Supplementary Material of Causal-Entity Reflected Egocentric Traffic Accident Video Synthesis

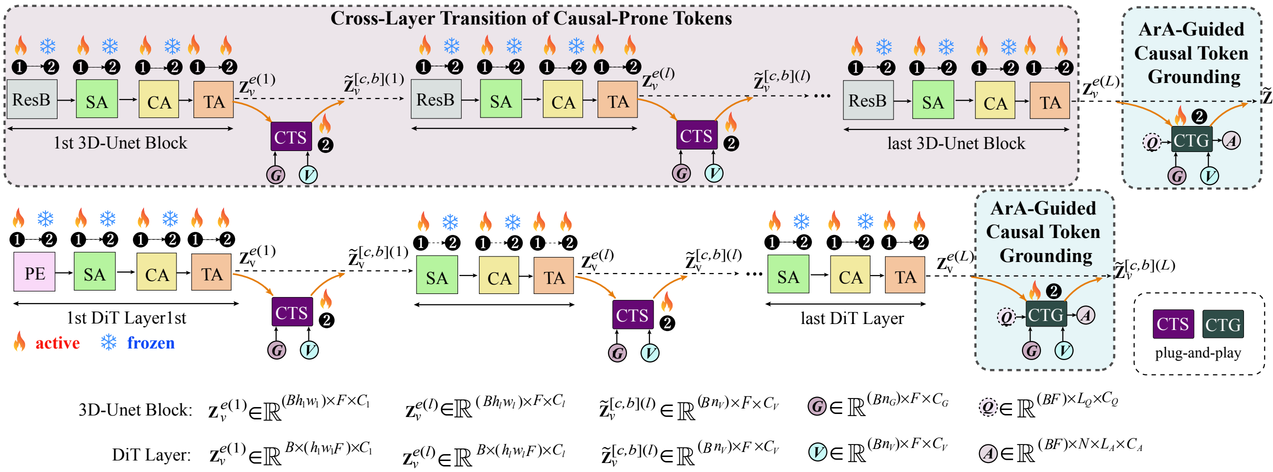
Cross-Layer Transition of Causal-Prone Tokens
ResB
SA
CA
TA
CTS
1st 3D-Unet Block
last 3D-Unet Block
ArA-Guided Causal Token Grounding
CTG
PE
1st DiT Layer1st
last DiT Layer
active
frozen
plug-and-play
3D-Unet Block: $\mathbf{z}_v^{e(1)} \in \mathbb{R}^{(Bh_lw_l)\times F\times C_l}$ $\mathbf{z}_v^{e(l)} \in \mathbb{R}^{(Bh_lw_l)\times F\times C_l}$ $\tilde{\mathbf{z}}_v^{[c,b](l)} \in \mathbb{R}^{(Bn_V)\times F\times C_V}$ $G \in \mathbb{R}^{(Bn_G)\times F\times C_G}$ $Q \in \mathbb{R}^{(BF)\times L_Q\times C_Q}$
DiT Layer: $\mathbf{z}_v^{e(1)} \in \mathbb{R}^{B\times(h_lw_lF)\times C_l}$ $\mathbf{z}_v^{e(l)} \in \mathbb{R}^{B\times(h_lw_lF)\times C_l}$ $\tilde{\mathbf{z}}_v^{[c,b](l)} \in \mathbb{R}^{(Bn_V)\times F\times C_V}$ $V \in \mathbb{R}^{(Bn_V)\times F\times C_V}$ $A \in \mathbb{R}^{(BF)\times N\times L_A\times C_A}$

Figure 1. **The detailed injection workflow of CTS and CTG** in Causal-VidSyn, interleaved with multiple SA, CA, and TA blocks in 3D-Unet or DiT-based video diffusion models. The dimension indexes are defined as: B denotes the batch size, $F = 16$ is the video frame length, the maximum question prompt length L_Q is set to 10, and the maximum answer prompt length L_A is set to 32. In each layer of 3D-Unet block or DiT, C_l, h_l, and w_l represent the channels, height, and width of the input noisy vision tokens $\boldsymbol{z}_v^{e(l)}$. C_V, C_G, C_Q, and C_A are the token channels of video V, gaze maps (G), the question (Q) prompt and optional answers (A), respectively. n_V and n_G represent the number of vision tokens and gaze map tokens. Here, we set $C_V = C_G = C_Q = C_A = 1024$ and $n_V = n_G = 256$ in this work.

1. More Details of Implementation

To be clear for re-reproduction, we provide the injection workflow of **CTS** and **CTG** in Fig. 1, where different attention modules, *i.e.*, **SA**, **CA**, and **TA**, are fed into low-rank adaptation (LoRA) trainer [2] for fast fine-tuning. In Stage-❷, we mainly freeze other modules and only fine-tune TA, **CTS** and **CTG**, for Tune-A-Video (TAV) [74], Latte-T+**CTS**+**CTG**+G[3] [48], and our Causal-VidSyn. For CogV-X-T+**CTS**+**CTG**+G, we follow the official training strategy of CogV-X [82], and update the parameters of query ($\mathbf{q}$), key ($\mathbf{k}$), value ($\mathbf{v}$) of CA and TA blocks.

