Causal-Entity Reflected Egocentric Traffic Accident Video Synthesis

Lei-Lei Li[1], Jianwu Fang[1†], Junbin Xiao[2], Shanmin Pang[1], Hongkai Yu[4], Chen Lv[3], Jianru Xue[1], and Tat-Seng Chua[2]

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

http://lotvsmmau.net/Causal-VidSyn *

Abstract

Egocentricly comprehending the causes and effects of car accidents is crucial for the safety of self-driving cars, and synthesizing causal-entity reflected accident videos can facilitate the capability test to respond to unaffordable accidents in reality. However, incorporating causal relations as seen in real-world videos into synthetic videos remains challenging. This work argues that precisely identifying the accident participants and capturing their related behaviors are of critical importance. In this regard, we propose a novel diffusion model Causal-VidSyn for synthesizing egocentric traffic accident videos. To enable causal entity grounding in video diffusion, Causal-VidSyn leverages the cause descriptions and driver fixations to identify the accident participants and behaviors, facilitated by accident reason answering and gaze-conditioned selection modules. To support Causal-VidSyn, we further construct ***Drive-Gaze****, the largest driver gaze dataset (with* ***1.54M*** *frames of fixations) in driving accident scenarios. Extensive experiments show that Causal-VidSyn surpasses state-of-the-art video diffusion models in terms of frame quality and causal sensitivity in various tasks, including accident video editing, normal-to-accident video diffusion, and text-to-video generation.*

1. Introduction

The emerging fully self-driving (FSD) technique [65], while bringing convenience to our daily life, has also led to various ethical, trustworthy, and economic disputes in handling car accidents [51]. Therefore, an egocentric comprehension of the car accident is paramount not only for improving self-driving safety but also for disambiguating accident responsibility. Yet, the scarcity of egocentric accident data severely hinders research in this field.

With the tremendous advancements in video diffusion models [11, 14, 24, 25, 32, 81, 83], synthetic videos could be a promising solution for data scarcity. However, the state-of-the-art (SOTA) video diffusion models are developed for common video generation; they often fail to generate egocentric accident videos that achieve a causal-entity reflected synthesis [48, 82]. As shown in Fig. 1, when facing a counterfactual text edit from "pedestrian" to "motorbike" collision, one of the SOTA model Abductive-OAVD [17] fails to generate the motorbike. While CogVideoX [82] responds to the needed motorbike, the requested collision is not reflected.

Figure 1. We show the inability of two state-of-the-art diffusion models (*i.e.*, Abductive-OAVD [17] and CogVideoX-2B [82]) for editing accident video content. Abductive-OAVD cannot generate the needed motorbike, while CogVideoX fails to reflect the collision situation. Our Causal-VidSyn accurately generates the collision motorbike and maintains the background scenes.

To overcome such issues, precisely identifying the causal entities (or objects) in accidents and capturing their accident-related behaviors is the key to success. However, accident scenarios often involve tiny objects in fast scene changes, which makes it extremely difficult to identify the objects, especially in the ego view, not to mention analyzing their nuanced behaviors related to the accidents. In this paper, we highlight the incorporation of two critical information cues: accident reason-collision descriptions and driver gaze fixations. The reason-collision descriptions contain rich information about the accidents, including the major participants and their misbehavior that resulted in the accidents. Despite being informative, finding the right reason for the accident is still challenging. Also, there is a modality gap between textual descriptions to visual accident appearances. As a remedy, driver gaze fixations provide direct visual attention to the accident regions, since human drivers can perceive the road hazard incisively based on their driving experience.

To effectively exploit such information for egocentric accident video diffusion, we propose the Causal-VidSyn, which makes the video diffusion backbone (*e.g.*, 3D-Unet) causal-grounded, facilitated by 1) an accident reason answer-

*†Corresponding author.

ing (ArA) module designed to retrieve the right accident reason and incorporate it into noise representation learning, and 2) a driver gaze-conditioned visual token selection mechanism to focus on causal participant regions. Additionally, to enhance diffusion learning in fast scene changes, we additionally learn to contrast forward and backward time order frame diffusion with reciprocal text and vision prompts, as creating a counterfactual intervention to forward text and visual prompts, via the exogenous noise e to help the causal scene learning associating reciprocal frame and text prompts.

Causal-VidSyn fulfills the causal-entity reflected video synthesis from ❶ internal diffusion recipe level and ❷ external knowledge level. Diffusion recipe level fulfills a reciprocal prompted frame diffusion (RPFD, §4.1). The knowledge level designs *causal-prone*[1] token selection blocks (**CTS**) and an ArA-guided causal token grounding block (**CTG**) (§4.2) to make the 3D-Unet backbone causal-grounded. Additionally, to support Causal-VidSyn, we construct the largest driver gaze dataset **Drive-Gaze**, which collects over 1.54 **million** frames of driver fixations for 9,727 accident scenarios. Notably, ArA and driver fixations are only involved in the training phase, and the testing phase only inputs the video or text prompts for accident video diffusion.

Extensive evaluations are carried out for three video diffusion tasks 1) accident video content editing on accident dataset (**DADA-2000** [16]), 2) normal-to-accident video diffusion on accident-free dataset (**BDD-A** [76]), and 3) text-to-video accident generation. The results show that Causal-VidSyn can surpass many state-of-the-art methods with higher frame quality, fidelity, and causal sensitivity. We also extend **CTS** and **CTG** to SOTA Transformer-based video diffusion models (*i.e.*, CogVideoX-2B [82] and Latte [48]), and demonstrate consistently significant improvements.

2. Related Work

2.1. Approaches for Ameliorating Diffusion Models

Enhancing Content Consistency. Video diffusion models have dominated the field of video generation in recent years [3, 27, 47, 68, 71, 72, 78] and take large efforts to overcome a knotty problem, *i.e.*, maintaining temporal logic and content consistency [3, 84, 89], such as StoryDiffusion [89], CogVideoX [82], T2V-Turbo-v2 [35], etc. Various temporal alignment modules are modeled to ensure consistency across video frames in [3, 10, 23, 57, 77, 84, 88].

Causality Discovery. Causal relations are natural in the real world [37, 43, 69, 80, 85]. Based on the discussion on causality and grounding [86], causality prefers a time-different grounding between the cause and effect elements. Recent causal diffusion models mainly focus on the counterfactual (what-if issue) estimation [33, 59] or contrastive causal-relation discovery [30, 60], *e.g.*, forward noising the causal relations of latent variables step-by-step [49, 59]. Compared to the domain-extraneous ways, considering driving scene knowledge is preferred in this work.

Considering Driving Scene Knowledge. Video diffusion in driving scenes has garnered widespread attention [20, 26, 41, 73, 79] for safe driving by frame or view synthesis. The multi-view panoramic consistency within captured videos, such as in DriverDreamer [70], DrivingDiffusion [38], Panacea [73], *etc.*, involve the road topology (*e.g.*, bird's eye view (BEV) [50, 54, 63]) and object locations to correlate the ego-scene relation in multi-view driving video diffusion. Nonetheless, current video diffusion models mainly encounter accident-free scenarios, while the knowledge of on-road accident videos is commonly absent. Recent works explore the accident video synthesis [17, 22, 34], while they mainly focus on the text-to-video generation, and do not explore the causality in video conditioned synthesis.

Table 1. Attribute comparison of available driver gaze datasets.

Datasets	Years	#Clips	#Frames	CoT-F	S-C	Sub.num	TA	R/S	Cites
3DDS [5]	2011	-	18K	in-lab	NOM	10		S	56
BDD-A [76]	2018	1,232	378K	in-lab	CR	45		R	181
DADA-2000 [15, 16]	2019	2,000	658K	in-lab	Acc	20		R	244
DR(eye)VE [53]	2019	74	555K	real-d	NOM	8		R	342
DGAZE [13]	2020	20	100K	in-lab	NOM	20		R	21
TrafficGaze [12]	2020	16	75K	in-lab	NOM	28		R	117
MAAD [21]	2021	8	60K	in-lab	NOM	23		R	13
Eye-car [1]	2021	21	31.5K	in-lab	Acc	20		R	60
PSAD [18]	2021	2,724	797K	in-lab	Acc	6		R	10
RainyGaze [66]	2022	16	81K	in-lab	NOM	30		R	16
CoCAtt [61]	2022	-	88K	in-lab	NOM	11		S	13
LBW [31]	2022	-	123K	real-d	NOM	28		R	27
DPoG [52]	2024	-	15.3K	real-d	NOM	11		R	1
Drive-Gaze	2025	**9,727**	**1.54 million**	in-lab	Acc	10	✓	R	-

CoT-F: collection form (in-lab or real driving (real-d)), **S-C**: scene categories (NOM-normal, CR-critical, and Acc-accident), **TA**: text annotations, **Sub.num**: subject number, **ReS**: frame resolution, **R/S**: real or synthetic videos. **Cites**: Google citations up to Mar. 07, 2025.

