Texvent: Asynchronous Event Data Simulation via Text Prompt

Ruofei Wang[1,3] Peiqi Duan[2] Ka Chun Cheung[3] Simon See[3] Boxin Shi[2] Renjie Wan[1*]
[1]Hong Kong Baptist University [2]Peking University [3]NVIDIA AI Technology Center, NVIDIA
ruofei@life.hkbu.edu.hk {duanqi0001, shiboxin}@pku.edu.cn
{chcheung, ssee}@nvidia.com renjiewan@hkbu.edu.hk

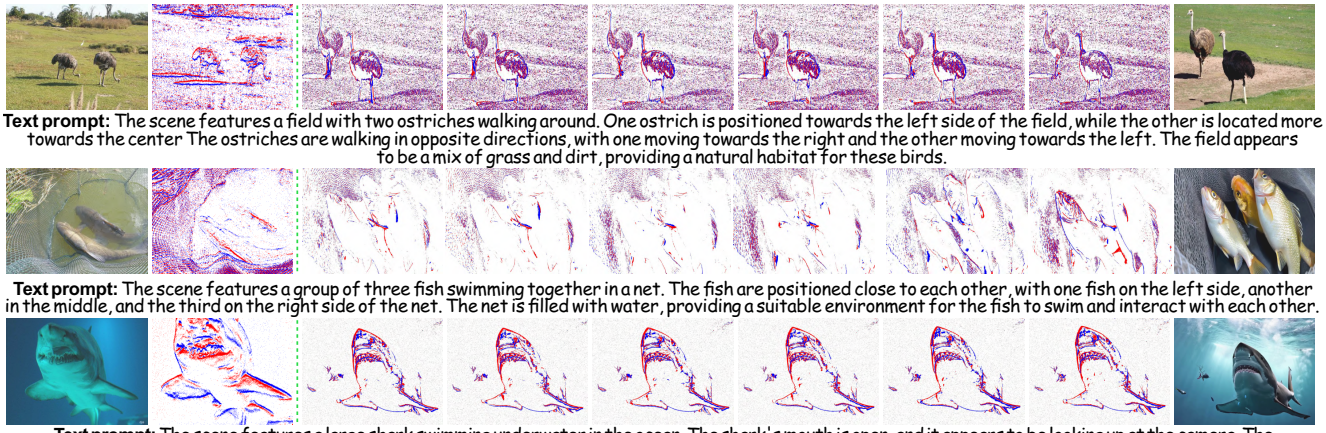

Figure 1. From left to right, the real image-event pair taken from the collected NT-ImageNet dataset, followed by the sequential event streams and a single frame produced by Texvent (https://github.com/rfww/texvent). Texvent can generate asynchronous, sparse, and high-resolution events with smooth transitions and rich backgrounds. Blue and red denote the positive and negative events, respectively.

Abstract

*Current event simulation methods focus on employing videos to synthesize new event data, suffering from costly video capture and limited scalability across viewpoints, motions, and lighting. To this end, we propose a **Text**-to-**event** simulation framework (**Texvent**) that can directly generate asynchronous event data from simple text prompts. Texvent first renders prompt-driven videos via multimodal large language models and subsequently applies a new physical simulator to generate event streams. Specifically, an adaptive brightness-aware frame interpolation approach is proposed to enhance the temporal resolution of the rendered videos. A balanced logarithmic intensity comparison strategy and a cache–based voltage refreshment mechanism are introduced into the simulator to generate event data. To narrow the sim-to-real gap, we also introduce background activity noise injection and dense time stamp reconstruction operations. Extensive experiments demonstrate Texvent's superior computational efficiency and its generation ability.*

1. Introduction

Event camera, a bio-inspired vision sensor, shows great advantages in terms of time latency, power consumption, and dynamic range compared with conventional cameras [9]. These advantages have established event-based learning as a prominent approach across diverse vision tasks [6, 21, 41, 46, 50, 53, 56]. However, the difficulties in large event dataset collection limit the event-based exploration, thereby inspiring an urgent need for event simulation [5].

Event simulation is the technique to generate synthetic event data, eliminating the need for a physical event camera [38]. Such a technique is capable of repurposing a non-event dataset for event-based downstream exploration [11]. Currently, most approaches [11, 19, 38] focus on video-to-event simulation since the continuous video frames can help

[1]Ruofei Wang and Renjie Wan are with the Department of Computer Science, Hong Kong Baptist University. [2]Peiqi Duan and Boxin Shi are with the State Key Laboratory of Multimedia Information Processing and the National Engineering Research Center of Visual Technology, School of Computer Science, Peking University. *Corresponding author.

to formulate the pixel brightness change, a prerequisite for event activation [4]. However, these approaches are constrained by the costly video capture and limited scalability across viewpoints, motions, and lighting. To this end, text-to-event (T2E) is proposed [37], where the event data can be generated under the guidance of simple text prompts. A pioneering approach integrates a text encoder [34], a diffusion model [40], and an autoencoder to achieve the gesture-specific T2E simulation [37]. However, training such specialized models still requires a large corpus of text-event pairs, which limits its feasibility. A training-free T2E method with general simulation ability is highly desirable, as it can reduce expensive data collection and facilitate rapid adaptation to diverse event domains.

A naive solution is to cascade an existing video-to-event simulator [11, 19, 38] after the video generator to synthesize event data. However, such a pipeline is prone to: **1)** low efficiency, due to the redundant bidirectional optical flow estimation during the frame interpolation, making it difficult to support applications that require large amounts of training data or real-time generation; and **2)** low fidelity, stemming from incomplete modeling of the discrepancies between the real event data capture and event data simulation, limiting the generalization ability and reliability of models trained on them. Thus, the ideal T2E method should be designed based on these two concerns.

In this paper, we present a training-free text-to-event simulation method (**Texvent**) that facilitates general event data simulation with only text prompts. The proposed architecture primarily consists of the high-frame-rate video generation and the efficient event data simulation. For the high frame-rate video generation, we adopt a brightness-aware interpolation approach that minimizes the redundant frame interpolation, thereby ensuring **efficiency**. To improve the **fidelity**, we present a new event simulator equipped with a balanced logarithmic intensity comparison strategy and a cache–based voltage refreshment mechanism. These units aim to tackle unfair event activation sensitivity under low- and high-light conditions and reduce event loss caused by frequent reference brightness updates, respectively. With these enhancements, Texvent enables high-fidelity simulation of event data, effectively supporting event-based downstream tasks, such as object recognition, event-to-image reconstruction, *etc*. Our main contributions are as follows:

- We propose a novel T2E framework (Texvent) that overcomes the challenges of event dataset collection, enabling effective support for event-based downstream tasks.
- We present the balanced logarithmic intensity comparison strategy and the cache–based voltage refreshment mechanism to avoid potential event data corruption.
- We introduce NT-ImageNet, a new text-event pair dataset comprising 5000 pairs, specifically curated for evaluating text-to-event simulation.

2. Related work

2.1. Video generation

Recently, Multimodal Large Language Models (MLLMs) have achieved great success in video generation, such as Sora [1], Cosmos [2], Wan [45], VideoPoet [25], video diffusion models [17], *etc*. These models enable realistic video generation with only text prompts, named text-to-video (T2V) generation. VDM [17] is the first T2V model that extends image diffusion models to video generation via employing a U-Net to model temporal details. Make-A-Video [44] decomposes video generation into text-image modeling and motion perception, where a cascade of diffusion models is used to render final high-quality videos. Ge *et al*. [10] propose a video noise prior that is tailored for finetuning a pretrained image diffusion model to increase temporal consistency. [3, 18, 51] insert various temporal layers into the text-to-image model to generate videos. VideodirectorGPT [28] leverages the large language model to generate structured plans, which subsequently guide the T2V model in the production of long video sequences. Apart from T2V models, some explorations focus on employing other modalities such as images [2, 52], videos [49], and control maps [23] to generate videos. Training all these models shows a high computational cost [28, 32]. Therefore, it is imperative to maximize the utility of such models by expanding their application across diverse domains, thereby fully unleashing their capabilities.