2. More Details of CTS and CTG

2.1. The Architecture of Sampling Adapter

As denoted in Fig. 1, $\boldsymbol{z}_v^{e(l)} \in \mathbb{R}^{(Bh_lw_l)\times F\times C_l}$. To match the dimension of $\boldsymbol{z}_v^{gate}$ in Eq. 4 stated in the main paper, a bilinear interpolation (BintP) is applied to adjust the token dimension as $\mathbb{R}^{(Bh_pw_p)\times F\times C_l}$, where the size of $h_p \times w_p$ equals the number of vision tokens n_V. Then, we apply a 2D convolution (kernel size: 1×1) to transform C_l to C_V. We reshape the token shape and output $\tilde{\boldsymbol{z}}_v^{e(l)} \in \mathbb{R}^{(Bn_V)\times F\times C_V}$, which is then fed into the **CTS** block.

From Fig. 1, it is worth noting that there is a change between the input and output of **CTS** block in the spatial dimension of tokens though the *token sampling adapter* (Eq. 6). This aims to enhance the versatility of the **CTS** block, *i.e.*, ensuring the noise representation $\boldsymbol{z}_v^e$ of different diffusion models (*e.g.*, Unet- or DiT-based) to adapt to the video representation $\boldsymbol{z}_v$ (Eq. 5) outputted by CLIP model [56] in **CTS** block after token sampling adapter (Eq. 6). Actually, this is universal in the cross-attention (CA) module of 3D-Unet backbones to fulfill the text-vision alignment with the extra input of vision or text tokens. We also have attempted to keep the resolution of $\boldsymbol{z}_v$ (Eq. 6) consistent with $\boldsymbol{z}_v^e$ (Eq. 3), while multi-layer **CTS** block will increase memory usage and computation cost exhaustively without versatility.

Therefore, we try our best to minimize the influence of **CTS** on the spatial relationships of the tokens in the cross-layer transition within the 3D-Unet backbone as much as possible. As shown in Fig. 1, the 1st **CTS** block is injected at the end of the 1st layer of 3D-Unet (DiT: the same way) backbone. In the inference phase, **CTS** and **CTG** are removed, where the output of the 1st layer of 3D-Unet (or DiT) backbone is directly fed into the next layer of backbones. Additionally, only the temporal attention (TA) is fine-tuned in Stage-❷, which has not influence on the position indices of spatial attention (SA) and crossing attention (CA) modules. Therefore, in the inference phase, the spatial

[2] https://github.com/cloneofsimo/lora

[3] The official Latte does not provide T2V training code, we take the implementation of https://github.com/lyogavin/train_your_own_sora for Latte's T2V training.

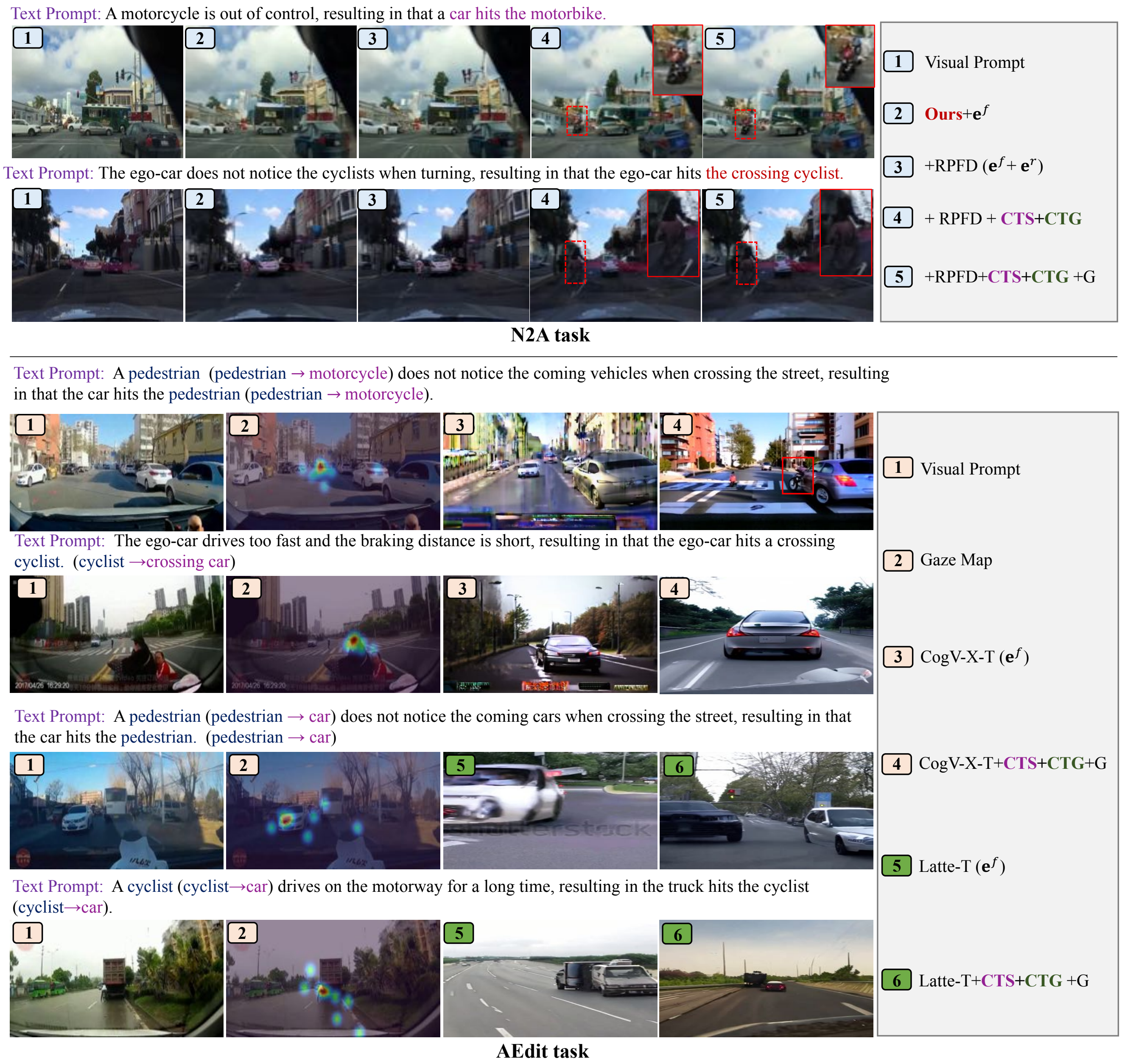
Text Prompt: A motorcycle is out of control, resulting in that a car hits the motorbike.
1
2
3
4
5
Text Prompt: The ego-car does not notice the cyclists when turning, resulting in that the ego-car hits the crossing cyclist.
1 Visual Prompt
2 Ours+$\mathbf{e}^f$
3 +RPFD ($\mathbf{e}^f$+ $\mathbf{e}^r$)
4 + RPFD + CTS+CTG
5 +RPFD+CTS+CTG +G
N2A task
Text Prompt: A pedestrian (pedestrian → motorcycle) does not notice the coming vehicles when crossing the street, resulting in that the car hits the pedestrian (pedestrian → motorcycle).
Text Prompt: The ego-car drives too fast and the braking distance is short, resulting in that the ego-car hits a crossing cyclist. (cyclist →crossing car)
Text Prompt: A pedestrian (pedestrian → car) does not notice the coming cars when crossing the street, resulting in that the car hits the pedestrian. (pedestrian → car)
Text Prompt: A cyclist (cyclist→car) drives on the motorway for a long time, resulting in the truck hits the cyclist (cyclist→car).
1 Visual Prompt
2 Gaze Map
3 CogV-X-T ($\mathbf{e}^f$)
4 CogV-X-T+CTS+CTG+G
5 Latte-T ($\mathbf{e}^f$)
6 Latte-T+CTS+CTG +G
AEdit task