2.2. Driver Gaze Datasets for Driving Videos

Tab. 1 presents the attribute analysis for available driver gaze datasets. Among them, BDD-A [76], DADA-2000 [15], and DR(eye)VE [53] are the top-3 most popular ones for critical, accident, and normal driving scenarios, respectively. DR(eye)VE [53], LBW [31], and DPoG [31] collect the gaze data in the real driving process. Most datasets focus on normal driving scenes. DADA-2000 [15], Eye-car [1], PSAD [18], and our Drive-Gaze concentrate on the driving accident scenarios. Drive-Gaze is the largest driver gaze dataset and owns the text description for accident reasons and collision descriptions.

3. Drive-Gaze Dataset

The data source of Drive-Gaze stems from the recently released multimodal egocentric accident video dataset, *i.e.*, MM-AU [17], which annotates the accident reason, collision type, and accident prevention descriptions for 11,727 accident videos. We find that DADA-2000 [15] is MM-AU's

[1] Causal-prone tokens mean that the selected tokens prefer better causal-sensitive text-vision alignment than non-selected ones.

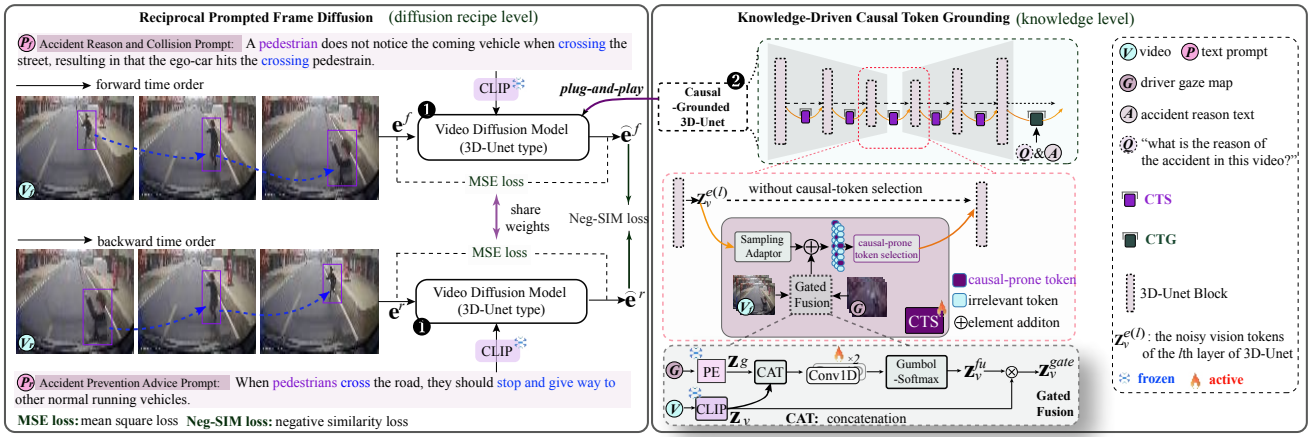

Figure 2. The training schema of **Causal-VidSyn**. It mainly includes three stages: The direct optimization of $\boldsymbol{e}^f$ with the forward time order diffusion (Stage-**0**, omitted in this figure), where the pre-trained Stable Diffusion [58] is used for model initialization. The reciprocal prompted frame diffusion (RPFD) in Stage-❶ contrasts two diffusion pathways conditioned by reciprocal time order vision frames (V_f&V_r) and semantically reciprocal text prompt (P_f&P_r). We take the 3D-Unet backbone [17] in the noise ($\boldsymbol{e}^f$&$\boldsymbol{e}^r$) representation learning at this stage. State-❷ injects *causal-prone token selection blocks* (**CTS**) and a *causal token grounding* (**CTG**) block respectively into the inner layers and end layer of the 3D-Unet module to fulfill a causal-grounded video diffusion helped by accident reason (Q) answering (A) and gaze (G)-conditioned token selection modules.

subset, having the driver fixation data already. We collect the diver gaze data for the remainder accident videos. We have ten subjects (4 females and 6 males aged from 21 to 26 years old with over two years of driver's licenses) to collect the driver fixations on 1.54 million frames of 9,727 videos (frame resolution: 1280×720) in three months. We use the desk-mounted eye-tracker Tobii Pro Fusion (250 Hz) to collect the driver fixations with well-calibrated eye vision.

▸**Annotation Details**: To avoid the eye fatigue of each subject, we divide 9,727 videos into 83 long videos with about 10 minutes each. In addition, we group similar accident types into one long video as much as possible to allow the subject to accumulate experience for capturing the dangerous object better. Each subject can only annotate one long video within one hour, and each long video is watched by all subjects, where at least one hour is maintained for resting the eyes for the next time watching. To obtain the gaze map, similar to DADA-2000 [15], each frame's final gaze map is obtained using a Gaussian filter (50×50 pixel kernel) to convolute all fixation points of subjects. To match the frame rate, we accumulate the fixation points in 250Hz frames of each subject to 30Hz frames.

▸**Utilization Ways**: We take all frames in Drive-Gaze for training use. Therefore, the video clips sampled in the diffusion process have pair-wise driver gaze maps and video frames. Additionally, Drive-Gaze can facilitate many egocentric traffic accident video understanding tasks, such as driver attention prediction [9, 19]; cognitive driving accident anticipation, like DRIVE [2] and CogTAA [36]; scanpath prediction in accident reason answering, like [8], and other gaze map involved accident video understanding tasks.

4. Methodology

The essence of Causal-VidSyn is to fulfill a causal-grounded video content learning when providing certain accident-related text descriptions. To begin with, we analyze the video diffusion process in egocentric accident scenarios.

For an accident video clip V coupling with an accident reason-collision description P, a noise representation $\boldsymbol{e}\sim\mathcal{N}(0,\boldsymbol{I})$ is leveraged for the video diffusion process. The *forward process* $q(\boldsymbol{z}_v^{e_k}|\boldsymbol{z}_v^0)$ gradually adds $\boldsymbol{e}$ to $\boldsymbol{z}_v^0$ and generates the sequential-noisy latent vision variables $\boldsymbol{z}_v^0, \boldsymbol{z}_v^{e_2}, \ldots, \boldsymbol{z}_v^{e_k}$ (k: diffusion step index), where

$$\boldsymbol{z}_v^{e_k} = \sqrt{\hat{\beta}_k}\boldsymbol{z}_v^0 + \sqrt{1-\hat{\beta}_k}\boldsymbol{e}, \hat{\beta}_k = \prod_{i=1}^{k}\alpha_i, \alpha_k = 1-\beta_k, \quad (1)$$

β_k is a parameterized schedule, and $\boldsymbol{z}_v^0$ is the pure video embedding of V. The *reverse process* $q(\boldsymbol{z}_v^{e_{k-1}}|\boldsymbol{z}_v^{e_k})$ recovers $\tilde{\boldsymbol{z}}_v^0$ by the following sampler step-by-step:

$$\tilde{\boldsymbol{z}}_v^{e_{k-1}} = \mu_{\hat{\boldsymbol{e}}}(\boldsymbol{z}_v^{e_k}, k) + \sigma_k \boldsymbol{e}, \quad (2)$$

where $\mu_{\hat{\boldsymbol{e}}}(\boldsymbol{z}_v^{e_k}, k) = \frac{1}{\sqrt{\hat{\beta}_k}}\left(\boldsymbol{z}_v^{e_k} - \frac{\beta_k}{\sqrt{1-\hat{\beta}_k}}\phi_{\hat{\boldsymbol{e}}}(k, \boldsymbol{z}_v^{e_k}, P)\right)$, and σ_k is the noise variance at the k^{th} step (k is omitted in following for simplicity). Let's denote $\tilde{V}$ as the generated clip by a trained video diffusion model $\phi_{\hat{\boldsymbol{e}}}$ conditioned by $\{V, P, \hat{\boldsymbol{e}}\}$. The goal is to denoise the causal relations of latent variables $\boldsymbol{z}_v^e$ step-by-step from V to $\tilde{V}$ and identify the causal-grounded (CG) latent vision representation $\boldsymbol{z}_v^c : \boldsymbol{z}_v^e|_{\text{CG}}$, responding well to the counterfactual change of text prompt P, optimized by the mean square loss (MSE):

$$\min_{\hat{\boldsymbol{e}}, \boldsymbol{z}_v^c} : \mathbb{E}_{\boldsymbol{e}\sim\mathcal{N}(0,\boldsymbol{I})}\|\boldsymbol{e} - \phi_{\hat{\boldsymbol{e}}}(k, \boldsymbol{z}_v^e, P)\|_2^2. \quad (3)$$

Manifestly, learning the causal entity within accident videos (*i.e.*, reasoning $\boldsymbol{z}_v^c$ associated with P) for video diffusion models is challenging because of the tiny and sudden change of causal entities. Hence, as shown in Fig. 2, this work formulates Causal-VidSyn as two progressive levels:

❶ **Diffusion recipe level** fulfills a contrastive forward and backward time order frame diffusion conditioned by semantically reciprocal text descriptions (*abbrev., reciprocal prompted frame diffusion*), *i.e.*, **forward**: accident reason and collision prompt; **backward**: accident prevention advice prompt. This recipe prefers to enhance the causal scene learning by global text-vision prompt intervention.