2.2. Event simulation

Event simulation is proposed to generate events from existing image and video datasets, effectively reducing the cost and labor of data collection [15, 22, 31, 57]. There are two main categories: image-to-event simulation and video-to-event simulation. For image-to-event, users take an event camera to shoot the images with a light movement during event capturing, where the image annotation is shared between two modalities. There are extensive image-to-event datasets, such as N-Caltech101 [36], CIFAR10-DVS [26], N-ImageNet [24]. This method still needs an event camera to record the activated event, thereby showing high hardware requirements. For video-to-event, it directly converts a video stream into event data [11, 19, 29, 38, 54], eliminating the necessity of event cameras. ESIM [38] is the first event camera simulator, which employs an adaptive sampling strategy to interpolate frames and then calculate the intensity difference between adjacent pixels to generate events. VID2E [11] employs the event simulator introduced in ESIM [38] to generate events, however, adopting the optical flow between pairs of video frames to interpolate intermediate samples. V2E [19] generates events from videos by considering the temporal noise, leak events, and pixel threshold mismatch, enhancing the realness of simu-

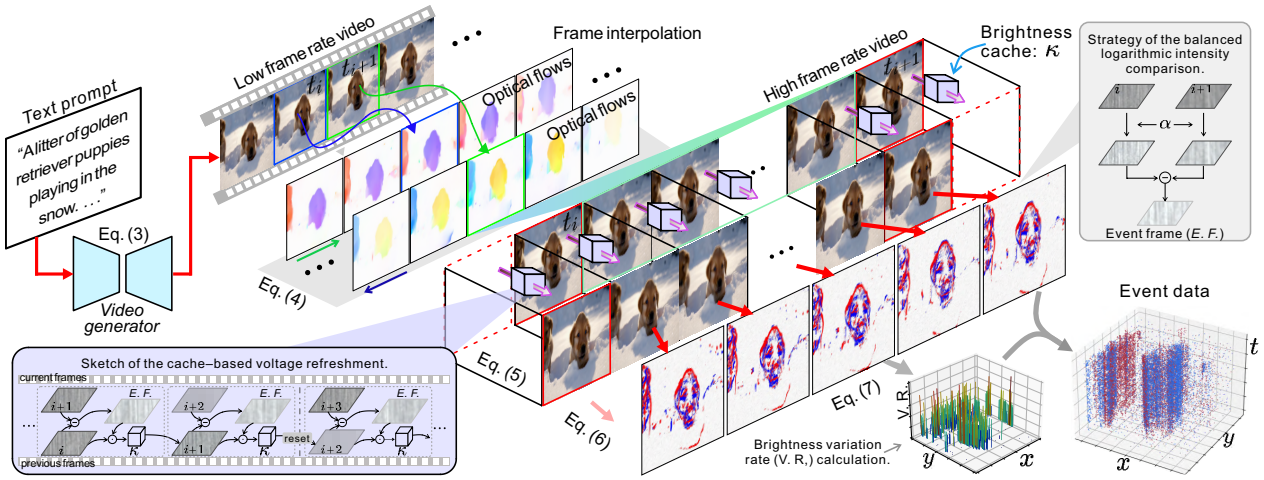

Figure 2. Framework of Texvent, including the high frame-rate video generation (Eq. 3, Eq. 4) and event simulation. During computing the event frame (E. F.), we present the brightness cache to store the brightness values at coordinates where no event data has been activated. These values still serve as the reference brightness values in the subsequent event frame generation (Eq. 5). Such a cache is periodically reset as null to avoid acting fake events in long-term event simulation. After injecting the background noise (Eq. 6), we calculate the brightness variation rate (V. R.) at each coordinate to reconstruct the dense time stamps (Eq. 7).

lated events. V2CE [54] proposes a local dynamic-aware timestamp inference strategy to recover event timestamps that enables the event to be distributed normally. By modeling the voltage variation as a Brownian motion with drift, DVS-Voltmeter [29] simulates event data from videos under a stochastic process that enhances the realness.

Although these conversion toolboxes can easily convert a video into event data, personalizing or generating the specific event streams is still challenging. Therefore, text-to-event is proposed to enable different users to generate their event data with a text prompt. Ott *et al*. [37] employ the diffusion model to embed the text prompts and train an autoencoder to reconstruct the event frames from the latent representations. It requires not only labeling the text prompt for each event sample but also converting the event stream into event frames to train the autoencoder. This limitation restricts users from personalizing their event data. Here, we propose harnessing the power of MLLMs to simulate event data, enabling open-world event simulation.

3. Methodology

3.1. Problem formulation

Event camera captures the visual content via the brightness variation, asynchronously generating the event data [47]. Specifically, an event $e_i = (\mathbf{x}_i, t_i, p_i)$ is activated when the log brightness change at pixel $\mathbf{x}_i = (x_i, y_i)$ has exceeded the contrast threshold δ, formulated as:

$$|L(\mathbf{x}_i, t_i) - L(\mathbf{x}_i, t_i - \Delta_t)| \geq \delta. \tag{1}$$

In Eq. 1, t_i denotes the time stamp. Δ_t is a time interval, indicating the minimum temporal resolution of an event sensor. $p_i \in \{-1, +1\}$ is the sign of the brightness change, called polarity. $p = +1$ denotes the positive event where the change of the logarithmic brightness is higher than $+\delta$. Conversely, $p = -1$ represents the negative event.

The objective of event simulation is to generate event data from existing digital contents (*e.g*., images, videos, *etc*.) instead of physical capturing, effectively reducing the cost of data collection [38]. Recently, with the increasing popularity of text-to-X generation, generating event data from text has become attractive, aka text-to-event (T2E) [37]. T2E generation can be formulated as:

$$\mathcal{E} = \mathcal{F}(\mathbf{T}), \tag{2}$$

where $\mathcal{E} = \{e_i\}_{i=1}^{n}$ denotes the generated event data, n is number of the generated events, $\mathcal{F}(\cdot)$ indicates the T2E generation function, and $\mathbf{T}$ is the text prompt that depicts the desired visual contents. In developing event simulation methodologies, two critical requirements warrant careful consideration. 1) Efficiency: The simulation method must execute with sufficient speed to simulate event data without introducing perceptible latency for end users. 2) Fidelity: The simulated events must maintain statistical consistency with real data to ensure the reliability of downstream tasks. These dual requirements fundamentally inform the design principles of our simulation framework.

3.2. Texvent

3.2.1. High frame-rate video generation

Prompt-driven video generation. As shown in Fig. 2, we synthesize the event data via a text-to-video generation model. Such a model first encodes the text prompt $\mathbf{T}$

as tokens and then decodes them as the image sequence: $\mathbf{I}_{t_{\{1:N\}}} \in \mathbb{R}^{N\times H\times W\times 3}$, which can be formulated as:

$$\mathbf{I}_{t_{\{1:N\}}} = D(E(\mathbf{T};\theta_e);\theta_d), \tag{3}$$

where N, H, and W denote the number, height, and width of video frames. $D(\cdot;\theta_d)$ and $E(\cdot;\theta_e)$ are the image decoder and text prompt encoder, respectively. Typically, videos generated by $D(\cdot;\theta_d)$ show a lower temporal resolution compared to the event camera data. Therefore, we need to interpolate these videos to achieve higher frame-rates, thereby reducing the potential event loss during the following event simulation process.

Brightness-aware frame interpolation. For frame interpolation, a critical issue is to determine the number of intermediate samples [11], as this impacts both the simulation efficiency and the time stamp of the final generated event data. In [11, 19, 38], this number is set based on bidirectional optical flows, where the relative displacement between intermediate frames is at most 1 pixel. Obviously, this adaptive manner leads to low efficiency once the video contains chaotic lighting variations. To address it, we propose an adaptive brightness-aware interpolation strategy to value the number via the brightness variation between two consecutive frames. Specifically, the number of intermediate frames (K_i) between $\mathbf{I}_{t_i}$ and $\mathbf{I}_{t_{i+1}}$ is denoted as:

$$K_i = \max(|L(\mathbf{I}_{t_i}) - L(\mathbf{I}_{t_{i+1}})|) \bmod \delta. \tag{4}$$

If there is an obvious brightness variation, K_i equally spaced intermediate frames are interpolated by RIFE [20], achieving a high temporal resolution. Conversely, if the brightness variation between $(\mathbf{I}_{t_i}, \mathbf{I}_{t_{i+1}})$ is lower than the contrast threshold δ, it indicates that no events are activated. Therefore, we don't need to interpolate these two video frames, thereby increasing the efficiency. After this adaptive interpolation, we have achieved a high frame-rate video ready for event data simulation.

3.2.2. Event data simulation

Event frame generation. The event camera circuit generates events by measuring the voltage variations between the reference voltage and current caused by the light signal hitting the pixel on the photoreceptor [19]. To simulate events, we compute the intensity signal variations at each pixel sampled from the video sequence with a high frame rate. The events are generated when the intensity variation exceeds the contrast threshold δ. The detailed event simulation formulation is as follows:

$$\prod_{i\in\{0:N-1\}} \prod_{j\in\{1:K_i\}} \left|L(\alpha+\mathbf{I}^{j}_{(t_i,t_{i+1})}) - L(\alpha+\kappa\diamond\mathbf{I}^{j-1}_{(t_i,t_{i+1})})\right| > \delta, \tag{5}$$

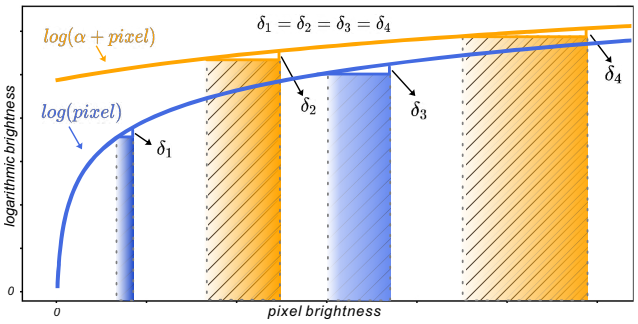

Figure 3. **Blue**: The original logarithmic function shows high sensitivity for low light variations. For the same contrast thresholds ($\delta_1 = \delta_3$), an event is triggered in low-light conditions with only a minimal brightness variation; while in high-light conditions, a variation of nearly four times is necessary to activate the event. **Orange**: Adding a balancing parameter (α) can alleviate this unfair event activation between the low- and high-light conditions. For the same contrast thresholds ($\delta_2 = \delta_4$), almost twice the brightness variation of low-light conditions can activate an event in high-light environments.

where α and κ denote a balancing parameter and brightness cache, respectively. $\diamond$ indicates the brightness updating operation. $L(\cdot)$ is the logarithmic function that simulates human retinas to calculate the intensity variation.