Figure 2. We visualize two N2A examples and four AEdit samples with the ablation response checking of our Causal-VidSyn , CogV-X-T [82], and Latte-T [48] with different causal-aware fine-tuning stages.

relationships of tokens are not disrupted.

2.2. Extending CTS and CTG to DiT-type VDMs

As for the injection of **CTS** and **CTG** in DiT-based video diffusion models (VDMs), as shown in Fig. 1, we denote the noisy vision tokens in the inner DiT layers as $\boldsymbol{z}_v^{e(l)} \in \mathbb{R}^{(B\times(h_l w_l F)\times C_l}$. Because there is no downscale or upscale operation in DiT layers, different from 3D-Unet, the sampling adapter just needs to reshape $\boldsymbol{z}_v^{e(l)}$ to match the dimension of $\boldsymbol{z}_v^{gate}$ without the BintP and Conv2D operations in 3D-Unet. Then, the gated fusion, **CTS** and **CTG** take the same structures.

To offer more evidence for the **CTS** and **CTG** in this work, we visualize more samples in Fig. 2 for checking their roles, where only the last frame in each generated clip is presented concisely. It can be observed that **CTS** and **CTG** modules can help CogV-X-T and Latte-T to recover the frame content (*e.g.*, tree, building, and sky regions) in original video frames and present active response to the changed text phrase (cyclist/pedestrian→ car). As for our Causal-