❷**Knowledge level** reforms the backbone (*e.g.*, 3D-Unet) to be causal-grounded in causal token learning explicitly helped by accident reason answering head (ArA) and driver gaze map fusion with local token intervention.

4.1. Reciprocal Prompted Frame Diffusion

Reciprocal prompted frame diffusion (RPFD) (stage-❶)is inspired by the greater attractiveness of the context of the critical scene or crash-prone objects than stable background scenes in egocentric accident videos [15]. We leverage the hypothesis [17] that the accident prevention text prompt can coincide with the accident dissipation process reflected by backward time order accident frames. Differently, we argue that the backward-order diffusion can be treated as creating a counterfactual intervention to forward text and visual prompts, and the exogenous noise $\boldsymbol{e}$ helps the causal scene associate reciprocal frame and text prompts.

With the reciprocal text and frame prompts, **different** text prompts should activate **different** visual content mainly associated with the causal entity in accident videos. Consequently, we contrast noise representation learning as

$$\mathcal{L}_{\text{ST1}} = \mathcal{L}_{\text{MSE}}(\boldsymbol{e}_f, \hat{\boldsymbol{e}}_f) + \mathcal{L}_{\text{MSE}}(\mathbf{e}_r, \hat{\boldsymbol{e}}_r) + \lambda \mathcal{L}_{\text{NS}}(\hat{\boldsymbol{e}}_f, \hat{\boldsymbol{e}}_r),$$
$$\mathcal{L}_{\text{NS}}(\hat{\boldsymbol{e}}_f, \hat{\boldsymbol{e}}_r) = \mathbb{E}\left(1 - \frac{\hat{\boldsymbol{e}}_f \cdot \hat{\boldsymbol{e}}_r}{||\hat{\boldsymbol{e}}_f|| ||\hat{\boldsymbol{e}}_r||}\right), \tag{4}$$

where $\hat{\mathbf{e}}_f$ and $\hat{\mathbf{e}}_r$ denote the reconstructed noise representation in contrastive two diffusion pathways with shared weights. $\mathcal{L}_{\text{NS}}$ is the negative similarity loss (inverse cosine similarity) to contrastively enhance the different visual content learning in RPFD conditioned by forward prompt (V_f,P_f) and backward prompt (V_r,P_r). $\mathcal{L}_{\text{MSE}}$ takes the form of Eq. 3, and λ=0.2 is a hyperparameter to balance the losses.

4.2. Knowledge-Driven Causal Token Grounding

Reciprocal prompted frame diffusion (RPFD) aims to suppress the influence of scene background on the frame-level diffusion recipe. To find the causal-entity reflected regions, we further reform the diffusion backbone to be hierarchically causal-grounded in the Stage-❷.

As shown in Fig. 2 and Fig. 3, the multi-layer 3D-Unet module in the diffusion model has multi-interleaved attention blocks, where we design **CTS** and **CTG** as flexible

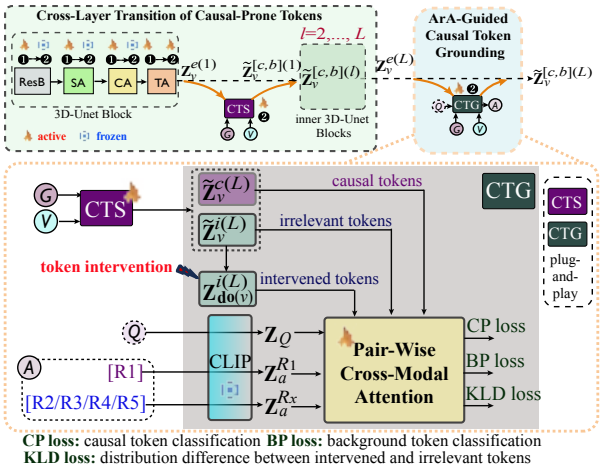

Figure 3. The details of Cross-Layer Causal-Prone Token Transition and ArA-Guided Causal Token Grounding in the 3D-Unet backbone, where the **CTG** block is illustrated. When **CTS** and **CTG** are removed in inference, $\tilde{\boldsymbol{z}}_v^{[c,b](l)}$=$\boldsymbol{z}_v^{e(l)}$, marked by "- - →".

blocks that can be plug-and-play injected in the inner layers or the end layer of the 3D-Unet [17] structured by residual block (**ResB**), spatial (**SA**), cross-modal (**CA**), and temporal attention (**TA**) blocks. Concisely, we take **CTS** as a bridge to fulfill cross-layer *causal-prone* token transition and involve an ArA head at **CTG** block to fulfill ArA-guided causal token grounding. Because of the sudden scene change in egocentric accident videos, we introduce the driver gaze to help the causal-prone token selection in **CTS**. Therefore, it is worth mentioning that **CTS** and **CTG** blocks can be seen as *powerful guider* to refine causal tokens layer-by-layer, which are only used in the training phase and removed in the inference stage.

Gating Allocation of Driver Gazes (Gated Fusion). As illustrated in Fig. 3, each **CTS** receives the tokenized driver gaze map $\boldsymbol{z}_g$, visual tokens $\mathbf{z}_v$ of video frames V, and the noisy vision representation $\boldsymbol{z}_v^{e(l)}$ from previous temporal attention (TA) in each diffusion step, where l indexes the inner layer of 3D-Unet module. Vision tokens $\boldsymbol{z}_v$ are encoded by a pre-trained frozen CLIP model [56]. Because of the sparse distribution of driver fixations, driver gaze tokens $\boldsymbol{z}_g$ are obtained by only the position embedding (PE) layer in the same CLIP model to avoid the zero-value issue in CLIP's deep layers. Here, the PE layer is fulfilled by a 2D convolution (*kernel size:* 14×14). To avoid the influence of gaze bias in the collection, we provide a gated fusion of $\boldsymbol{z}_g$ and $\boldsymbol{z}_v$ by

$$\boldsymbol{z}_v^{fu} = \text{Gumbel-Softmax}(\text{Conv1Ds}(\text{cat}(\boldsymbol{z}_v, \boldsymbol{z}_g))),$$
$$\boldsymbol{z}_v^{gate} = \boldsymbol{z}_v \otimes \boldsymbol{z}_v^{fu}, \text{'}\otimes - \texttt{element multiplication}, \tag{5}$$

where Conv1Ds(.) denotes two layers of 1D-convolution with *relu* operation for reducing the token channel dimension after concatenation (*i.e.*, cat(,,)) of vision and gaze tokens. Gumbel-Softmax(.) ensures the values of $\boldsymbol{z}_v^{fu}$ at each token

dimension summing to 1 for gated token selection [29].

Causal-Prone Token Selection (CTS) and Transition. In Stage-❷, the noisy vision representation $\mathbf{z}_v^e$ passes through multiple layers of the 3D-Unet module interleaved with SA, TA, and CA blocks. Therefore, reasoning the causal tokens $\boldsymbol{z}_v^c$ in $\boldsymbol{z}_v^e$ from the deep layer structure of 3D-Unet module is challenging. Hence, **CTS** blocks can be vividly understood as pulling out the causal tokens to be grounded layer by layer, where gaze maps are adopted to strengthen the pulling force by gated allocation.

Cross-Layer Token Sampling Adaptor: As shown in Fig. 3, we inject **CTS** block after each TA layer, helped by the temporal token transition ability. Assume the noisy vision representation outputted by the temporal attention (TA) layer is $\boldsymbol{z}_v^{e(l)}$ at the l^{th} ($l = 1, 2, ..., L$) inner layer of the 3D-Unet module, which is firstly processed through a *sampling adaptor* for resizing it to adapt to the cross-layer Unet-scale change and match the CLIP output dimension [56] adopted in gated fusion $\boldsymbol{z}_v^{gate}$, achieved by a bilinear interpolation (BintP) and a Conv2D (1×1) operation as

$$\tilde{\boldsymbol{z}}_v^{e(l)} = \text{Conv2D}(\text{BintP}(\boldsymbol{z}_v^{e(l)})), \quad \hat{\boldsymbol{z}}_v^{\text{CP}(l)} = \tilde{\boldsymbol{z}}_v^{e(l)} \oplus \boldsymbol{z}_v^{gate}, \tag{6}$$

where $\hat{\boldsymbol{z}}_v^{\text{CP}(l)}$ is the element addition $\oplus$ of $\tilde{\boldsymbol{z}}_v^{e(l)}$ and $\boldsymbol{z}_v^{gate}$.

Causal-Prone Token Selection: We compute the token importance score encoded from each video frame by *softmax*(MLP($\hat{\boldsymbol{z}}_v^{\text{CP}(l)}$)) with multilayers of perceptrons (MLP) and select tokens with top-d scores, where we set d empirically as a quarter of tokens within the single frame as the casual-prone tokens $\tilde{\boldsymbol{z}}_v^{c(l)}$ based on the long-tailed region distribution of causal objects and won't miss too many causal tokens (Here we omit the frame index for simplicity). The remaining ones are treated as the background tokens $\tilde{\boldsymbol{z}}_v^{b(l)}$. The combination $\tilde{\boldsymbol{z}}_v^{[c,b](l)}$ of $\tilde{\boldsymbol{z}}_v^{c(l)}$ and $\tilde{\boldsymbol{z}}_v^{b(l)}$ is fed into the ResB block in the next 3D-Unet block (l<L). If we encounter the L^{th} layer, $\boldsymbol{z}_v^{e(L)}$ is grounded by the following ArA head.