Such a mechanism leads to the event sensor showing higher sensitivity for low light variations than high light conditions. However, conventional videos are typically captured using low dynamic range vision sensors, which struggle to accurately capture details in high-light areas. Thus, it's hard to simulate the event data from the highlighted objects in the video. As shown in Fig. 3, the brightness variation in high light conditions is almost three times that of low light, which can successfully generate an event (**Blue**). To balance this activation sensitivity, we add a *balancing parameter* α during computing the pixel brightness variation. This strategy can effectively alleviate the imbalance in different light conditions while retaining the property of biological retinas, as shown in Fig. 3 (**Orange**).

Instead of directly calculating the brightness variation from two consecutive frames, we compare the current frame with a calibrated frame to generate events. This approach can effectively prevent the loss of potential event data caused by frequent brightness updating. Specifically, a brightness cache κ is presented to store the past brightness for generating the calibrated frame. We sample the brightness value at $\mathbf{x}$ without activating events yet from κ to update the brightness at the same coordinates of $\mathbf{I}^{j-1}_{(t_i,t_{i+1})}$. This operation denotes that we only modify the reference brightness at a pixel where an event is activated, consistent with the voltage updating of the event camera circuit [27]. Based on thresholding the intensity signal variations, we

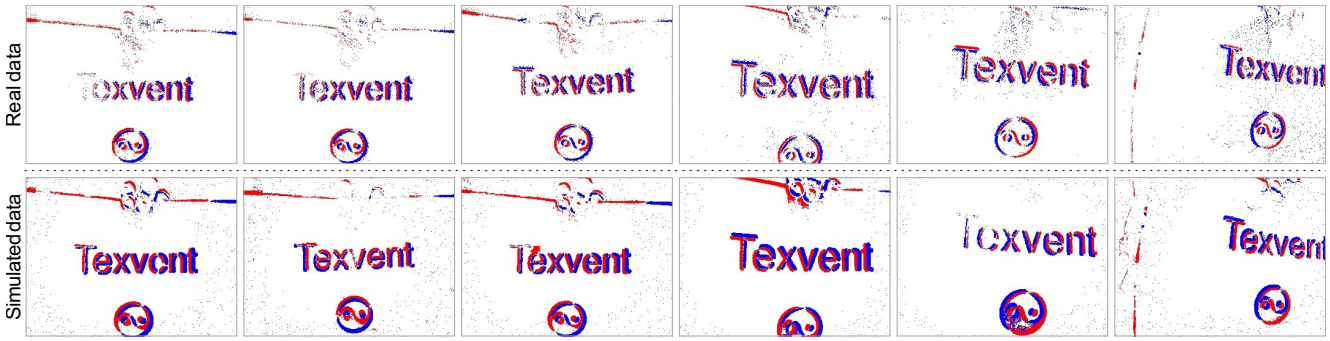

Figure 4. Comparison between the real data (1st row) and our simulated data (2nd row). The event-video collection system consists of a DAVIS346 sensor and an RGB camera (480p, 30fps) followed by [7]. Blue and red denote the positive and negative events, respectively.

can obtain a series of event frames that contain information about coordinates, polarities, and compressed time stamps. However, these event frames only contain actual positive and negative events, showing a noticeable difference from the real event data. We need to add noise to this clean simulated data to reduce the sim-to-real gap.

Noise addition. For an event camera, the pixel photoreceptor and change detector circuits are easily affected by the temperature [35], resulting in Background Activity (BA) noise [14]. To narrow the gap between the simulated and real event data, VID2E [11] randomly samples contrast thresholds for positive and negative events to introduce BA noise. However, setting the dynamic contrast thresholds may lead to disregarding real events. Therefore, we propose to add additional Poisson noise to enhance the practicality under two considerations. 1) For noise intensity, we generate the Poisson noise based on the fill factor of the event sensor, ensuring that different event sensors produce varying levels of noise. 2) For injection position, we suggest an adaptive injection strategy that prioritizes adding noise to the low-light background regions, as lower brightness makes noise activation easier. The detailed perturbation process is denoted as:

$$\mathcal{E} = \mathcal{E} \cdot (1 - \mathbf{M}) + \mathbf{M} \cdot \textbf{Poisson}(\lambda_1 \lambda_2), \tag{6}$$

where $\mathbf{M} = (\mathbf{I}_{t_{i+1}} < \sigma) \cdot (\Delta_L < \delta)$ denotes the mask that locates the low-light ($\mathbf{I}_{t_{i+1}} < \sigma$) and background ($\Delta_L < \delta$) region of the simulated event dat $\mathcal{E}$. Parameter σ is introduced to divide the high- and low-light areas from the current frame $\mathbf{I}_{t_{i+1}}$. $\Delta_L = |L(\alpha + \mathbf{I}_{t_{i+1}}) - L(\alpha + \kappa \diamond \mathbf{I}_{t_i})|$ is the absolute brightness difference between two consecutive frames. $\circledcirc$ denotes adding background noise into the background region of the event data to avoid corrupting the real event data. **Poisson**$(\cdot)$ indicates the noise generation function, where λ_1 denotes the probability of noise event activation at a pixel and λ_2 is a scaling parameter for fitting other sensor parameters.

Time stamp reconstruction. After noise addition, we reconstruct the dense time stamps for each event contained in the event frame. The time stamp denotes the time of an event when it is activated, which is mainly decided by the rate of brightness variation. Therefore, a reasonable assumption is that the larger the brightness change in a fixed time bin, the earlier the event data is activated. We focus on employing the brightness variation calculated by Eq. 5 for time stamp reconstruction. Specifically, the time stamp of each simulated event data is denoted as:

$$t^{\mathbf{x}_i} = \gamma \times (t_{i+1} - t_i)(1 - \frac{\Delta_L^{\mathbf{x}_i} - \min(\Delta_L)}{\max(\Delta_L) - \min(\Delta_L)}) + t_i, \tag{7}$$

where Δ_L denotes the absolute brightness difference and γ indicates the scaling parameter to ensure each event is activated at the microsecond level. t_i and t_{i+1} denote the time stamp of two consecutive frames, respectively. $\mathbf{x}_i = (x_i, y_i)$ is the coordinate. The polarity p is $+1$ when the logarithmic brightness change is higher than $+\delta$, while $p = -1$ once the change is lower than $-\delta$. Finally, a new data $e = \{\mathbf{x}_i, t_i, p_i\}$ can be obtained by incorporating these three components. By assembling a sequence of such events, we achieve the event simulation from a text prompt.

3.3. Implementation details

We implement our training-free text-to-event simulation framework (Texvent) using Cosmos-1.0-Diffusion-7B-Text2World [2], having also experimented with Wan [45], CogVideoX [51], and Open-Sora [55]. A single NVIDIA H100 GPU is employed to render the event data. We set the accumulation time to $1/FPS$ to aggregate individual events into an event stream. It's fixed across all experiments. The FPS of Cosmos, Wan, CogVideoX, and Open-Sora are 24, 16, 16, and 24, respectively. For the depth map and warped event, E2Depth [16] and Contrast Maximization Framework [43] are employed, respectively.

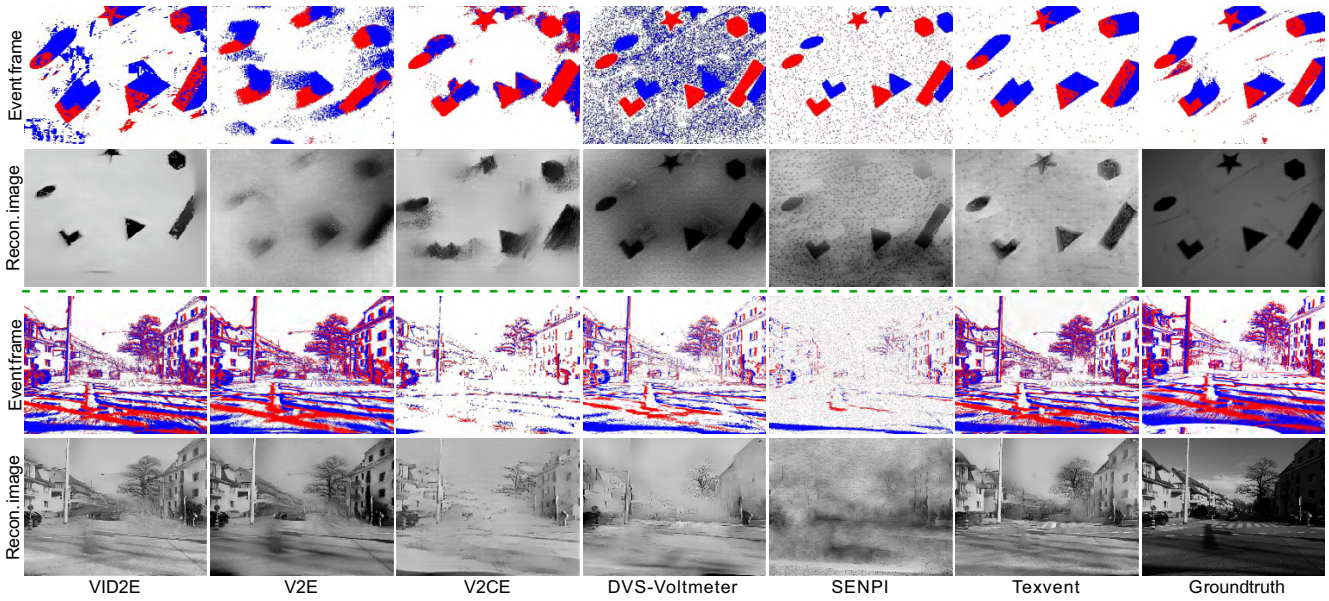

Figure 5. Visualization results of the event frame and its corresponding reconstructed (Recon.) image. E2VID [39] is employed to convert the event data into images. Blue and red denote the positive and negative events, respectively.