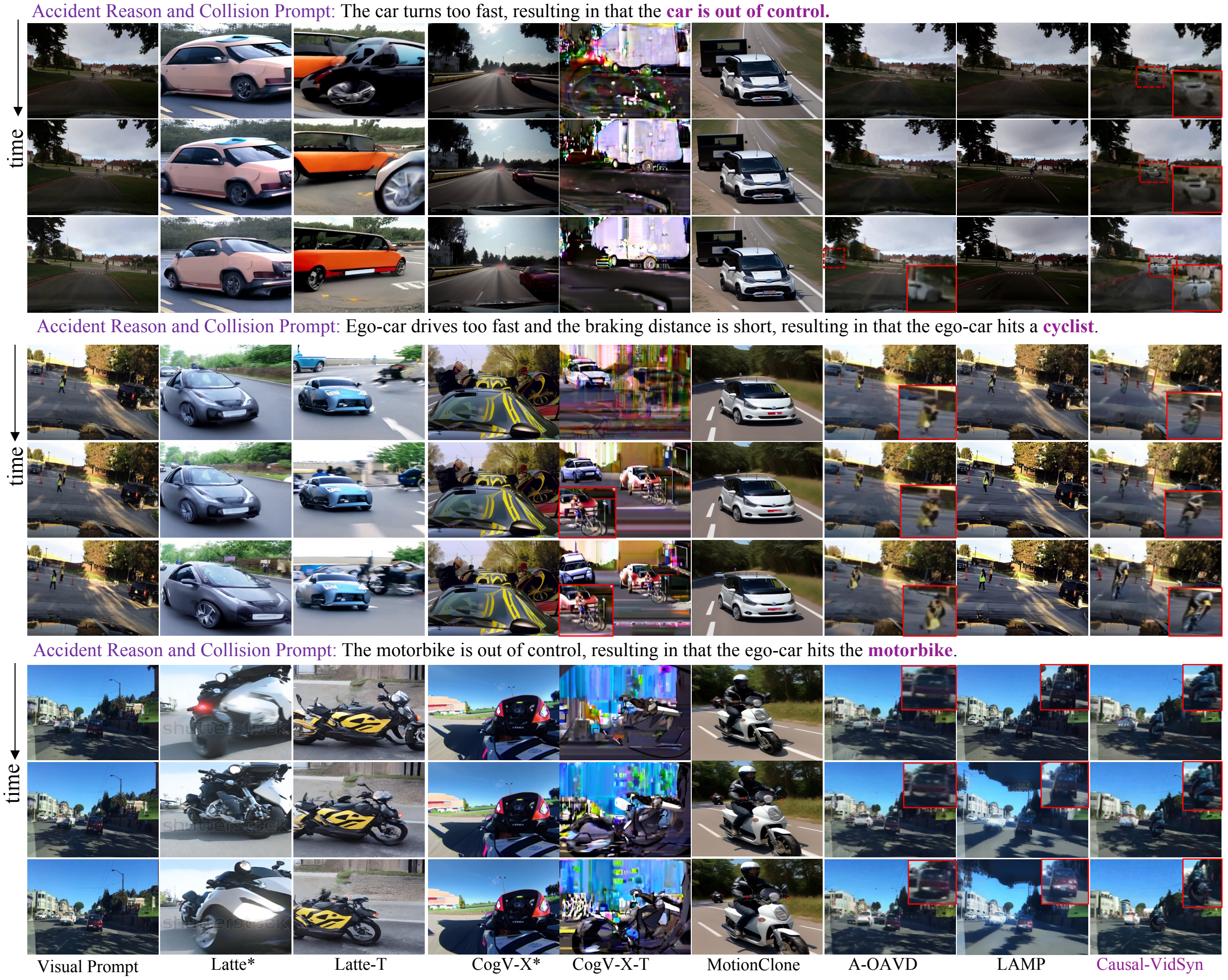
Accident Reason and Collision Prompt: The car turns too fast, resulting in that the car is out of control.
time
Accident Reason and Collision Prompt: Ego-car drives too fast and the braking distance is short, resulting in that the ego-car hits a cyclist.
time
Accident Reason and Collision Prompt: The motorbike is out of control, resulting in that the ego-car hits the motorbike.
time
Visual Prompt
Latte*
Latte-T
CogV-X*
CogV-X-T
MotionClone
A-OAVD
LAMP
Causal-VidSyn

Figure 3. 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).

VidSyn, **CTS** and **CTG** perform well for active response corresponding to the given text prompt. In summary, the **CTS** and **CTG** modules can fine-tune the video diffusion models to identify the critical object effectively and rule out the influence of background scenes.

3. More Evaluations of Causal-VidSyn

For a solid evaluation, we offer more evaluations mainly from the visualizations of N2A, T2V, and AEdit tasks.

3.1. More Visualizations on N2A and T2V Tasks

N2A Evaluation: We present more ego-car involved visualizations of the N2A task in Fig. 3. It can be observed that the "large object issue" is manifest in MotionClone [42], and DiT-based methods, *i.e.*, Latte* [48] and CogV-X* [82], despite of the clear video frames. Additionally, the frame style changes dramatically for Latte* and CogVideoX*. After training by the egocentric accident videos (*i.e.*, CogV-X*→CogV-X-T), the content structure (*e.g.*, the trees and sky) in the generated frames becomes closer to the original visual prompt. MotionClone is not consistently active for the given text prompt, such as the "cyclist" of the 1^{st} sample in Fig. 3, and generates an irrelevant car. A-OAVD [17] and LAMP [75] show failures on the presented samples where the expected content change in most examples does not appear, such as the cyclist and motorbike in the 2^{nd} and 3^{rd} samples in Fig. 3. Our Causal-VidSyn can change the critical objects with the best response, *w.r.t.*, the given text prompt, while maintaining the frame background well.

T2V Evaluation: We present more T2V visualizations on ego-car involved accidents in Fig. 4, where we compare