ArA-Guided Causal Token Grounding (CTG). ArA-guided causal token grounding leverages the insight of the VideoQA task [4, 7, 46], while differently we design **CTG** block in token-level grounding and prefer an ArA-guided video diffusion. The answers are multi-choice with the accurate one [R_1], and four disturbing ones [R_2, R_3, R_4, R_5] corresponding to the egocentric accident video clip [V].

As illustrated in Fig. 3, if we receive $\boldsymbol{z}_v^{e(L)}$, it also involves the gate-conditioned **CTS** block and obtains $\tilde{\boldsymbol{z}}_v^{c(L)}$ and $\tilde{\boldsymbol{z}}_v^{b(L)}$. Differently, we take a token intervention on $\tilde{\boldsymbol{z}}_v^{b(L)}$ as $\tilde{\boldsymbol{z}}_{\boldsymbol{do}(v)}^{b(L)}$ by randomly masking a quarter of tokens to noise. With these token representations, the same cross-modal attention (CA) in 3D-Unet is adopted to classify the causal tokens and background tokens aligned by the text tokens of the accurate reason-question pair ($\boldsymbol{z}_a^{R_1}$,$\boldsymbol{z}_Q$) and disturbing accident reason-question token pairs ($\boldsymbol{z}_a^{R_{x(x\neq 1)}}, \boldsymbol{z}_Q$).

We employ cross-entropy loss (XE) [39, 40] for the causal and background token classification. Notably, the XE for background tokens follows a uniform distribution $\mathcal{U}(0, 1)$ over all irrelevant answer candidates. The distribution difference between the intervened tokens $\tilde{\boldsymbol{z}}_{\boldsymbol{do}(v)}^{b(L)}$ and $\tilde{\boldsymbol{z}}_v^{b(L)}$ is measured by the Kullback-Leibler Divergence (KLD) [40]. Thus, the loss function of ArA grounding is

$$\mathcal{L}_{\text{ArA}} = \mathcal{L}_{\text{XE}}^{c} + \mathcal{L}_{\text{XE}}^{b} + \mathcal{L}_{\text{KLD}}^{\boldsymbol{do}(b)}, \tag{7}$$

and $\mathcal{L}_{\text{KLD}}^{\boldsymbol{do}(b)} = \text{KLD}(\text{CA}(\boldsymbol{z}_a^{R_x}|\tilde{\boldsymbol{z}}_v^{b(L)}, \boldsymbol{z}_Q), \text{CA}(\boldsymbol{z}_a^{R_x}|\tilde{\boldsymbol{z}}_{\boldsymbol{do}(v)}^{b(L)}, \boldsymbol{z}_Q))$.

4.3. Training and Inference

‣**Training Recipe**: Stage-**0**, Stage-❶, and Stage-❷ are progressively trained, and we use the same MSE loss in Eq. 3 for noise representation learning. In Stage-**0**, only the forward time order video diffusion (*i.e.*, only with $\boldsymbol{e}^f$) is conducted on the 3D-Unet backbone and trained for 10,000 training steps. Stage-❶ re-loads the pre-trained parameters in Stage-**0**, and fine-tunes the 3D-Unet module with contrastive noise learning (*i.e.*, $\boldsymbol{e}^f$ and $\boldsymbol{e}^r$) by 10,000 steps, and then the trained parameters are reloaded in Stage-❷ and further trained 10,000 steps after adding **CTS** and **CTG** modules. In Stage-❷, the total loss involved the ArA head is:

$$\mathcal{L}_{\text{ST2}} = \mathcal{L}_{\text{MSE}}(\boldsymbol{e}^f, \hat{\boldsymbol{e}}^f) + \gamma \mathcal{L}_{\text{ArA}}, \tag{8}$$

where $\boldsymbol{e}^f \sim \mathcal{N}(0, \boldsymbol{I})$ is randomly re-initialized and γ=0.3 balances the weight of two terms.

‣**Inference**: We provide the video-to-video (V2V) and text-to-video (T2V) synthesizing choices. Notably, **CTS** and **CTG** blocks are plug-and-play, which are removed in the inference stage, *i.e.*, once two levels of fine-tuning ❶ and ❷ are completed, the ArA head and driver gaze maps are removed. In the V2V mode, we can implement accident content editing and normal-to-accident video diffusion conditioned by the video clip input (V) and text prompt (P). T2V mode starts from an initial random noise and generates the expected video frames conditioned by the text prompt (P). In the reverse diffusion process, we adopt the common DDIM scheduler [62] to decode the video frames.

5. Experiments

5.1. Tasks and Datasets

We extensively evaluate the performance by three promising video generation tasks in testing: 1) normal-to-accident video diffusion (**N2A**), 2) accident video content editing (**AEdit**), and 3) text-to-accident video generation (**T2V**). Because the accident commonly appears suddenly with a very short time window [17], we generate 16 frames to show the near-crash to crashing (NC-2-C) process concisely.

‣**N2A** is a first-launched task in this field, which can be used to create labels for the detection of crash-prone objects and

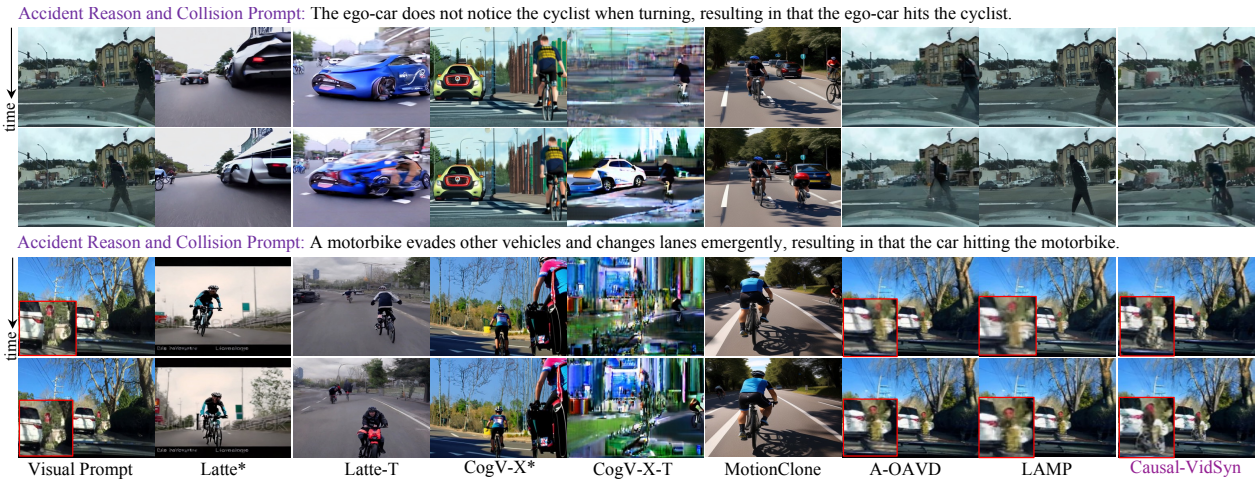

Figure 4. Sample visualizations of N2A task by Latte* [48], Latte-T [48], CogV-X* [82], CogV-X-T [82], MotionClone [42], A-OAVD [17], LAMP [75], and our Causal-VidSyn (Best viewed in zoom mode).

accident anticipation on accident-free AV testing platforms, *e.g.,* nuScenes [6] and accelerate self-driving car testing. This work chooses a safety-critical dataset BDD-A [76] to evaluate whether the critical scene context is aligned with the accident-related texts. We sample the end 16 frames (appearing critical objects) of 2,000 videos as visual prompts.

‣**AEdit** task checks the causal-sensitivity by tiny text changes (*e.g.*, "*pedestrian*"→ "*cyclist*") reflecting the causal-entities. In this task, the background scenes are preferred to be unchanged. In this task, we adopt the DADA-2000 dataset [16], which also provides the driver fixations for checking whether the noticed objects aligned with text prompts are edited. We sample 3,000 clips in DADA-2000 from the NC-2-C frame windows in inference.

‣**T2V** is well-known and we generate 500 clips conditioned by accident reason-collision text prompts in DADA-2000.

‣**Training Dataset**. We sample 6,492 clips from the NC-2-C windows in 9,727 accident videos of **MM-AU** [17] with pair-wise accident reason-collision and prevention text descriptions, randomly sampled from the near-crash to crashing frame windows to maintain diversity.

