod Transp, 2016, 24: 284–303

2 Kalra N, Paddock S M. Driving to safety: how many miles of driving would it take to demonstrate autonomous vehicle reliability? Transp Res Part A Policy Pract, 2016, 94: 182–193
3 Fang J, Qiao J, Xue J, et al. Vision-based traffic accident detection and anticipation: a survey. IEEE Trans Circuits Syst Video Technol, 2024, 34: 1983–1999
4 Jiang Y, Zhang X, Xu X, et al. Event-triggered shared lateral control for safe-maneuver of intelligent vehicles. Sci China Inf Sci, 2021, 64: 172203
5 Liu J, Guo H, Song L, et al. Driver-automation shared steering control for highly automated vehicles. Sci China Inf Sci, 2020, 63: 190201
6 Wang L, Cheng K, Lei S, et al. Dc-Gaussian: Improving 3d Gaussian splatting for reflective dash cam videos. 2024. ArXiv:2405.17705
7 Guan Y, Liao H, Wang C, et al. Domain-enhanced dual-branch model for efficient and interpretable accident anticipation. Commun Transp Res, 2025, 5: 100214
8 Cheong J Y, Simon C, Kim C S, et al. Reflection removal under fast forward camera motion. IEEE Trans Image Process, 2017, 26: 6061–6073
9 Sadeky S, Al-Hamadiy A, Michaelisy B, et al. Real-time automatic traffic accident recognition using HFG. In: Proceedings of 2010 20th International Conference on Pattern Recognition, 2010. 3348–3351
10 Wang T, Chen K, Chen G, et al. GSC: a graph and spatio-temporal continuity based framework for accident anticipation. IEEE Trans Intell Veh, 2023, 9: 2249–2261
11 Chan F H, Chen Y T, Xiang Y, et al. Anticipating accidents in dashcam videos. In: Proceedings of the 13th Asian Conference on Computer Vision, Taipei, 2017. 136–153
12 Bao W, Yu Q, Kong Y. Uncertainty-based traffic accident anticipation with spatio-temporal relational learning. In: Proceedings of ACM Multimedia Conference, 2020
13 Karim M M, Li Y, Qin R, et al. A dynamic spatial-temporal attention network for early anticipation of traffic accidents. IEEE Trans Intell Transp Syst, 2022, 23: 9590–9600
14 Fatima M, Khan M U K, Kyung C M. Global feature aggregation for accident anticipation. In: Proceedings of 2020 25th International Conference on Pattern Recognition (ICPR), 2021. 2809–2816
15 Song W, Li S, Chang T, et al. Dynamic attention augmented graph network for video accident anticipation. Pattern Recognition, 2024, 147: 110071
16 Chen S, Dong J, Ha P Y J, et al. Graph neural network and reinforcement learning for multi-agent cooperative control of connected autonomous vehicles. Comput-Aided Civil Infrastructure Eng, 2021, 36: 838–857
17 Ou J, Xia J, Wang Y, et al. A data-driven approach to determining freeway incident impact areas with fuzzy and graph theory-based clustering. Comput-Aided Civil Infrastructure Eng, 2020, 35: 178–199
18 Liao H, Li Y, Wang C, et al. When, where, and what? A benchmark for accident anticipation and localization with large language models. In: Proceedings of the 32nd ACM International Conference on Multimedia, 2024. 8–17
19 Thakur N, Gouripeddi P, Li B. Graph (graph): a nested graph-based framework for early accident anticipation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2024. 7533–7541
20 Liu W, Zhang T, Lu Y, et al. THAT-Net: two-layer hidden state aggregation based two-stream network for traffic accident prediction. Inf Sci, 2023, 634: 744–760
21 Liao H, Li Y, Li Z, et al. Real-time accident anticipation for autonomous driving through monocular depth-enhanced 3D modeling. Accident Anal Prevention, 2024, 207: 107760
22 Liang M, Yang B, Hu R, et al. Learning lane graph representations for motion forecasting. In: Proceedings of European Conference on Computer Vision, 2020. 541–556
23 Mohamed A, Qian K, Elhoseiny M, et al. Social-STGCNN: a social spatio-temporal graph convolutional neural network for human trajectory prediction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020. 14424–14432
24 Gesnouin J, Pechberti S, Stanciulescu B, et al. Assessing cross-dataset generalization of pedestrian crossing predictors. In: Proceedings of 2022 IEEE Intelligent Vehicles Symposium (IV), 2022. 419–426
25 Martin M, Roitberg A, Haurilet M, et al. Drive&act: a multi-modal dataset for fine-grained driver behavior recognition in autonomous vehicles. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019. 2801–2810
26 Wu Z, Pan S, Chen F, et al. A comprehensive survey on graph neural networks. IEEE Trans Neural Netw Learn Syst, 2020, 32: 4–24
27 Xiao D, Dianati M, Geiger W G, et al. Review of graph-based hazardous event detection methods for autonomous driving systems. IEEE Trans Intell Transp Syst, 2023, 24: 4697–4715
28 Zhang J, Huang J, Jin S, et al. Vision-language models for vision tasks: a survey. IEEE Trans Pattern Anal Mach Intell, 2024, 46: 5625–5644
29 Radford A, Kim J W, Hallacy C, et al. Learning transferable visual models from natural language supervision. In: Proceedings of International Conference on Machine Learning, 2021. 8748–8763
30 Liao H, Shen H, Li Z, et al. GPT-4 enhanced multimodal grounding for autonomous driving: leveraging cross-modal attention with large language models. Commun Transp Res, 2024, 4: 100116