Table 1. Quantitative evaluation of event frames (E. F.) / reconstructed images (R. I.) in terms of mean squared error (MSE), structural similarity (SSIM), and the calibrated perceptual loss (LPIPS). The best and second-best scores are highlighted in **bold** and underlined.

	VID2E [11] E. F. / R. I.	V2E [19] E. F. / R. I.	V2CE [54] E. F. / R. I.	DVS-Voltmeter [29] E. F. / R. I.	SENPI [13] E. F. / R. I.	Texvent E. F. / R. I.
MSE↓	0.116 / 0.188	0.142 / 0.099	0.082 / 0.151	0.276 / **0.096**	0.186 / 0.104	**0.045** / 0.116
SSIM↑	0.430 / 0.387	0.299 / 0.420	**0.552** / 0.392	0.085 / 0.149	0.095 / 0.251	0.488 / **0.472**
LPIPS↓	0.406 / 0.381	0.603 / 0.422	0.383 / 0.451	0.972 / 0.354	0.820 / 0.561	**0.339** / **0.296**

4. Experiment

4.1. Experimental setup

Dataset. To evaluate various event simulators, the video-to-event datasets: Event Camera Dataset (ECD) [33] and DSEC [12], are employed in our experiments. Additionally, we collect a text-event pair dataset: NT-ImageNet to test the performance of our method on text-to-event simulation. Specifically, the event streams are directly sampled from the validation part of N-ImageNet [24]. Then, we sample corresponding images from the ImageNet dataset [42] and employ the LLaVA-v1.5-13B [30] with the text prompt: "Caption the image in detail." to generate the text prompt.

Baselines. We adopt current video-to-event methods: VID2E [11], V2E [19], V2CE [54], DVS-Voltmeter [29], and SENPI [13], in our experiments. The method [37] is not included due to a lack of official implementations. For video generators, we test various multimodal large language models, including Cosmos [2], Wan [45], Open-Sora [55], and CogVideoX [51]. Image reconstruction methods include E2VID [39], HyperE2VID [8], and ETNet [48].

Evaluation. We present *three* evaluation strategies to evaluate our method. For *frame-level*, we test the image metrics [47] of the event-to-video reconstruction results to assess the temporal consistency and visual quality of our simulated event data. The image metrics are PSNR, SSIM, LPIPS, and MSE, respectively. For *point-level*, we test the Event Quality Score (EQS) [5] on our collected NT-ImageNet dataset to quantitatively measure the fidelity and statistical characteristics between the simulated and real event data. For *application-level*, we test our simulated event data on downstream tasks (*e.g.*, objection recognition, image reconstruction, depth estimation, *etc.*) to validate its practical utility across various event-based models.

4.2. Comparison on video-to-event

Real-world evaluation. To validate our event simulator in real-world scenarios, we constructed a dual-camera system (DAVIS346 sensor + RGB camera) followed by [7]. The RGB camera footage is directly input into our simulator, generating synthetic events. This experimental setup allowed us to directly evaluate the fidelity of our simulated

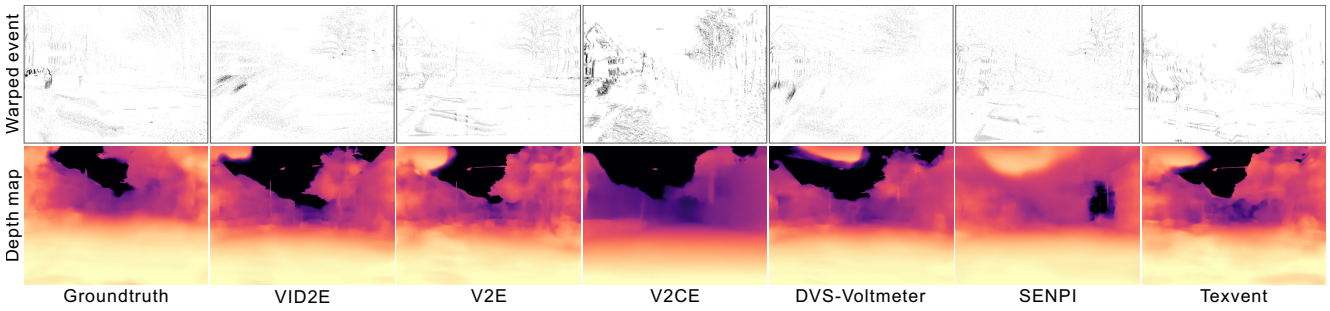

Figure 6. Warped event and depth map of simulated event data. Discrepancies between groundtruth and simulated event data arise from the misalignment between raw event data and corresponding video sequences in the DSEC dataset [12].

events against real event data under practical conditions. As shown in Fig. 4, we observed that our simulated events closely matched the temporal dynamics and spatial distribution characteristics of the real event data. This real-world validation demonstrates the practical applicability and reliability of our simulation approach.

Qualitative evaluation. To evaluate our simulator for video-to-event generation, we compare it to VID2E [11], V2E [19], V2CE [54], DVS-Voltmeter [29], and SENPI [13] on ECD [33] and DSEC [12] datasets. The simulated event and the corresponding reconstructed image of each simulator are shown in Fig. 5. VID2E [11] and V2E [19] generate relatively sparse events with noticeable gaps in the event distribution. V2CE [54] produces events in a low temporal resolution that leads to some events being lost in the motion parts, neither for the SENPI [13]. DVS-Voltmeter [29] shows scattered noise-like patterns that corrupt natural event distributions. In comparison, our simulator produces event patterns that more closely resemble the ground truth, with balanced density and clear object boundaries. As for the fidelity of simulated data from the reconstruction level, our method still performs better than other comparison baselines. V2E [19] shows some blurring and loss of details in the reconstructed image.

Quantitative evaluation. Table 1 quantitatively evaluates the event frames and reconstructed images of different simulators in terms of MSE, SSIM, and LPIPS. For event frames, our method achieves the best MSE (0.045) and competitive SSIM (0.488), while maintaining the lowest LPIPS score by 0.339, indicating high-quality event simulation. For reconstructed images, our approach shows balanced performance with the highest SSIM (0.472) and the best LPIPS (0.296). DVS-Voltmeter [29] achieves the lowest MSE by 0.096, which is slightly better than our simulator by 0.02. These results validate that our method not only generates accurate event data but also ensures high-fidelity image reconstruction, outperforming existing methods.

Table 2. Event Quality Score (EQS) [5] and runtime of different simulators. The best and second-best scores are highlighted in **bold** and underlined. DVS. denotes the DVS-Voltmeter.

	VID2E [11]	V2E [19]	V2CE [54]	DVS. [29]	SENPI [13]	Texvent
EQS↑	0.8597	0.8138	0.8642	0.8573	0.8824	**0.8851**
Time (s)	2.1228	2.1652	0.0950	0.6919	**0.0573**	0.0653

Table 3. Comparison to various image reconstruction methods without (✗) and with (✓) our augmented event data on Event Camera dataset [33]. Only 5% event data are generated by Texvent in this experiment. Best results are **bolded**.

Method	✗ / ✓	PSNR↑	SSIM↑	LPIPS↓	MSE↓
E2VID [39]	✗	22.1127	0.5897	0.3104	0.0070
	✓	**22.8475**	**0.6320**	**0.2791**	**0.0063**
HyperE2VID [8]	✗	21.9488	0.5973	0.2421	0.0077
	✓	**23.3000**	**0.6624**	**0.1656**	**0.0061**
ETNet [48]	✗	21.1987	0.5629	**0.2013**	0.0087
	✓	**21.3771**	**0.5630**	0.2112	**0.0084**

4.3. Comparison on text-to-event

Fig. 1 presents a comparative analysis between real event data captured by SAMSUNG Gen3 [24] (first column) and our simulated results (subsequent columns). The qualitative results demonstrate that our simulator generates event data with high fidelity across diverse scenarios, including dynamic scenes such as ostriches traversing grassland, sharks moving through oceanic environments, *etc*. Table 2 shows quantitative results of our method and five baselines in terms of event quality score (EQS) [5] and efficiency. Our method achieves the highest EQS by 0.8851 on the NT-ImageNet dataset, which is better than the SENPI [13] by 2.7%. Among these six simulators, SENPI [13] is the fastest method to simulate the event data from a video sequence, taking 0.0573 seconds to process each frame pair. Our method is the second fastest, which only spends 0.0653 seconds generating an event stream. VID2E [11] and V2E [19] show low efficiency due to repetitive optical flow estimation during frame interpolation.