5.2. Implementation Details

All the experiments of Causal-VidSyn and other Unet-based diffusion models are conducted using two NVIDIA RTX3090 GPUs with each 22GB RAM. The frame resolution is 224×224. The learning rate of each stage is $1e-5$ using the Adam solver with β_1=0.9 and β_2=0.999. We also involve Diffusion Transformer (DiT)-based methods for comparison. Because of the large frame resolution (CogVideoX-2B (CogV-X) [82]: 720×480; Latte [48]: 512×512) in diffusion, they are trained on one NVIDIA A800 GPU with 80GB RAM. The number of diffusion steps k in each stage is set to $1,000$. The batch size for training Causal-VidSyn is 2, and the layers L of the 3D-Unet module is 12 with a symmetric structure.

Methods	**N2A** (BDD-A [76] (2000))			**T2V** (500)		Backbone
	CLIP$_s$↑	FVD↓	TempC↑	CLIP$_s$↑	TempC↑	
T2V-zero [32]$_{\text{CVPR2023}}$	26.0	11754.8	0.992	24.8	0.929	Unet
Free-bloom [28]$_{\text{NeurIPS2024}}$	25.1	8280.1	0.990	22.1	0.841	Unet
CoVideo [87]$_{\text{ICLR2024}}$	24.2	9906.1	0.930	-	-	Unet
Latte* [48]$_{\text{Arxiv2024}}$	25.8	11066.2	0.993	22.8	0.977	DiT
CogV-X* [82]$_{\text{ICLR2025}}$	25.6	9075.6	0.992	26.0	**0.994**	DiT
MotionClone [42]$_{\text{ICLR2025}}$	24.9	11554.9	0.994	-	-	Unet
TAV [74]$_{\text{ICCV2023}}$	24.1	9305.3	0.902	23.5	0.820	Unet
A-OAVD [17]$_{\text{CVPR2024}}$	25.8	6378.9	0.992	26.5	0.977	Unet
LAMP [75]$_{\text{CVPR2024}}$	25.4	6208.2	0.991	-	-	Unet
Latte-T [48]$_{\text{Arxiv2024}}$	25.3	11196.6	0.993	25.6	0.977	DiT
CogV-X-T [82]$_{\text{ICLR2025}}$	24.6	10094.0	0.993	25.3	0.981	DiT
AVD2 [34]$_{\text{ICRA2025}}$	-	-	-	27.3	0.976	DiT
w/o [G+**CTS**&**CTG**+RPFD]	25.8	6378.9	0.992	26.5	0.976	Unet
w/o [G+**CTS**&**CTG**]	25.0	7225.5	**0.997**	26.4	0.949	Unet
w/o [G]	26.0	6249.3	0.992	27.4	0.945	Unet
Causal-VidSyn [Full Train]	**26.5**	**6192.3**	0.994	**27.5**	0.944	Unet

The numbers in the brackets denote the sample scale in inference.

Table 2. Performance on N2A and T2V tasks (**bold** font: the best).

‣**Metrics**. Following the popular models [68, 82], we evaluate the frame quality by Fréchet video distance (**FVD** [67]), causal-sensitivity by clip score (**CLIP**$_s$) [17, 78], frame content consistency by temporal consistency (**TempC** [55]).

‣**Competitors**. In comparison, we choose many SOTA video diffusion models because of their training efficiency and their ability to generate high-quality videos. They are divided as training-free methods, *i.e.*, ControlVideo (*Abbrev.*, CoVideo) [87], T2V-zero [32], Free-bloom [28], Latte* [48], CogVideoX*(*Abbrev.*, CogV-X*) [82], and MotionClone [42], and training methods including Tune-A-Video (*Abbrev.*, TAV) [74], Abductive-OAVD (*Abbrev.*, A-OAVD) [17], LAMP [75] and the trained version of Latte-T and CogVideoX-T (*abbrev.,* CogV-X-T). The training-free ones solely leverage their pre-trained models and directly infer the generation, and the training ones are fine-tuned 10,000 training steps with their official setting for involving more egocentric accident knowledge for fair comparison, via the forward time order frame diffusion by the same training set as our Causal-VidSyn. Besides, we further involve a new work, AVD2 [34], targeting accident video synthesis (trained

Methods	AEdit (DADA [16] (3000))			
	CLIP$_s$↑	FVD↓	TempC↑	**Afd**↑ (%)
TAV [74]$_{ICCV2023}$	23.8	10076.2	0.909	-
Latte-T [48]$_{Arxiv2024}$	28.2	12377.3	0.945	-
CogV-X-T [82]$_{ICLR2025}$	25.3	11420.5	0.945	-
A-OAVD [17] $_{CVPR2024}$	26.9	5358.2	0.947	49.4
LAMP [75]$_{CVPR2024}$	26.1	6191.6	**0.971**	34.7
w/o [G+**CTS**&**CTG**+RPFD]	26.9	5358.8	0.947	47.1
w/o [G+**CTS**&**CTG**]	28.6	6374.8	0.942	50.3
w/o [G]	28.6	5368.2	0.939	50.4
Causal-VidSyn [Full Train]	**28.7**	**5352.9**	0.940	**55.4**

Table 3. The video generation performance of SOTA video diffusion models on the AEdit task (**bold** font: the best).

Figure 5. **Afd** is the ratio of IOU$(,) > 0$ of all checks. IOU: the intersection over union of two bounding boxes.

by MM-AU dataset [17]) in T2V evaluation.

5.3. Main Results on N2A and T2A Tasks

N2A Task. In this task, A-OAVD [17], CogV-X* [82], and our Causal-VidSyn are top three solutions showing better text-vision alignment than others, as shown in Tab. 2. We also visualize the generated frames in Fig. 4, and we can see that only our Causal-VidSyn can generate the critical cyclist while maintaining the frame style (the 1^{st} example in Fig. 4). In addition, our Causal-VidSyn can generate an expected collision while maintaining the background reflected in the 2^{nd} sample of Fig. 4. State-of-the-art methods, *i.e.*, Latte* [48], CogV-X* [82], and MotionClone [42] generate clear frames with active response, while the frame background is not maintained with irrelevant object styles. After involving more accident knowledge in CogV-X-T (CogV-X*→CogV-X-T), it generates artifacts mainly because more accident knowledge is required for fitting its large-scale DiT-based parameters. Notably, the objects in the generated frames of CogV-X*, Latte-T, and MotionClone are very large (named as **large object issue**), which causes large but unreasonable CLIP$_s$ values.

T2V Task. Because there is no video frame reference, the T2V task here is to evaluate the ability for semantic alignment in text-to-video generation, and the quantitative results are focused here. From Tab. 2, our Causal-VidSyn is the best for semantic alignment, and CogV-X* generates the best temporal consistency while the CLIP$_s$ value has a large gap to our model. Qualitative visualizations on the T2V task are shown in the supplementary file.

5.4. Main Results on AEdit Task

To measure the portions of cases where the expected objects are identified and edited in the video frames (causal-entity reflected), we introduce a new metric, affordance (**Afd**), that stems from Add-it image editor [64]. We present Fig. 5 to display the computing way of Afd. Notably, we follow [64] and adopt the GroundDINO [45] to fulfill the text-guided object detection, while we utilize the gazed regions instead of the object bounding boxes to match a human-preferable causal-entity editing checking based on the positive perception of critical objects by human attention [15, 44].

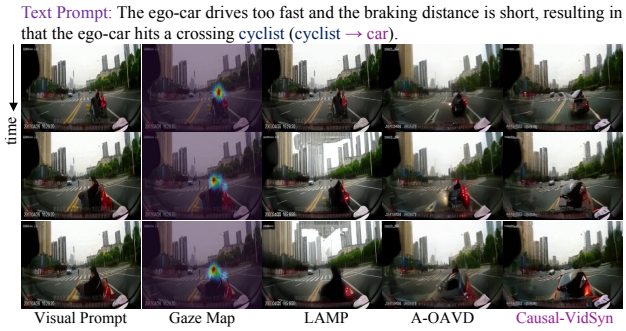

Figure 6. We visualize AEdit results of one crossing situation by LAMP [75], A-OAVD [17], and our Causal-VidSyn.

Based on the comparison in the N2A task, we can see that only LAMP, A-OAVD, and our Causal-VidSyn can maintain the frame style and the background scene context. Although Latte-T generates a high CLIP$_s$ (28.2) in the AEdit task, it is mainly caused by the *large object issue*, as shown by the results in Tab. 3. In this task, we visualize one example of DADA-2000 in Fig. 6, where we change the "cyclist" to "car" in the text prompt and display the driver gaze maps to show the attentive objects. We can see that A-OAVD and our model show active response while the car shape is clearer in Causal-VidSyn. LAMP [75] with a motion-consistency constraint can maintain the background style while it generates distorted and blurred content. From the **Afd** scores in Tab. 3, our Causal-VidSyn can identify the fixed and crash-prone objects better and show stronger causal-sensitive object editing ability than LAMP and A-OAVD.