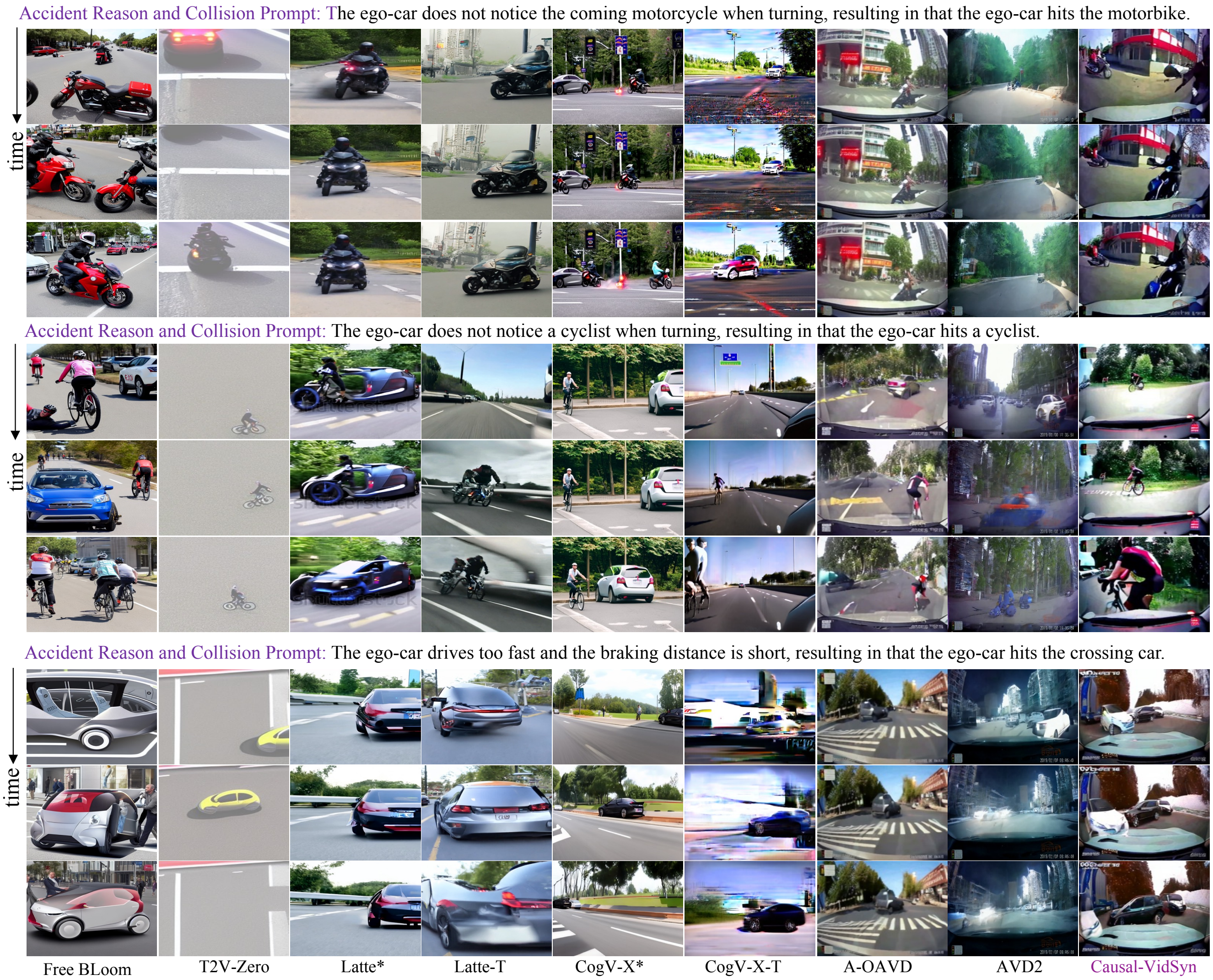
Accident Reason and Collision Prompt: The ego-car does not notice the coming motorcycle when turning, resulting in that the ego-car hits the motorbike.
time
Accident Reason and Collision Prompt: The ego-car does not notice a cyclist when turning, resulting in that the ego-car hits a cyclist.
time
Accident Reason and Collision Prompt: The ego-car drives too fast and the braking distance is short, resulting in that the ego-car hits the crossing car.
time
Free BLoom
T2V-Zero
Latte*
Latte-T
CogV-X*
CogV-X-T
A-OAVD
AVD2
Causal-VidSyn

Figure 4. Sample visualizations of T2V task by Free-bloom [28], T2V-Zero [32], Latte* [48], Latte-T [48], CogV-X* [82], CogV-X-T [82], A-OAVD [17], AVD2 [34], and our Causal-VidSyn (Best viewed in zoom mode).

Free-bloom [28], T2V-Zero [32], Latte* [48], Latte-T [48], CogV-X* [82], CogV-X-T [82], A-OAVD [17], AVD2 [34], and our Causal-VidSyn. From these visualization results, we can see that Free-bloom [28] likely generates multiple and similar objects while the collision situations do not occur. The results of T2V-Zero [32] show more surveillance views. AVD2 [34], fine-tuned Sora[4] by MM-AU dataset [17] can generate a similar frame style, while the expected collision is not well exhibited. In the T2V task, we can observe that the CogV-X* and Latte* and the other training-free methods cannot reflect the collisions well. As for A-OAVD, the expected near-crash scenarios appear while the collisions are only generated by our Causal-VidSyn. From these results, it concludes that egocentric traffic accident knowledge is limited in these text-to-video generation methods.

3.2. More Visualizations on the AEdit Task

The AEdit task is directly to show the ability for causal-entity reflected accident video synthesis by counterfactual text prompt change. In Fig. 5, we present four samples with the comparison of LAMP [75], A-OAVD [17], and our Causal-VidSyn. LAMP [75] prefers collision-free and commonly focuses on the consistency of background scenes. A-OAVD [17] presents active responses while the target shape is not complete and the locations of the created targets have a larger distance to the ego-car compared with our Causal-VidSyn. When facing strong or dark light conditions, A-OAVD cannot edit the expected video content well, *w.r.t.*, the counterfactual text phrase. As for our Causal-VidSyn, it

[4] https://openai.com/sora/.