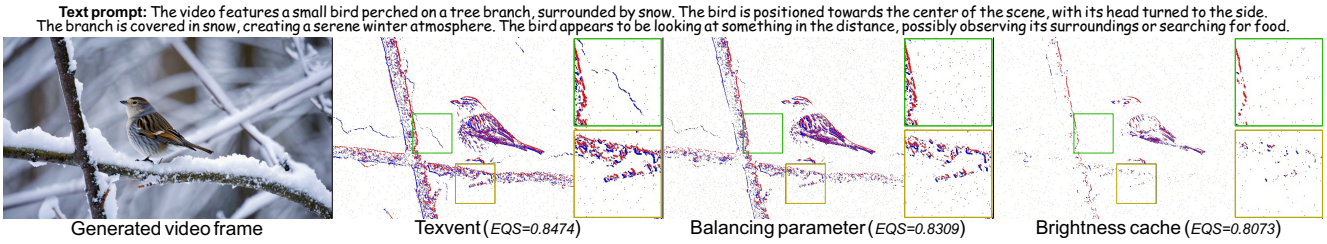

Figure 7. Ablation study about the balancing parameter and brightness cache. Quantitative results are shown in the Supplementary Material. Blue and red denote the positive and negative events, respectively.

4.4. Comparison on downstream task

Image reconstruction. As shown in Table 3, the quality of reconstructed images has been successfully improved with only augmenting 5% new cases. Among the methods, HyperE2VID [8] outperforms the others, achieving the highest PSNR and SSIM, and the lowest LPIPS and MSE. The improvement percentages are 6.16%, 10.9%, 31.6% and 20.9%, respectively. E2VID [39] and ETNet show moderate performance, but almost all metrics are refined by employing our augmented event data. Only the LPIPS of ETNet [48] is reduced by 0.0099, likely due to limited training on perceptual details. Totally, Table 3 suggests that even a small increase in training data can lead to substantial improvements in image reconstruction quality, underscoring the value of Texvent in this context.

Depth map estimation. Here, we evaluate the quality of various simulated event data from views of depth map estimation, as shown in Fig. 6. Totally, Texvent performs better than other baselines. The warped events are the cleanest and most sharply aligned, with minimal ghosting. And the depth map is smooth yet preserves scene structure with clear near–far separation. V2CE [54] follows, showing reasonably sharp warped events and a coherent depth map, but with more sparsity and small artifacts at edges. This likely stems from its lower resolution. VID2E [11] and V2E [19] exhibit noticeable misalignment in the warped events (double edges, missing details), streaks, and holes in the depth maps. DVS-Voltmeter [29] achieves the noisy warped event and a coarse depth map. SENPI [13] lands in the middle, recovering some structure but with residual ghosting in warped events and blocky artifacts in the depth.

5. Discussion

The event camera employs the logarithmic function to model the sensitivity difference for low- and high-light conditions. However, for event simulation, the video is captured by a conventional camera that shows fewer high-light information due to its low dynamic range. As a result, the noticeable edge changes cannot be captured by directly using the unbalanced logarithmic function (green box in Fig. 7). The EQS [5] is decreased from 0.8474 to 0.8309. In the event camera circuit, the reference voltage is updated only when an event is activated [19]. To prevent the potential event loss caused by the frame-by-frame intensity calculation, the brightness cache is proposed. As depicted in the orange box of Fig. 7, some events are not activated when removing such a cache, resulting in a substantial EQS drop of 4.01%, showcasing the importance of our brightness cache.

6. Conclusion

In this paper, we introduce a new text-to-event simulation framework, Texvent, utilizing multimodal large language models. Our approach requires no training and demonstrates high efficiency and accuracy in video-to-event simulations. Moreover, the proposed event simulator is developed as a plug-and-play module that can be easily compatible with different video generation models and real standard cameras. A new text-event pair dataset is collected for evaluating the text-to-event simulators. Comprehensive experiments on various datasets and real-world scenarios effectively showcase the advantages of our method.

7. Acknowledgment

This work was carried out at the Renjie Group. Renjie Group is supported by the National Natural Science Foundation of China under Grant No. 62302415, Guangdong Basic and Applied Basic Research Foundation under Grant No. 2024A1515012822, and the Research Grant Council (RGC) of Hong Kong SAR, under a GRF Grant 12203124 and an ECS Grant 22201125. This work was also supported by Beijing Natural Science Foundation (Grant No. L233024), Beijing Municipal Science & Technology Commission, Administrative Commission of Zhongguancun Science Park (Grant No. Z241100003524012), and National Natural Science Foundation of China (Grant No. 62136001). Peiqi Duan was supported by National Natural Science Foundation of China (Grant No. 62402014), China National Postdoctoral Program for Innovative Talents (Grant No. BX20230010) and China Postdoctoral Science Foundation (Grant No. 2023M740076). This work was also supported by Beijing Municipal Science & Technology Program No. Z251100007125021.

References

[1] OpenAI. https://openai.com/sora/, 2024. 2

[2] Niket Agarwal, Arslan Ali, Maciej Bala, Yogesh Balaji, Erik Barker, Tiffany Cai, Prithvijit Chattopadhyay, Yongxin Chen, Yin Cui, Yifan Ding, et al. Cosmos world foundation model platform for physical AI. *arXiv preprint arXiv:2501.03575*, 2025. 2, 5, 6

[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 *Proc. CVPR*, pages 22563–22575, 2023. 2

[4] Bharatesh Chakravarthi, Aayush Atul Verma, Kostas Daniilidis, Cornelia Fermuller, and Yezhou Yang. Recent event camera innovations: A survey. In *Proc. ECCVW*, pages 342–376, 2024. 2

[5] Kaustav Chanda, Aayush Verma, Arpitsinh Vaghela, Yezhou Yang, and Bharatesh Chakravarthi. Event quality score (eqs): Assessing the realism of simulated event camera streams via distance in latent space. In *Proc. CVPRW*, pages 5105–5113, 2025. 1, 6, 7, 8

[6] Peiqi Duan, Zihao W Wang, Xinyu Zhou, Yi Ma, and Boxin Shi. EventZoom: Learning to denoise and super resolve neuromorphic events. In *Proc. CVPR*, pages 12824–12833, 2021. 1

[7] Peiqi Duan, Boyu Li, Yixin Yang, Hanyue Lou, Minggui Teng, Xinyu Zhou, Yi Ma, and Boxin Shi. EventAid: Benchmarking event-aided image/video enhancement algorithms with real-captured hybrid dataset. *TPAMI*, 2025. 5, 6

[8] Burak Ercan, Onur Eker, Canberk Saglam, Aykut Erdem, and Erkut Erdem. Hypere2vid: Improving event-based video reconstruction via hypernetworks. *TIP*, 33:1826–1837, 2024. 6, 7, 8

[9] Guillermo Gallego, Tobi Delbrück, Garrick Orchard, Chiara Bartolozzi, Brian Taba, Andrea Censi, Stefan Leutenegger, Andrew J Davison, Jörg Conradt, Kostas Daniilidis, et al. Event-based vision: A survey. *TPAMI*, 44(1):154–180, 2020. 1

[10] Songwei Ge, Seungjun Nah, Guilin Liu, Tyler Poon, Andrew Tao, Bryan Catanzaro, David Jacobs, Jia-Bin Huang, Ming-Yu Liu, and Yogesh Balaji. Preserve your own correlation: A noise prior for video diffusion models. In *Proc. ICCV*, pages 22930–22941, 2023. 2

[11] Daniel Gehrig, Mathias Gehrig, Javier Hidalgo-Carrió, and Davide Scaramuzza. Video to events: Recycling video datasets for event cameras. In *Proc. CVPR*, pages 3586–3595, 2020. 1, 2, 4, 5, 6, 7, 8

[12] Mathias Gehrig, Willem Aarents, Daniel Gehrig, and Davide Scaramuzza. DSEC: A stereo event camera dataset for driving scenarios. *RA-L*, 6(3):4947–4954, 2021. 6, 7

[13] Joseph L Greene, Adrish Kar, Ignacio Galindo, Elijah Quiles, Elliott Chen, and Matthew Anderson. A pytorch-enabled tool for synthetic event camera data generation and algorithm development. In *Synthetic Data for Artificial Intelligence and Machine Learning: Tools, Techniques, and Applications III*, pages 117–137. SPIE, 2025. 6, 7, 8

[14] Shasha Guo and Tobi Delbruck. Low cost and latency event camera background activity denoising. *TPAMI*, 45(1):785–795, 2022. 5

[15] Haiqian Han, Jiacheng Lyu, Jianing Li, Henglu Wei, Cheng Li, Yajing Wei, Shu Chen, and Xiangyang Ji. Physical-based event camera simulator. In *Proc. ECCV*, pages 19–35, 2024. 2

[16] Javier Hidalgo-Carrió, Daniel Gehrig, and Davide Scaramuzza. Learning monocular dense depth from events. In *3DV*, pages 534–542, 2020. 5

[17] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. In *Proc. NeurIPS*, 2022. 2

[18] Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. CogVideo: Large-scale pretraining for text-to-video generation via transformers. In *Proc. ICLR*, 2023. 2