5.5. Diagnostic Experiments

Roles of RPFD (Eq. 4), CTS, CTG, and Gaze Maps. We evaluate different causal-aware modules in Stage-❶ and ❷. We gradually dismantle the gaze maps (G), **CTS**&**CTG**, and RPFD in the fine-tuning phase and re-train the model with the same training dataset. Tab. 2 and Tab. 3 show the ablation results for N2A-T2V tasks and the AEdit task, respectively. From these results, all components are positive and RPFD seems to play a significant role in the AEdit task claimed by the increase of CLIP$_s$ value. Removing **CTS**&**CTG** displays a significant degradation. The N2A sample (creating a "crossing cyclist") in Fig. 7 further verifies the positive role of **CTS**&**CTG**. Gaze maps can localize the critical objects, but the accident causes can better explain their behaviors.

Portability of CTS&CTG. We also graft the causal-aware modules to fine-tune three other SOTA methods, *i.e.,* Unet-based TAV, DiT-based Latte, and CogV-X. TAV follows the

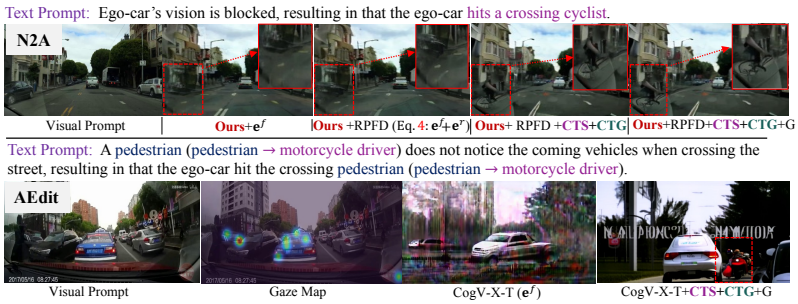

Figure 7. Visualizations of one N2A example and an AEdit sample with the ablation checking of Causal-VidSyn and CogV-X-T [82].

Methods	AEdit		T2V	Tun.-Params
	$CLIP_s$↑	FVD↓	$CLIP_s$↑	
TAV [74] (Unet)	23.8	10076.2	23.5	-
+RPFD	25.7	8701.4	23.7	0.24B (26%)
+RPFD+**CTS**(-G) &**CTG**	27.5	8694.3	24.6	0.24B (26%)
+RPFD+**CTS**(+G) &**CTG**	**27.9**	**8567.5**	**25.4**	0.24B (26%)
Latte-T [48] (DiT)	28.2	12377.3	25.6	-
+**CTS**(+G) &**CTG**	**28.5**	**11316.9**	**26.3**	1.06B(100%)
CogV-X-T [82] (DiT)	25.3	11420.5	25.3	-
+**CTS**(+G) &**CTG**	**28.2**	**10892.9**	**29.2**	0.05B (2.9%)

Tun.-Params: fine-tuned parameters (rate%).

Table 4. The diagnostic evaluation of TAV [74], Latte-T [48], and CogV-X-T [82] on **AEdit** and **T2V** tasks, with the fine-tuning by different causal-aware modules.

Methods	**N2A** (BDD-A(2000))		**AEdit** (DADA (3000))		
	$CLIP_s$↑	FVD↓	$CLIP_s$↑	FVD↓	**Afd**↑
A-OAVD [17]	25.8	6378.9	26.9	5358.2	49.4
Causal-VidSyn [Full Train]	**26.5**	**6192.3**	**28.7**	**5352.9**	**55.4**
Downscale Layers w/o **CTS**	26.2	6449.6	28.3	5476.4	49.9
Upscale Layers w/o **CTS**	26.3	6408.9	28.3	5449.4	54.9
with only **CTG**	26.0	6512.9	26.9	5507.8	50.2

Table 5. The ablation studies by our Causal-VidSyn ,*w.r.t.*, different injection layers of **CTS** on the 3D-Unet module.

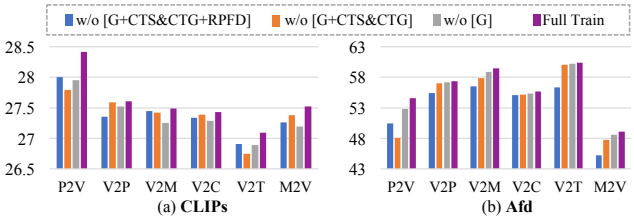

Figure 8. The $CLIP_s$ and Afd values of Causal-VidSyn, *w.r.t.*, fine-grained causal-entity editing from "pedestrian" to "vehicle" (P2V), "vehicle" to "pedestrian" (V2P), "vehicle" to "motorcycle/motorbike" (V2M), "vehicle" to "cyclist" (V2C), "vehicle to "truck" (V2T), and "motorcycle/motorbike" to "vehicle" (M2V).

same fine-tuning stages of Causal-VidSyn. For DiT-based ones, because they do not have ResB block and the Unet-scale change, the **CTS** block is taken as an extra bridge between previous TA and next SA block, and the **CTG** block is also attached at the end of DiT structure (see more details in the supplement). Then, Latte-T and CogV-T are further fine-tuned with 20,000 steps. We gradually add each module and check the performance gains. Tab. 4 demonstrates the effectiveness of the **CTS**&**CTG** clearly, especially for the AEdit task by CogV-X-T ($CLIP_s$ +2.9). Notably, although CogV-T generates artifacts as analyzed in Fig. 4, **CTS**&**CTG** can fine-tune it to recover the content structure clearly, as shown by the AEdit example in Fig. 7.

▸**Roles of Different Injection Layers of CTS**. As aforementioned, **CTS** and **CTG** can be understood as pulling out the causal tokens to be grounded layer by layer. Therefore, we offer more analysis for which part of the layers in the 3D-Unet module is dominant for selecting the causal tokens. We remove the **CTS** blocks from the downscale layers and upscale layers in the training phase, respectively. They are named as "Downscale Layers w/o **CTS**" and "Upscale Layers w/o **CTS**". In addition, we also maintain the final **CTG** only and remove all **CTS** blocks in the inner 3D-Unet layers. From Tab. 5, we can observe that removing **CTS** in downscale layers has more degree of performance degradation, which indicates that the **CTS** is more important in the downscale layers than upscale layers. This observation verifies: *deeper is better for inserting the **CTS** key*.

▸**Fine-Grained Causal-Entity Editing**. We present Fig. 8 for the analysis of the fine-grained causal-entity editing with the same ablation versions of Causal-VidSyn in Tab. 2 and Tab. 3. From the statistics, it is clear that the driver gaze helps find causal-entity reflected by the sudden decreases of $CLIP_s$ scores when removing driver gaze assistance. In addition, for the "vehicle" to "truck" (V2T) group, the Afd values are larger than other groups, which is mainly caused by the large scale of the truck and can easily obtain a large IOU score (Fig. 5). Additionally, the $CLIP_s$ values show inconsistent gain when removing different causal-aware modules on different groups. We think that this is caused by the text-vision semantic alignment contributed by the unmatched behaviors (not changed) of newly converted object types (different objects with distinct motion patterns). However, compared with the [Full Train] version, the causal-aware modules are positive for Causal-VidSyn.

6. Conclusions

This work presents a new egocentric traffic accident video diffusion model Causal-VidSyn, which enables the causal grounding in video diffusion via leveraging the cause descriptions and driver fixations to identify the accident participants and behaviors facilitated by accident reason answering and gaze-conditioned token selection modules (**CTS**&**CTG**). The extensive experiments and analysis show the designed **CTS**&**CTG** are plug-and-play with clear effectiveness and portability and make Causal-VidSyn with better frame quality and causal-sensitivity for N2A, AEdit, and T2V tasks. In the future, we will investigate video diffusion for other accident understanding tasks, such as accident anticipation.

Acknowledgment: This work is supported by NSFC (W2411052 and 62273057), and the Outstanding Youth Foundation of Shaanxi Province (2025JC-JCQN-092).

References

[1] Sonia Baee, Erfan Pakdamanian, Inki Kim, Lu Feng, Vicente Ordonez, and Laura E. Barnes. MEDIRL: predicting the visual attention of drivers via maximum entropy deep inverse reinforcement learning. In *ICCV*, pages 13158–13168, 2021. 2

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

[3] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *CVPR*, pages 22563–22575, 2023. 2

[4] Lukas Blübaum and Stefan Heindorf. Causal question answering with reinforcement learning. In *WWW*, pages 2204–2215, 2024. 5

[5] Ali Borji, Dicky N. Sihite, and Laurent Itti. Computational modeling of top-down visual attention in interactive environments. In *BMVC*, pages 1–12, 2011. 2

[6] Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *CVPR*, pages 11618–11628, 2020. 6

[7] Long Chen, Yuhang Zheng, Yulei Niu, Hanwang Zhang, and Jun Xiao. Counterfactual samples synthesizing and training for robust visual question answering. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(11):13218–13234, 2023. 5

[8] Xianyu Chen, Ming Jiang, and Qi Zhao. Predicting human scanpaths in visual question answering. In *CVPR*, pages 10876–10885, 2021. 3

[9] Yilong Chen, Zhixiong Nan, and Tao Xiang. Fblnet: Feedback loop network for driver attention prediction. In *ICCV*, pages 13325–13334, 2023. 3