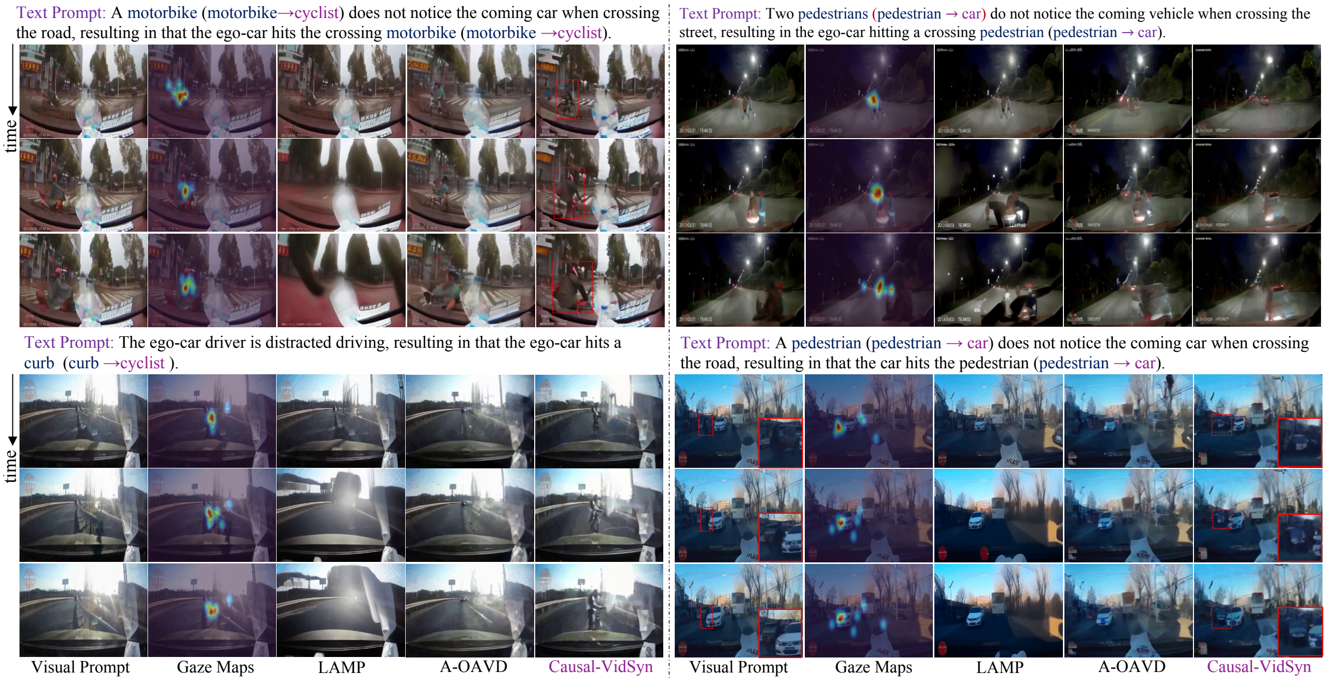
Text Prompt: A motorbike (motorbike→cyclist) does not notice the coming car when crossing the road, resulting in that the ego-car hits the crossing motorbike (motorbike →cyclist).
Text Prompt: Two pedestrians (pedestrian → car) do not notice the coming vehicle when crossing the street, resulting in the ego-car hitting a crossing pedestrian (pedestrian → car).
time
Text Prompt: The ego-car driver is distracted driving, resulting in that the ego-car hits a curb (curb →cyclist).
Text Prompt: A pedestrian (pedestrian → car) does not notice the coming car when crossing the road, resulting in that the car hits the pedestrian (pedestrian → car).
time
Visual Prompt
Gaze Maps
LAMP
A-OAVD
Causal-VidSyn
Visual Prompt
Gaze Maps
LAMP
A-OAVD
Causal-VidSyn

Figure 5. We visualize more AEdit results of four different situations (rainy, dark, strong light, and shadow) by LAMP [75], A-OAVD [17], and our Causal-VidSyn.

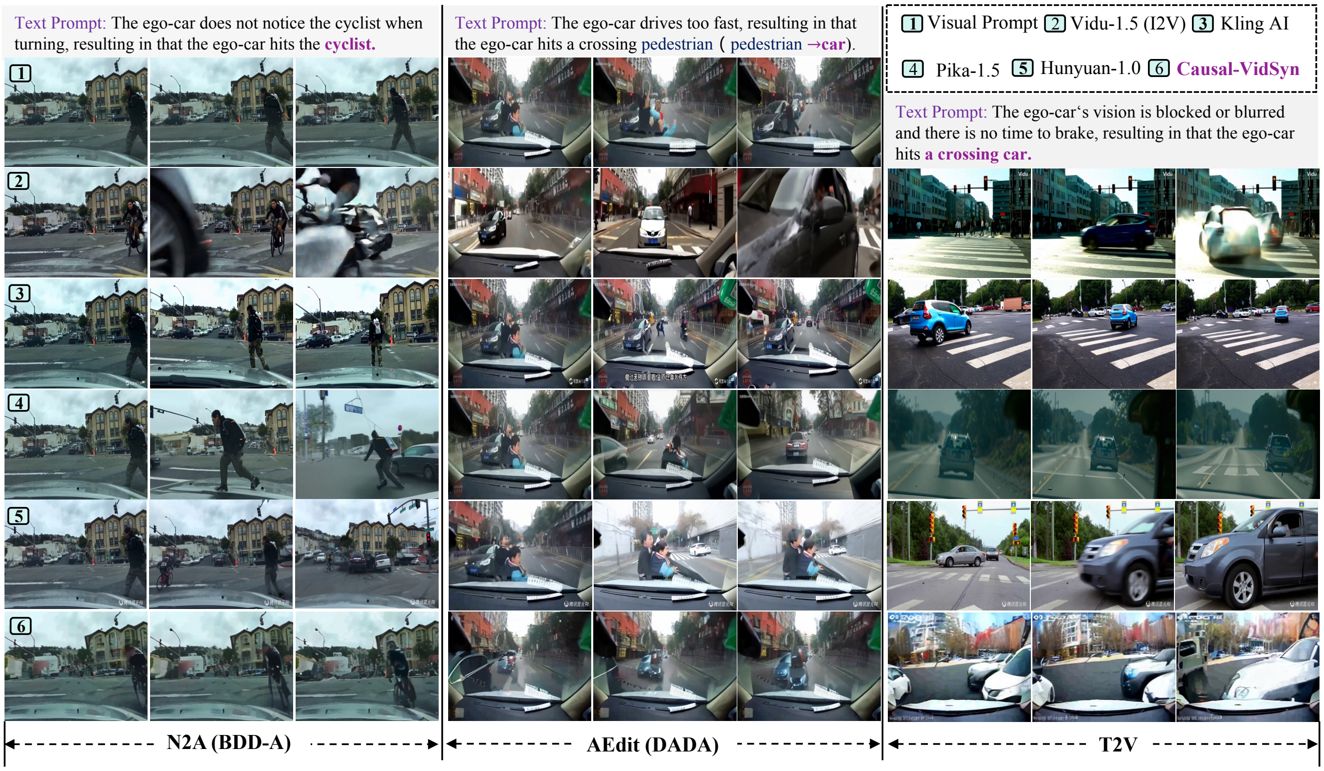
Text Prompt: The ego-car does not notice the cyclist when turning, resulting in that the ego-car hits the cyclist.
Text Prompt: The ego-car drives too fast, resulting in that the ego-car hits a crossing pedestrian (pedestrian →car).
1 Visual Prompt 2 Vidu-1.5 (I2V) 3 Kling AI
4 Pika-1.5 5 Hunyuan-1.0 6 Causal-VidSyn
Text Prompt: The ego-car's vision is blocked or blurred and there is no time to brake, resulting in that the ego-car hits a crossing car.
1
2
3
4
5
6
N2A (BDD-A)
AEdit (DADA)
T2V