[19] Yuhuang Hu, Shih-Chii Liu, and Tobi Delbruck. V2E: From video frames to realistic dvs events. In *Proc. CVPR*, pages 1312–1321, 2021. 1, 2, 4, 6, 7, 8

[20] Zhewei Huang, Tianyuan Zhang, Wen Heng, Boxin Shi, and Shuchang Zhou. Real-time intermediate flow estimation for video frame interpolation. In *Proc. ECCV*, pages 624–642, 2022. 4

[21] Jianping Jiang, Jiahe Li, Baowen Zhang, Xiaoming Deng, and Boxin Shi. Evhandpose: Event-based 3d hand pose estimation with sparse supervision. *TPAMI*, 46(9):6416–6430, 2024. 1

[22] Damien Joubert, Alexandre Marcireau, Nic Ralph, Andrew Jolley, André Van Schaik, and Gregory Cohen. Event camera simulator improvements via characterized parameters. *Frontiers in Neuroscience*, 15:702765, 2021. 2

[23] Johanna Karras, Aleksander Holynski, Ting-Chun Wang, and Ira Kemelmacher-Shlizerman. Dreampose: Fashion image-to-video synthesis via stable diffusion. In *Proc. ICCV*, pages 22623–22633, 2023. 2

[24] Junho Kim, Jaehyeok Bae, Gangin Park, Dongsu Zhang, and Young Min Kim. N-imagenet: Towards robust, fine-grained object recognition with event cameras. In *Proc. ICCV*, pages 2146–2156, 2021. 2, 6, 7

[25] Dan Kondratyuk, Lijun Yu, Xiuye Gu, Jose Lezama, Jonathan Huang, Grant Schindler, Rachel Hornung, Vighnesh Birodkar, Jimmy Yan, Ming-Chang Chiu, Krishna Somandepalli, Hassan Akbari, Yair Alon, Yong Cheng, Joshua V. Dillon, Agrim Gupta, Meera Hahn, Anja Hauth, David Hendon, Alonso Martinez, David Minnen, Mikhail Sirotenko, Kihyuk Sohn, Xuan Yang, Hartwig Adam, Ming-Hsuan Yang, Irfan Essa, Huisheng Wang, David A Ross, Bryan Seybold, and Lu Jiang. VideoPoet: A large language model for zero-shot video generation. In *Proc. ICML*, 2024. 2

[26] Hongmin Li, Hanchao Liu, Xiangyang Ji, Guoqi Li, and Luping Shi. CIFAR10-DVS: an event-stream dataset for object classification. *Frontiers in neuroscience*, 11:309, 2017. 2

[27] Patrick Lichtsteiner, Christoph Posch, and Tobi Delbruck. A 128×128 120db 15μs latency asynchronous temporal contrast vision sensor. *JSSC*, 43(2):566–576, 2008. 4

[28] Han Lin, Abhay Zala, Jaemin Cho, and Mohit Bansal. VideodirectorGPT: Consistent multi-scene video generation via LLM-guided planning. In *COLM*, 2024. 2

[29] Songnan Lin, Ye Ma, Zhenhua Guo, and Bihan Wen. Dvs-voltmeter: Stochastic process-based event simulator for dynamic vision sensors. In *Proc. ECCV*, pages 578–593, 2022. 2, 3, 6, 7, 8

[30] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Proc. NeurIPS*, 2023. 6

[31] Hanyue Lou, Jinxiu Liang, Minggui Teng, Yi Wang, and Boxin Shi. V2V: Scaling event-based vision through efficient video-to-voxel simulation. In *Proc. NeurIPS*, 2025. 2

[32] Ziyuan Luo, Yangyi Zhao, Ka Chun Cheung, Simon See, and Renjie Wan. ImageSentinel: Protecting visual datasets from unauthorized retrieval-augmented image generation. In *Proc. NeurIPS*, 2025. 2

[33] Elias Mueggler, Henri Rebecq, Guillermo Gallego, Tobi Delbruck, and Davide Scaramuzza. The event-camera dataset and simulator: Event-based data for pose estimation, visual odometry, and slam. *IJRR*, 36(2):142–149, 2017. 6, 7

[34] Bolin Ni, Houwen Peng, Minghao Chen, Songyang Zhang, Gaofeng Meng, Jianlong Fu, Shiming Xiang, and Haibin Ling. Expanding language-image pretrained models for general video recognition. In *Proc. ECCV*, pages 1–18, 2022. 2

[35] Yuji Nozaki and Tobi Delbruck. Temperature and parasitic photocurrent effects in dynamic vision sensors. *TED*, 64(8): 3239–3245, 2017. 5

[36] Garrick Orchard, Ajinkya Jayawant, Gregory K Cohen, and Nitish Thakor. Converting static image datasets to spiking neuromorphic datasets using saccades. *Frontiers in neuroscience*, 9:437, 2015. 2

[37] Joachim Ott, Zuowen Wang, and Shih-Chii Liu. Text-to-Events: Synthetic event camera streams from conditional text input. In *NICE*, pages 1–10, 2024. 2, 3, 6

[38] Henri Rebecq, Daniel Gehrig, and Davide Scaramuzza. ESIM: an open event camera simulator. In *CoRL*, pages 969–982, 2018. 1, 2, 3, 4

[39] Henri Rebecq, René Ranftl, Vladlen Koltun, and Davide Scaramuzza. Events-to-video: Bringing modern computer vision to event cameras. In *Proc. CVPR*, pages 3857–3866, 2019. 6, 7, 8

[40] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proc. CVPR*, pages 10684–10695, 2022. 2

[41] Viktor Rudnev, Mohamed Elgharib, Christian Theobalt, and Vladislav Golyanik. EventNeRF: Neural radiance fields from a single colour event camera. In *Proc. CVPR*, pages 4992–5002, 2023. 1

[42] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *IJCV*, 115:211–252, 2015. 6

[43] Shintaro Shiba, Yannick Klose, Yoshimitsu Aoki, and Guillermo Gallego. Secrets of event-based optical flow, depth and ego-motion estimation by contrast maximization. *TPAMI*, 46(12):7742–7759, 2024. 5

[44] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-A-Video: Text-to-video generation without text-video data. In *Proc. ICLR*, 2023. 2

[45] Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025. 2, 5, 6

[46] Ruofei Wang, Qing Guo, Haoliang Li, and Renjie Wan. Event Trojan: Asynchronous event-based backdoor attacks. In *Proc. ECCV*, pages 315–332, 2024. 1

[47] Ruofei Wang, Peiqi Duan, Boxin Shi, and Renjie Wan. Asynchronous event error-minimizing noise for safeguarding event dataset. In *Proc. ICCV*, pages 10141–10150, 2025. 3, 6

[48] Wenming Weng, Yueyi Zhang, and Zhiwei Xiong. Event-based video reconstruction using transformer. In *Proc. ICCV*, pages 2563–2572, 2021. 6, 7, 8

[49] Shuai Yang, Yifan Zhou, Ziwei Liu, and Chen Change Loy. Rerender a video: Zero-shot text-guided video-to-video translation. In *SIGGRAPH Asia*, pages 1–11, 2023. 2

[50] Yixin Yang, Jinxiu Liang, Bohan Yu, Yan Chen, Jimmy S Ren, and Boxin Shi. Latency correction for event-guided deblurring and frame interpolation. In *Proc. CVPR*, pages 24977–24986, 2024. 1

[51] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, Da Yin, Yuxuan.Zhang, Weihan Wang, Yean Cheng, Bin Xu, Xiaotao Gu, Yuxiao Dong, and Jie Tang. CogVideoX: Text-to-video diffusion models with an expert transformer. In *Proc. ICLR*, 2025. 2, 5, 6

[52] Shengming Yin, Chenfei Wu, Jian Liang, Jie Shi, Houqiang Li, Gong Ming, and Nan Duan. Dragnuwa: Fine-grained control in video generation by integrating text, image, and trajectory. *arXiv preprint arXiv:2308.08089*, 2023. 2

[53] Bohan Yu, Jieji Ren, Jin Han, Feishi Wang, Jinxiu Liang, and Boxin Shi. EventPS: Real-time photometric stereo using an event camera. In *Proc. CVPR*, pages 9602–9611, 2024. 1

[54] Zhongyang Zhang, Shuyang Cui, Kaidong Chai, Haowen Yu, Subhasis Dasgupta, Upal Mahbub, and Tauhidur Rahman. V2CE: Video to continuous events simulator. In *ICRA*, pages 12455–12461, 2024. 2, 3, 6, 7, 8

[55] Zangwei Zheng, Xiangyu Peng, Tianji Yang, Chenhui Shen, Shenggui Li, Hongxin Liu, Yukun Zhou, Tianyi Li, and Yang You. Open-Sora: Democratizing efficient video production for all. *arXiv preprint arXiv:2412.20404*, 2024. 5, 6

[56] Xinyu Zhou, Peiqi Duan, Yi Ma, and Boxin Shi. EvUnroll: Neuromorphic events based rolling shutter image correction. In *Proc. CVPR*, pages 17775–17784, 2022. 1

[57] Alex Zihao Zhu, Ziyun Wang, Kaung Khant, and Kostas Daniilidis. EventGAN: Leveraging large scale image datasets for event cameras. In *Proc. ICCP*, pages 1–11, 2021. 2

Texvent: Asynchronous Event Data Simulation via Text Prompt

Supplementary Material

8. Event camera circuit

Fig. 8 shows a simplified schematic of the event camera circuit. As shown in sub-figure (a), the voltage V_p is generated when the light hits the photoreceptor, logarithmically increasing with the light intensity. Then, an inverting amplifier (-A) is employed to amplify the change in log intensity from the value memorized after the last event was activated. Finally, two voltage comparators detect the increase or decrease in log intensity that exceeds the threshold ($\delta = \{\delta_{on}, \delta_{off}\}$). The principle of operation is illustrated in (c). Once the voltage change reaches the ON or OFF threshold, the event camera activates an ON or OFF event and resets the capacitor C.