[10] Zhikai Chen, Fuchen Long, Zhaofan Qiu, Ting Yao, Wengang Zhou, Jiebo Luo, and Tao Mei. Learning spatial adaptation and temporal coherence in diffusion models for video super-resolution. In *CVPR*, pages 9232–9241, 2024. 2

[11] Ernie Chu, Tzuhsuan Huang, Shuo-Yen Lin, and Jun-Cheng Chen. Medm: Mediating image diffusion models for video-to-video translation with temporal correspondence guidance. In *AAAI*, pages 1353–1361, 2024. 1

[12] Tao Deng, Hongmei Yan, Long Qin, Thuyen Ngo, and B. S. Manjunath. How do drivers allocate their potential attention? driving fixation prediction via convolutional neural networks. *IEEE Trans. Intell. Transp. Syst.*, 21(5):2146–2154, 2020. 2

[13] Isha Dua, Thrupthi Ann John, Riya Gupta, and C. V. Jawahar. DGAZE: driver gaze mapping on road. In *IROS*, pages 5946–5953, 2020. 2

[14] Patrick Esser, Johnathan Chiu, Parmida Atighehchian, Jonathan Granskog, and Anastasis Germanidis. Structure and content-guided video synthesis with diffusion models. In *ICCV*, pages 7312–7322, 2023. 1

[15] Jianwu Fang, Dingxin Yan, Jiahuan Qiao, Jianru Xue, He Wang, and Sen Li. DADA-2000: can driving accident be predicted by driver attention? analyzed by A benchmark. In *ITSC*, pages 4303–4309, 2019. 2, 3, 4, 7

[16] Jianwu Fang, Dingxin Yan, Jiahuan Qiao, Jianru Xue, and Hongkai Yu. DADA: driver attention prediction in driving accident scenarios. *IEEE Trans. Intell. Transp. Syst.*, 23(6): 4959–4971, 2022. 2, 6, 7

[17] Jianwu Fang, Lei-Lei Li, Junfei Zhou, Junbin Xiao, Hongkai Yu, Chen Lv, Jianru Xue, and Tat-Seng Chua. Abductive ego-view accident video understanding for safe driving perception. In *CVPR*, pages 22030–22040, 2024. 1, 2, 3, 4, 5, 6, 7, 8

[18] Shun Gan, Quan Li, Qingfan Wang, WenTao Chen, Detong Qin, and Bingbing Nie. Constructing personalized situation awareness dataset for hazard perception, comprehension, projection, and action of drivers. In *ITSC*, pages 1697–1704, 2021. 2

[19] Shun Gan, Xizhe Pei, Yulong Ge, Qingfan Wang, Shi Shang, Shengbo Eben Li, and Bingbing Nie. Multisource adaption for driver attention prediction in arbitrary driving scenes. *IEEE Trans. Intell. Transp. Syst.*, 23(11):20912 – 20925, 2022. 3

[20] Ruiyuan Gao, Kai Chen, Enze Xie, Lanqing Hong, Zhenguo Li, Dit-Yan Yeung, and Qiang Xu. Magicdrive: Street view generation with diverse 3D geometry control. In *ICLR*, 2024. 2

[21] Deepak Gopinath, Guy Rosman, Simon Stent, Katsuya Terahata, Luke Fletcher, Brenna Argall, and John Leonard. Maad: A model and dataset for" attended awareness" in driving. In *ICCV*, pages 3426–3436, 2021. 2

[22] Zipeng Guo, Yuchen Zhou, and Chao Gou. Drivinggen: Efficient safety-critical driving video generation with latent diffusion models. In *ICME*, pages 1–6, 2024. 2

[23] Sai Sree Harsha, Ambareesh Revanur, Dhwanit Agarwal, and Shradha Agrawal. Genvideo: One-shot target-image and shape aware video editing using t2i diffusion models. pages 7559–7568, 2024. 2

[24] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *NeurIPS*, 33:6840–6851, 2020. 1

[25] Jonathan Ho, Tim Salimans, Alexey A. Gritsenko, William Chan, Mohammad Norouzi, and David J. Fleet. Video diffusion models. In *NeurIPS*, 2022. 1

[26] Anthony Hu, Lloyd Russell, Hudson Yeo, Zak Murez, George Fedoseev, Alex Kendall, Jamie Shotton, and Gianluca Corrado. Gaia-1: A generative world model for autonomous driving. *arXiv preprint arXiv:2309.17080*, 2023. 2

[27] Runyi Hu, Jie Zhang, Yiming Li, Jiwei Li, Qing Guo, Han Qiu, and Tianwei Zhang. Videoshield: Regulating diffusion-based video generation models via watermarking. In *ICLR*, 2025. 2

[28] Hanzhuo Huang, Yufan Feng, Cheng Shi, Lan Xu, Jingyi Yu, and Sibei Yang. Free-bloom: Zero-shot text-to-video generator with llm director and ldm animator. *NeurIPS*, 36, 2024. 6, 4

[29] Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. In *ICLR*, 2017. 5

[30] Amir Mohammad Karimi-Mamaghan, Andrea Dittadi, Stefan Bauer, Karl Henrik Johansson, and Francesco Quinzan. Diffusion-based causal representation learning. *Entropy*, 26 (7):556, 2024. 2

[31] Isaac Kasahara, Simon Stent, and Hyun Soo Park. Look both ways: Self-supervising driver gaze estimation and road scene saliency. In *ECCV*, pages 126–142, 2022. 2

[32] Levon Khachatryan et al. Text2video-zero: Text-to-image diffusion models are zero-shot video generators. In *ICCV*, pages 15908–15918, 2023. 1, 6, 4

[33] Aneesh Komanduri, Chen Zhao, Feng Chen, and Xintao Wu. Causal diffusion autoencoders: Toward counterfactual generation via diffusion probabilistic models. In *CVPR Workshop*, 2024. 2

[34] Cheng Li, Keyuan Zhou, Tong Liu, Yu Wang, Mingqiao Zhuang, Huan-ang Gao, Bu Jin, and Hao Zhao. Avd2: Accident video diffusion for accident video description. In *ICRA*, 2025. 2, 6, 4

[35] Jiachen Li, Qian Long, Jian Zheng, Xiaofeng Gao, Robinson Piramuthu, Wenhu Chen, and William Yang Wang. T2v-turbo-v2: Enhancing video generation model post-training through data, reward, and conditional guidance design. In *ICLR*, 2025. 2

[36] Lei-Lei Li, Jianwu Fang, and Jianru Xue. Cognitive traffic accident anticipation. *IEEE Intell. Transp. Syst. Mag.*, 16(5): 17–32, 2024. 3, 6

[37] Xiaochuan Li, Baoyu Fan, Runze Zhang, Liang Jin, Di Wang, Zhenhua Guo, Yaqian Zhao, and Rengang Li. Image content generation with causal reasoning. In *AAAI*, pages 13646–13654, 2024. 2

[38] Xiaofan Li, Yifu Zhang, and Xiaoqing Ye. Drivingdiffusion: Layout-guided multi-view driving scenarios video generation with latent diffusion model. In *ECCV*, pages 469–485, 2024. 2

[39] Yicong Li, Xiang Wang, Junbin Xiao, and Tat-Seng Chua. Equivariant and invariant grounding for video question answering. In *ACM MM*, pages 4714–4722, 2022. 5

[40] Yicong Li, Xiang Wang, Junbin Xiao, Wei Ji, and Tat-Seng Chua. Invariant grounding for video question answering. In *CVPR*, pages 2928–2937, 2022. 5

[41] Bencheng Liao, Shaoyu Chen, Haoran Yin, Bo Jiang, Cheng Wang, Sixu Yan, Xinbang Zhang, Xiangyu Li, Ying Zhang, Qian Zhang, and Xinggang Wang. Diffusiondrive: Truncated diffusion model for end-to-end autonomous driving. In *CVPR*, 2025. 2

[42] Pengyang Ling, Jiazi Bu, Pan Zhang, Xiaoyi Dong, Yuhang Zang, Tong Wu, Huaian Chen, Jiaqi Wang, and Yi Jin. Motionclone: Training-free motion cloning for controllable video generation. In *ICLR*, 2025. 6, 7, 3

[43] Hongcheng Liu, Pingjie Wang, Zhiyuan Zhu, Yanfeng Wang, and Yu Wang. Ce-vdg: Counterfactual entropy-based bias reduction for video-grounded dialogue generation. In *LREC-COLING*, pages 2958–2968, 2024. 2

[44] Ruyang Liu, Jingjia Huang, Thomas H. Li, and Ge Li. Causality compensated attention for contextual biased visual recognition. In *ICLR*, 2023. 7

[45] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. Grounding DINO: marrying DINO with grounded pre-training for open-set object detection. In *ECCV*, pages 38–55, 2024. 7

[46] Yang Liu, Guanbin Li, and Liang Lin. Cross-modal causal relational reasoning for event-level visual question answering. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(10):11624–11641, 2023. 5

[47] Zhengxiong Luo, Dayou Chen, Yingya Zhang, Yan Huang, Liang Wang, Yujun Shen, Deli Zhao, Jingren Zhou, and Tieniu Tan. Videofusion: Decomposed diffusion models for high-quality video generation. In *CVPR*, pages 10209–10218, 2023. 2