31 Lohner A, Compagno F, Francis J, et al. Enhancing vision-language models with scene graphs for traffic accident understanding. In: Proceedings of 2024 IEEE International Automated Vehicle Validation Conference (IAVVC), 2024. 1–7
32 Gao P, Geng S, Zhang R, et al. CLIP-adapter: better vision-language models with feature adapters. Int J Comput Vis, 2024, 132: 581–595
33 Fang H, Xiong P, Xu L, et al. Clip2video: mastering video-text retrieval via image clip. 2021. ArXiv:2106.11097
34 Zellers R, Lu X, Hessel J, et al. Merlot: multimodal neural script knowledge models. In: Proceedings of Advances in Neural Information Processing Systems, 2021. 34: 23634–23651
35 Bar-Tal O, Ofri-Amar D, Fridman R, et al. Text2live: text-driven layered image and video editing. In: Proceedings of European Conference on Computer Vision, 2022. 707–723
36 Wendland P, Birkenbihl C, Gomez-Freixa M, et al. Generation of realistic synthetic data using multimodal neural ordinary differential equations. npj Digit Med, 2022, 5: 122
37 Bai J, Bai S, Chu Y, et al. Qwen technical report. 2023. ArXiv:2309.16609
38 Shinar D. Accident/crash causation and analysis. In: Traffic Safety and Human Behavior. Bognor Regis: Emerald Publishing Limited, 2017. 1039–1083
39 Wang H, Yu Y, Yuan Q. Application of Dijkstra algorithm in robot path-planning. In: Proceedings of 2011 Second International Conference on Mechanic Automation and Control Engineering, 2011. 1067–1069
40 Hu Y, Yang J, Chen L, et al. Planning-oriented autonomous driving. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023. 17853–17862
41 Caesar H, Bankiti V, Lang A H, et al. nuScenes: a multimodal dataset for autonomous driving. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020. 11621–11631
42 Li C, Zhou K, Liu T, et al. Avd2: Accident video diffusion for accident video description. 2025. ArXiv:2502.14801
43 Obukhov A, Krasnyanskiy M. Quality assessment method for GAN based on modified metrics inception score and fréchet inception distance. In: Proceedings of the Computational Methods in Systems and Software, 2020. 102–114
44 Ge S, Mahapatra A, Parmar G, et al. On the content bias in fréchet video distance. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024. 7277–7288
45 Korhonen J, You J. Peak signal-to-noise ratio revisited: Is simple beautiful? In: Proceedings of 2012 Fourth International Workshop on

Quality of Multimedia Experience, 2012. 37–38
46 Brunet D, Vrscay E R, Zhou Wang E R. On the mathematical properties of the structural similarity index. IEEE Trans Image Process, 2011, 21: 1488–1499
47 Xu H, Ghosh G, Huang P Y, et al. VideoCLIP: contrastive pre-training for zero-shot video-text understanding. 2021. ArXiv:2109.14084
48 Sohan M, Sai Ram T, Rami Reddy C V. A review on YOLOv8 and its advancements. In: Proceedings of International Conference on Data Intelligence and Cognitive Informatics, 2024. 529–545
49 Zhang Y, Sun P, Jiang Y, et al. ByteTrack: multi-object tracking by associating every detection box. In: Proceedings of European Conference on Computer Vision, 2022. 1–21
50 Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. 2014. ArXiv:1409.1556
51 Bhat S F, Birkl R, Wofk D, et al. ZoeDepth: zero-shot transfer by combining relative and metric depth. 2023. ArXiv:2302.12288
52 Aftan S, Shah H. A survey on BERT and its applications. In: Proceedings of 2023 20th Learning and Technology Conference (L&T), 2023. 161–166
53 Yao Y, Xu M, Wang Y, et al. Unsupervised traffic accident detection in first-person videos. In: Proceedings of 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2019. 273–280
54 Fang J, Yan D, Qiao J, et al. DADA-2000: can driving accident be predicted by driver attention analyzed by a benchmark. In: Proceedings of 2019 IEEE Intelligent Transportation Systems Conference (ITSC), 2019. 4303–4309
55 Shah A P, Lamare J B, Nguyen-Anh T, et al. CADP: a novel dataset for CCTV traffic camera based accident analysis. In: Proceedings of 2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), 2018. 1–9
56 Yao Y, Wang X, Xu M, et al. DoTA: unsupervised detection of traffic anomaly in driving videos. IEEE Trans Pattern Anal Mach Intell, 2022, 45: 444–459
57 Che Z, Li G, Li T, et al. D^2-City: a large-scale dashcam video dataset of diverse traffic scenarios. 2019. ArXiv:1904.01975
58 Yu F, Chen H, Wang X, et al. BDD100K: a diverse driving dataset for heterogeneous multitask learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020. 2636–2645
59 Tong Z, Song Y, Wang J, et al. VideoMAE: masked autoencoders are data-efficient learners for self-supervised video pre-training. In: Proceedings of Advances in Neural Information Processing Systems, 2022. 35: 10078–10093
60 Suzuki T, Kataoka H, Aoki Y, et al. Anticipating traffic accidents with adaptive loss and large-scale incident db. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018. 3521–3529
61 Arnab A, Dehghani M, Heigold G, et al. ViViT: a video vision transformer. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021. 6836–6846