The digital circuit of reset and refractory period is depicted in Fig. 8 (b). A reset pulse is generated when the pixel receives row and column acknowledge signals: RA and CA. The charge on the capacitor C is released quickly. C' in the circuit is the capacitive element of the reset circuit, which forms a time length with the bias current I_{refr}, thereby controlling the refractory period. By adjusting the I_{refr}, the charging rate of C' can be changed, thereby changing the refractory period, which is used to balance continuous event triggering. Based on this mechanism, we propose the concept of brightness cache in event simulation, effectively improving the simulation fidelity.

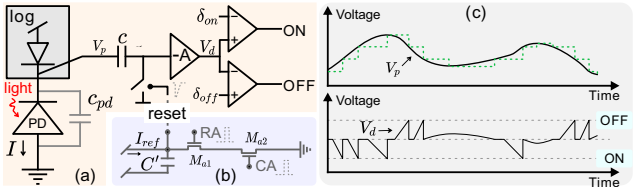

Figure 8. (a) Circuit of the event camera. (b) Reset and refractory period. (c) The principle of operation. Figure adapted from [8, 11].

9. Additional experiments

9.1. Ablation study

In Table 4, we conduct the quantitative evaluation to measure the effect of the balancing parameter, brightness cache, and frame interpolation. For the balancing parameter, the default α is set to 30 in Texvent. In Table 4, we set it to 0 (E1), 0.5× (E2), 2× (E3) respectively to test the EQS [2] of the corresponding generated event data. Setting α with 2× leads to lower EQS since the larger the balancing parameter, the smoother the brightness change scale, as shown in Fig. 9. In the sub-figure E3, the brightness change shows

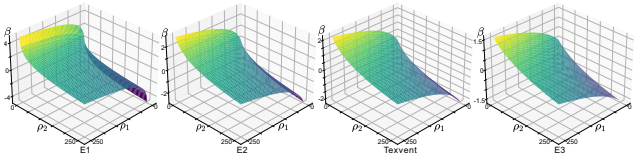

Figure 9. Visualization of equation $\beta = log(\alpha+\rho_1) - log(\alpha+\rho_2)$ with different balancing parameters (α). ρ_1 and ρ_2 denote the pixel value, subject to ($0 \sim 255$). This equation is derived from Eq. (1) in the main paper. E1: $\alpha = 0$. E2: $\alpha = 0.5\times$. E3: $\alpha = 2\times$. Texvent: $\alpha = 30$.

similar scales in low light and high light conditions, which conflicts with the basic mechanism of event cameras.

For the brightness cache, we conduct tests by disabling the cache and implementing a global cache in experiments E4 and E5 of Table 4, respectively. The adoption of a global cache results in the simulation of false events, which consequently diminishes the EQS. The global cache means that a brightness cache works for all video frames. Compared with Texvent, removing the brightness cache leads to potential event missing, reducing the EQS from 0.9086 to 0.8597.

For frame interpolation, we respectively remove interpolation (E6), interpolate it with a fixed number (10) (E7), and adopt the interpolator used in existing event simulators (Super Slomo [9]) (E8) to evaluate the importance of our interpolation strategy. Using the Super Slomo would interpolate redundant frames, which not only reduces the efficiency but also corrupts the event simulation fidelity. Thus, it's greatly important to propose an adaptive interpolation strategy based on brightness change for simulating event data.

To evaluate the effectiveness of our noise addition, we compare our Texvent with three settings: without adding noise (E9), adding Gaussian noise (E10), and randomly adding noise into the whole region of the simulated event data (E11). As shown in Table 5, the EQS of the simulated data is reduced by a large value when removing our noise addition strategy. This experiment effectively demonstrates the importance of the proposed noise addition strategy in our methodology section.

To demonstrate the importance of our time stamp reconstruction, we employ a random initialization method to assign the time stamp, denoted as E12 in Table 5. The EQS of the simulated event data is decreased from 0.9086 to 0.9065 when altering the time stamp initialization in a random way. This situation effectively demonstrates the effectiveness of our time stamp reconstruction strategy.

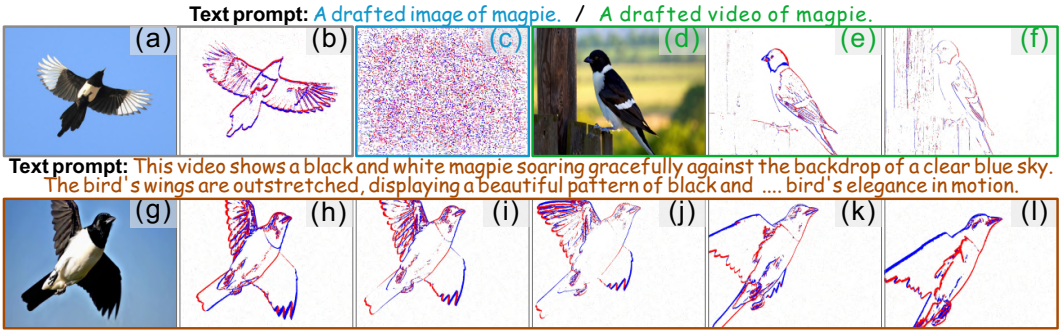

Figure 10. (a)-(b): Ground-truth image-event pair. (c): Event frame of Event decoder [14] + EventBind [21] using blue prompt. (d)-(f): Event frame of Texvent using green text prompt. (g)-(l): Event frame of Texvent using orange text prompt. The orange text prompt is generated by LLaVA-v1.5-13B [13].

Table 4. Ablation studies about the balancing parameter α (E1-E3), brightness cache κ (E4-E5), and frame interpolation strategy K (E6-E8). Local cache (L. C.) denotes a cache only used for an image pair (before interpolation), while global cache (G. C.) is used for all video sequences. B. A. denotes our proposed brightness-aware interpolation strategy. "Fixed" denotes interpolating a fixed number (10) of intermediate frames into each image pair. S. L. indicates the Super Slomo [9], which is used in VID2E [3]. "—" denotes the result tested by removing the corresponding operations. The best and second-best scores are highlighted in **bold** and underlined.

	E1	E2	E3	E4	E5	E6	E7	E8	Texvent
α	0	15	60	30	30	30	30	30	30
κ	L. C.	L. C.	L. C.	~~L. C.~~	G. C.	L. C.	L. C.	L. C.	L. C.
K	B. A.	B. A.	B. A.	B. A.	B. A.	~~B. A.~~	Fixed	S. L.	B. A.
EQS↑	<u>0.8639</u>	0.8637	0.8531	0.8597	0.8508	0.8435	0.8638	0.8314	**0.9086**

Table 5. Ablation studies about the proposed noise addition (E9-E11) and time time stamp reconstruction (E12). For noise addition, we implement E9-E11 by 'without noise', 'Gaussian noise', and 'random noise', respectively. E12 denotes initializing time stamps of simulated event data randomly.

	E9	E10	E11	E12	Texvent
EQS↑	0.8400	0.8142	0.8402	0.9065	0.9086

9.2. Baseline comparison

To the best of our knowledge, only a limited number of text-to-event baselines have been proposed [14] (without open-sourcing). We reproduce [14] via combining an event decoder [14] with EventBind [21] for T2E. Specifically, we pre-train an event encoder-decoder net [14] on the N-ImageNet dataset [10], then use its decoder to generate event data from the text embeddings extracted by EventBind [21]. However, as shown in Fig. 10 (c), this baseline fails to generate reliable event data. This is mainly because the limited text-event pairs lead to a sub-optimal text embedding module, thereby affecting the following event decoding. In contrast, our Texvent discards the text-event pairs while also generating high-quality event data, as shown in the case (e-f) of Fig. 10. Even with a very simple text prompt (green), Texvent still produces simulated event data that closely matches the real data (b). Providing a more detailed prompt (orange) will add richer motion details. These results show that Texvent's performance does not rely heavily on the prompt quality or complexity.

9.3. Effectiveness on object recognition

We evaluate the usefulness of our Texvent on the object recognition task. We sample the first 10 classes from the N-ImageNet dataset and augment 20% new samples to train four different models. As shown in Table 6, our simulated data can improve the accuracy of different classifiers effectively. Compared with VID2E [3], our Texvent achieves better performance, particularly improving the accuracy of ResNet34 from 0.514 to 0.6320. For fair comparison, we train these models with 30 epochs. This evaluation demonstrates the usefulness of our method on mainstream tasks.