Figure 6. Comparisons between Causal-VidSyn and four commercial model: Pika-1.5, Vidu-1.5 (I2V), Kling-AI, and HunyuanVideo-I2V.

shows a better causal-sensitive object editing, *w.r.t.*, counterfactual text modification, than other models.

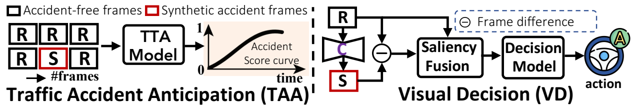

Figure 7. The TAA and decision Tasks. **R** is the real data, and **S** is the synthesis data driven by **C** (our Causal-VidSyn model).

4. Comparison with Commercial Models

To verify the SOTA performance of our Causal-VidSyn, we also take four popular and famous commercial models, including Pika-1.5[5], Vidu-1.5[6], Kling AI[7], and newly released HunyuanVideo-I2V[8], for egocentric traffic accident video generation. Fig. 6 displays one sample for N2A, AEdit, and T2V tasks, respectively. From the results, it is interesting that only Vidu-1.5 conditioned by images (*abbrev.,* Vidu-1.5 (I2V)) can generate egocentric accidents well for the examples in all tasks. The text prompt-driven vision change is not well exhibited for Pika-1.5, Kling-AI, and HunyuanVideo-I2V for the N2A and AEdit tasks. For example, the pedestrian in the first example in Fig. 6 is not changed to an expected cyclist. Vidu-1.5 (I2V) can reflect the behavior or visual content change, while it does not display an active response of "pedestrian"→"car". HunyuanVideo-I2V generates a good generation for the T2V task, while in N2A and AEdit tasks, it does not show the expected video content change. For Kling-AI, it contrarily shows an accident dissipation process, where the near-crash objects go far away from the ego-car given the text prompt. As for our Causal-VidSyn, the expected egocentric accident situation is outputted. Certainly, these comparisons cannot represent all situations, but they can exhibit the causal sensitivity of our model for accident video content editing and egocentric accident video generation.

5. Downstream Task Explorations

We explore two downstream tasks using our synthetic data: traffic accident anticipation (TAA) and visual decision (VD). Fig. 7 shows the pipelines of them. In the TAA task, we take Cog-TAA [36] as the baseline, and take the accident-free BDD-A [76] datasets to synthesize accident frames by LAMP, OAVD, and our model (**C**), respectively, obtaining the same-scale accident and accident-free sample pairs to Cog-TAA for training.

The same DADA-2000 test set [16] is adopted. For the VD task, we implement a visual decision work (VD-OIA)[9] by adopting the same training and testing frames on action decision from VD-OIA, while differently, we load our model (**C**) on the training data (**R**) to obtain a difference map (**R**-**C**(**S**)) for better important object learning (named as VD-OIA*). For the TAA task, we use the same AP, AUC, and $\text{TTA}_{0.5}$ of [36] for an accident frame and temporal occurrence determination, and the Action mF1, and Action F1-all metrics in VD-OIA for VD task. From Tab. 1 and Fig. 8, our model achieves better performance than the baselines. Notably, Cog-TAA (Ours) trained on accident-free videos achieves better AP and $\text{TTA}_{0.5}$ values than Cog-TAA trained on manually-labeled accident videos.

Baseline	DADA-2000 [16]		
	AP↑	AUC↑	$\text{TTA}_{0.5}$↑
Cog-TAA [36] w/o *T&Att*	0.701	**0.774**	3.742
Cog-TAA (LAMP)	0.698	0.504	3.210
Cog-TAA (OAVD)	0.735	0.654	3.746
Cog-TAA (Ours)	**0.758**	0.689	**3.762**

Table 1. The TAA task results.