[48] Xin Ma, Yaohui Wang, Gengyun Jia, Xinyuan Chen, Ziwei Liu, Yuan-Fang Li, Cunjian Chen, and Yu Qiao. Latte: Latent diffusion transformer for video generation. *CoRR*, abs/2401.03048, 2024. 1, 2, 6, 7, 8, 3, 4

[49] Hieu Man, Franck Dernoncourt, and Thien Huu Nguyen. Mastering context-to-label representation transformation for event causality identification with diffusion models. In *AAAI*, pages 18760–18768, 2024. 2

[50] Yifan Mao, Jian Liu, and Xianming Liu. Stealing stable diffusion prior for robust monocular depth estimation. *arXiv preprint arXiv:2403.05056*, 2024. 2

[51] Nature. Safe driving cars. *Nat. Mach. Intell.*, 4(2):95–96, 2022. 1

[52] Dat Viet Thanh Nguyen, Anh Tran, Nam Vu, Cuong Pham, and Minh Hoai. Driver attention tracking and analysis. *arXiv preprint arXiv:2404.07122*, 2024. 2

[53] Andrea Palazzi, Davide Abati, Simone Calderara, Francesco Solera, and Rita Cucchiara. Predicting the driver's focus of attention: The dr(eye)ve project. *IEEE Trans. Pattern Anal. Mach. Intell.*, 41(7):1720–1733, 2019. 2

[54] Ethan Pronovost, Kai Wang, and Nick Roy. Generating driving scenes with diffusion. *arXiv preprint arXiv:2305.18452*, 2023. 2

[55] Chenyang Qi, Xiaodong Cun, Yong Zhang, Chenyang Lei, Xintao Wang, Ying Shan, and Qifeng Chen. Fatezero: Fusing attentions for zero-shot text-based video editing. In *ICCV*, pages 15886–15896, 2023. 6

[56] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763, 2021. 4, 5, 1

[57] Megani Rajendran, Chek Tien Tan, Indriyati Atmosukarto, Aik Beng Ng, and Simon See. Review on synergizing the metaverse and ai-driven synthetic data: enhancing virtual realms and activity recognition in computer vision. *Vis. Intell.*, 2(1):27, 2024. 2

[58] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10674–10685, 2022. 3

[59] Pedro Sanchez and Sotirios A. Tsaftaris. Diffusion causal models for counterfactual estimation. In *CLeaR*, pages 647–668, 2022. 2

[60] Pedro Sanchez, Xiao Liu, Alison Q. O'Neil, and Sotirios A. Tsaftaris. Diffusion models for causal discovery via topological ordering. In *ICLR*, 2023. 2

[61] Yuan Shen, Niviru Wijayaratne, Pranav Sriram, Aamir Hasan, Peter Du, and Katherine Driggs Campbell. Cocatt: A cognitive-conditioned driver attention dataset. In *ITSC*, pages 32–39, 2022. 2

[62] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *ICLR*, 2021. 5

[63] Jinming Su, Songen Gu, Yiting Duan, Xingyue Chen, and Junfeng Luo. Text2street: Controllable text-to-image generation for street views. *arXiv preprint arXiv:2402.04504*, 2024. 2

[64] Yoad Tewel, Rinon Gal, Dvir Samuel, Yuval Atzmon, Lior Wolf, and Gal Chechik. Add-it: Training-free object insertion in images with pretrained diffusion models. In *ICLR*, 2025. 7

[65] Cherise Threewitt. U. s. news, 10 vehicles that are almost self-driving in 2024. https://cars.usnews.com/cars-trucks/advice/cars-that-are-almost-self-driving, 2024. 1

[66] Han Tian, Tao Deng, and Hongmei Yan. Driving as well as on a sunny day? predicting driver's fixation in rainy weather conditions via a dual-branch visual model. *IEEE CAA J. Autom. Sinica*, 9(7):1335–1338, 2022. 2

[67] Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphaël Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. *CoRR*, abs/1812.01717, 2018. 6

[68] Jianyi Wang, Zhijie Lin, Meng Wei, Yang Zhao, Ceyuan Yang, Fei Xiao, Chen Change Loy, and Lu Jiang. Seedvr: Seeding infinity in diffusion transformer towards generic video restoration. In *CVPR*, 2025. 2, 6

[69] Xiang Wang, Yingxin Wu, An Zhang, Fuli Feng, Xiangnan He, and Tat-Seng Chua. Reinforced causal explainer for graph neural networks. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45 (2):2297–2309, 2023. 2

[70] Xiaofeng Wang, Zheng Zhu, Guan Huang, Xinze Chen, and Jiwen Lu. Drivedreamer: Towards real-world-driven world models for autonomous driving. In *ECCV*, 2024. 2

[71] Yaohui Wang, Xinyuan Chen, Xin Ma, Shangchen Zhou, Ziqi Huang, Yi Wang, Ceyuan Yang, Yinan He, Jiashuo Yu, Peiqing Yang, et al. Lavie: High-quality video generation with cascaded latent diffusion models. *Int. J. Comput. Vis.*, 2024. 2

[72] Zihan Wang, Ziliang Xiong, Hongying Tang, and Xiaobing Yuan. Detail-enhancing framework for reference-based image super-resolution. *arXiv preprint arXiv:2405.00431*, 2024. 2

[73] Yuqing Wen, Yucheng Zhao, Yingfei Liu, Fan Jia, Yanhui Wang, Chong Luo, Chi Zhang, Tiancai Wang, Xiaoyan Sun, and Xiangyu Zhang. Panacea: Panoramic and controllable video generation for autonomous driving. In *CVPR*, pages 6902–6912, 2024. 2

[74] Jay Zhangjie Wu et al. Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. In *ICCV*, pages 7589–7599, 2023. 6, 7, 8, 1

[75] Ruiqi Wu, Liangyu Chen, Tong Yang, Chunle Guo, Chongyi Li, and Xiangyu Zhang. Lamp: Learn a motion pattern for few-shot video generation. In *CVPR*, pages 7089–7098, 2024. 6, 7, 3, 4, 5

[76] Ye Xia, Danqing Zhang, Jinkyu Kim, Ken Nakayama, Karl Zipser, and David Whitney. Predicting driver attention in critical situations. In *ACCV*, pages 658–674, 2018. 2, 6

[77] Jinxi Xiang, Ricong Huang, Jun Zhang, Guanbin Li, Xiao Han, and Yang Wei. Versvideo: Leveraging enhanced temporal diffusion models for versatile video generation. In *ICLR*, 2023. 2

[78] Zeqi Xiao, Yifan Zhou, Shuai Yang, and Xingang Pan. Video diffusion models are training-free motion interpreter and controller. *CoRR*, abs/2405.14864, 2024. 2, 6

[79] Yunzhi Yan, Zhen Xu, Haotong Lin, Haian Jin, Haoyu Guo, Yida Wang, Kun Zhan, Xianpeng Lang, Hujun Bao, Xiaowei Zhou, and Sida Peng. Streetcrafter: Street view synthesis with controllable video diffusion models. In *CVPR*, 2025. 2

[80] Xu Yang, Hanwang Zhang, and Jianfei Cai. Deconfounded image captioning: A causal retrospect. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(11):12996–13010, 2023. 2

[81] Xiangpeng Yang, Linchao Zhu, Hehe Fan, and Yi Yang. Videograin: Modulating space-time attention for multi-grained video editing. 2025. 1

[82] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. In *ICLR*, 2025. 1, 2, 6, 7, 8, 3, 4

[83] Xi Ye and Guillaume-Alexandre Bilodeau. Stdiff: Spatio-temporal diffusion for continuous stochastic video prediction. In *AAAI*, pages 6666–6674, 2024. 1

[84] Xin Yuan, Jinoo Baek, Keyang Xu, Omer Tov, and Hongliang Fei. Inflation with diffusion: Efficient temporal adaptation for text-to-video super-resolution. In *WACV*, pages 489–496, 2024. 2

[85] An Zhang, Fangfu Liu, Wenchang Ma, Zhibo Cai, Xiang Wang, and Tat-Seng Chua. Boosting causal discovery via adaptive sample reweighting. In *ICLR*, 2023. 2

[86] Wenjun Zhang. How to unify grounding and causation. *Synthese*, 202(24):1–18, 2023. 2

[87] Yabo Zhang, Yuxiang Wei, Dongsheng Jiang, Xiaopeng Zhang, Wangmeng Zuo, and Qi Tian. Controlvideo: Training-free controllable text-to-video generation. In *ICLR*, 2024. 6

[88] Zicheng Zhang, Bonan Li, Xuecheng Nie, Congying Han, Tiande Guo, and Luoqi Liu. Towards consistent video editing with text-to-image diffusion models. In *NeurIPS*, 2023. 2

[89] Yupeng Zhou, Daquan Zhou, Ming-Ming Cheng, Jiashi Feng, and Qibin Hou. Storydiffusion: Consistent self-attention for long-range image and video generation. In *ICLR*, 2024. 2