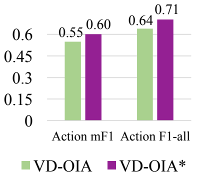

Figure 8. The VD task results.

6. Failure Case Analysis

In addition, we also show the limitations of our Causal-VidSyn by analyzing some failure cases, as shown in Fig. 9. In this analysis, we take several samples in N2A and AEdit tasks because of the demand for causal-sensitivity checking. For the failure cases in the N2A task, we can observe that the expected video content change does not appear because of the severe illumination (strong or dark light conditions) and rare object-scene interaction (*i.e., a cyclist drives on the motorway for a long time*). These failure cases inspire two possible insights: 1) Light or weather conditions can be further involved in the text prompt and model designs in future research; 2) The appropriate selection of text prompt in video diffusion is important, especially in the mixed traffic scenes. The text and visual frame prompts need to be well-paired for realistic video diffusion. Maybe, the dense video captioning approach (*e.g.*, Pllava[10]) can be introduced with a collision phrase ("hitting") injection.

Additionally, for the AEdit task, besides the severe illumination issue, the large-scale objects (occupying large regions) are hard to be edited (*e.g.*, the cars with "overtaking" behavior). For these situations, some object detectors can be further adopted, and the causal-entity editing in the egocentric accident video synthesis needs to consider some object-level insights. However, involving more conditions may restrict the flexibility of the causal token selection and grounding process. Therefore, this remains an open problem and needs to be considered in the future.

From the extensive visualization and ablation analysis, the superiority of our Causal-VidSyn is evidently verified.

[5] Release time: Oct 2, 2024, pika.art/

[6] Release time: Nov 13, 2024, vidu.studio/zh

[7] Release time: Jul 25, 2024, klingai.kuaishou.com/

[8] Release time: Mar 06, 2025, https://github.com/Tencent/HunyuanVideo

[9] https://twizwei.github.io/bddoia_project/

[10] https://pllava.github.io/

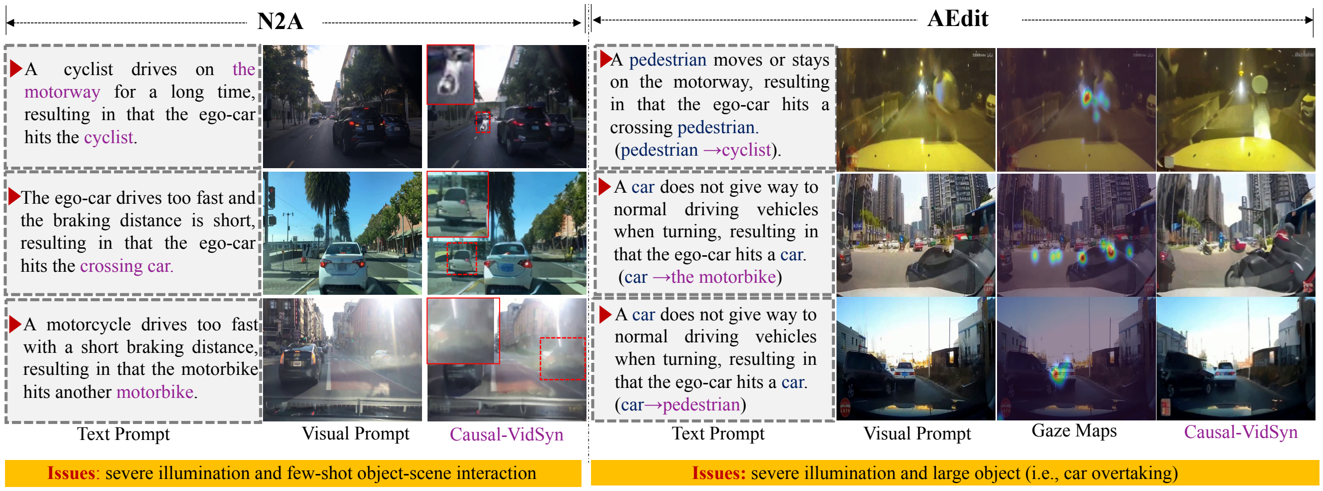
N2A
AEdit
A cyclist drives on the motorway for a long time, resulting in that the ego-car hits the cyclist.
The ego-car drives too fast and the braking distance is short, resulting in that the ego-car hits the crossing car.
A motorcycle drives too fast with a short braking distance, resulting in that the motorbike hits another motorbike.
A pedestrian moves or stays on the motorway, resulting in that the ego-car hits a crossing pedestrian. (pedestrian →cyclist).
A car does not give way to normal driving vehicles when turning, resulting in that the ego-car hits a car. (car →the motorbike)
A car does not give way to normal driving vehicles when turning, resulting in that the ego-car hits a car. (car→pedestrian)
Text Prompt
Visual Prompt
Causal-VidSyn
Text Prompt
Visual Prompt
Gaze Maps
Causal-VidSyn
Issues: severe illumination and few-shot object-scene interaction
Issues: severe illumination and large object (i.e., car overtaking)

Figure 9. Some failure cases in N2A and AEdit tasks.

7. Ethics Statement

The misuse including the creation of deceptive accident content for evidence collection may have negative societal impacts, and we advocate positive use for deep accident understanding, such as accident anticipation. In addition, we claim that all authors have solid contributions to this work.