9.4. Noise influence in image reconstruction

In our main experiment, we employ an image reconstruction method to evaluate the fidelity of the simulated event data. To evaluate the noise sensitivity of the employed image reconstruction method (*i.e.*, E2VID), we randomly perturb ρ percent of events to test the MSE, SSIM, and LPIPS. As shown in Table 7, the larger the perturbation rate, the higher

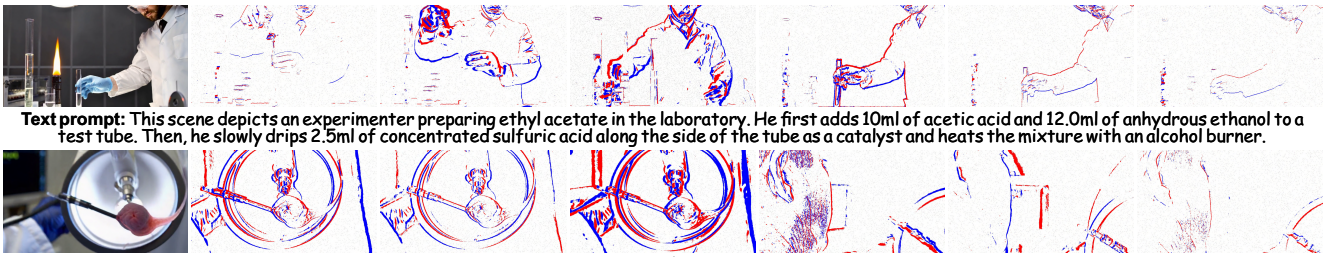

Figure 11. Failure cases of Texvent. Cosmos-1.0-Diffusion-7B-Text2World is employed as the video generator. Texvent struggles to simulate accurate event data in certain rare scenarios due to the limitations of current large video generators in specialized applications. This challenge can be effectively overcome by utilizing task-specific large video generators.

Table 6. Effectiveness evaluation of our Texvent in the objection recognition task. The video-to-event simulation method, VID2E, is employed as a comparison baseline.

	ResNet34	ResNet152	ViTB	SqueezeNet
Original	0.4980	0.4680	0.3840	0.3860
with VID2E	0.5140	0.4820	0.3820	0.3920
with Texvent	0.6320	0.5060	0.3940	0.4000

Table 7. Noise sensitivity evaluation of E2VID.

	$\rho = 1\%$	$\rho = 5\%$	$\rho = 10\%$	$\rho = 20\%$	$\rho = 30\%$
MSE↓	0.0035	0.0057	0.0080	0.0083	0.0168
SSIM↑	0.8958	0.7657	0.6811	0.5397	0.3652
LPIPS↓	0.0598	0.1512	0.2156	0.3426	0.5353

Table 8. Fidelity evaluation of simulated event data with a different image reconstruction method, ETNet [17].

	VID2E	V2E	V2CE	DVS.	SENPI	Texvent
MSE↓	0.2745	0.1354	0.2614	0.0836	0.0675	0.0799
SSIM↑	0.5521	0.6187	0.4545	0.6602	0.1976	0.6893
LPIPS↓	0.3555	0.3154	0.4731	0.2172	0.4989	0.2074

the MSE loss and the lower the SSIM and LPIPS. Thus, the E2VID is sensitive to the random noise, showing that different reconstruction methods will achieve different evaluation results. Apart from E2VID, we have employed ETNet [17] to reconstruct images from simulated event data, and the results are shown in Table 8. Our method still shows higher fidelity than other baselines.

9.5. Downstream evaluation

To evaluate the fidelity of the simulated event data, we directly test each simulator with a pretrained event-based classification model [10] on the N-ImageNet dataset. We employ the binary event image as the representation method and then test the top-1 and top-5 accuracy, respectively. As shown in Table 9, the pretrained classifier (Original) achieves the Acc@1 and Acc@5 by 0.6920 and 0.9156, respectively. The classifier achieves reduced performance on the simulated event data, denoting the domain gap between the realistic and simulated data. Texvent achieves the Acc@1 by 0.3067 and the highest Acc@5 by 0.5458. VID2E [3] achieves the highest Acc@1 by 0.3587, higher than ours by 4.18%. Other simulators all achieve the Acc@1 under 20%. This experiment demonstrates that exploration about ensuring the fidelity of simulated event data is needed. We will continuously focus on narrowing this gap and evaluating the promotion achieved by Texvent through real data evaluation.

9.6. Failure cases

Texvent employs the general multimodal LLM to simulate event data, which may fail in some specific scenarios, such as shown in Fig. 11. When we ask Texvent to simulate event data for a chemistry experiment, this highly professional experimental step and strict operation requirements limit the video generator from generating accurate video frames. Therefore, the simulated event data shows low fidelity according to the provided text prompt. For example, we show that there is an alcohol burner positioned beneath the test tube; however, the rendered frames depict the burner next to the tube, failing to heat the mixture. Also, the detailed liquor volume and strict experimental operation are not highlighted. In the second row, the text prompt focuses on the colonoscopy, while the simulated event data only include a polyp without modeling the real scene of the intestinal wall. These cases occur when users simulate event data for some rare application-specific scenarios. This is an out-of-distribution challenge that is present in current video generation models. It's infeasible to train a single world model that can accurately generate anything. To mitigate this problem, we can collect some task-specific data to fine-tune the Texvent or introduce a detection strategy to filter out failure cases, thereby preventing downstream training utility. In Texvent, the video generator is open-sourced, which can be easily updated and altered to enable

Figure 12. Estimated optical flow maps of the ground truth and various simulated event data. The color wheel is shown in the bottom-left corner of the ground truth. Texvent demonstrates better flow consistency with the ground truth compared to other baselines. The spatial misalignment between the ground truth map and the simulated event data arises from the misalignment between the event sensor and the RGB camera in the DSEC dataset [5].

Table 9. Classification accuracy of different simulators. We employ the official models released in the N-ImageNet project [10] with the representation of "Binary Event Image". Acc@1 and Acc@5 denote the top-1 and top-5 accuracy, respectively. The best and second-best scores are highlighted in **bold** and underlined.

	Original	VID2E [3]	V2E [8]	V2CE [19]	DVS-Voltmeter [19]	SENPI [6]	Texvent
Acc@1	0.6920	**0.3585**	0.3251	0.1611	0.1817	0.1686	0.3067
Acc@5	0.9156	0.4760	0.4873	0.3500	0.4589	0.4175	**0.5458**

task-specific applications.

9.7. Details of the NT-ImageNet dataset

Table 10 provides detailed statistics of our NT-ImageNet's characteristics. The NT-ImageNet dataset has 5000 text-event pairs distributed in 100 classes with different event densities and diverse motion types. Detailed collection steps are shown in the experimental details of the main paper to enhance the reproducibility.

10. Various visualizations

In this section, we show extensive results about our method under different multimodal large language models, including Cosmos [1], Wan [16], Open-Sora [20], and CogVideoX [18]. In Fig. 13, we evaluate our simulator among five kinds of generators in the autonomous driving scenario. From the first row to the last row, we display the first video frame, simulated event frames, and the last video frame, respectively. 7B and 14 B denote the parameter size of the employed large models. Our Textvent can accurately simulate event data for moving cars, which may promote the exploration of event-based autonomous driving. In addition, we test Texvent in a more challenging scenario: a construction site, as shown in Fig. 14. Although the scene is greatly complex, Texvent can still simulate event data from these frames to represent motion information. Extensive experiments demonstrate that our Texvent can simulate various event streams based on the text prompts.

11. Broader impact

The proposed event simulation method, Texvent, has the potential to create significant broader impacts in various domains. For instance, Texvent can accelerate the research and development in event-based vision. Event cameras are relatively expensive, and real-world event data collection can be challenging due to hardware limitations and the need for specific conditions (*e.g.*, high-speed motion or dynamic lighting). Our Texvent can provide researchers with a cost-effective and scalable way to generate synthetic event data, facilitating the training and development of event-based algorithms. Apart from positive impacts, over-reliance on the proposed method may have some negative impacts. The simulated event data may be unrealistic if the generated videos do not accurately mimic real-world dynamics. This could lead to overfitting of models to synthetic data, reducing their performance in real-world scenarios. Indeed, Texvent has the potential to promote event-based vision research and applications by making data more accessible and scalable. However, careful consideration of its limitations is essential to ensure responsible and effective usage.

12. Limitations and future work

Texvent is the first training-free text-to-event simulation framework that can synthesize various event data via simple text prompts. We mainly focus on studying the fidelity of simulated data while causing some limitations in overall efficiency, generalization ability, and evaluation metrics. We will explore the following directions in the future:

- Our Texvent employs a multimodal large language model to narrow the gap between the text description and event data, thereby the overall efficiency of the pipeline is limited by the selected model. To address this, we propose implementing event data simulation directly from the latent tokenizers generated by large models. This approach could significantly reduce the computational overhead associated with rendering video frames, allowing for more

Table 10. Details of the collected NT-ImageNet dataset.

Characteristic	Range / Coverage in NT-ImageNet
Class Number	100 classes, covering a broad spectrum of object categories.
Number/Class	50 text-event pairs per class (total: 5000 pairs).
Event Density	Sparse to dense; average events per stream: [3k, 170k].
Motion Types	Dynamic scenes, single/multiple object motion, complex, and naturalistic movement.
Illumination	Wide range: daylight, low-light, indoor/outdoor, underwater, varying shadows, highlights, etc.
Text length	Automatically generated text captions: (min: 39, max: 124).

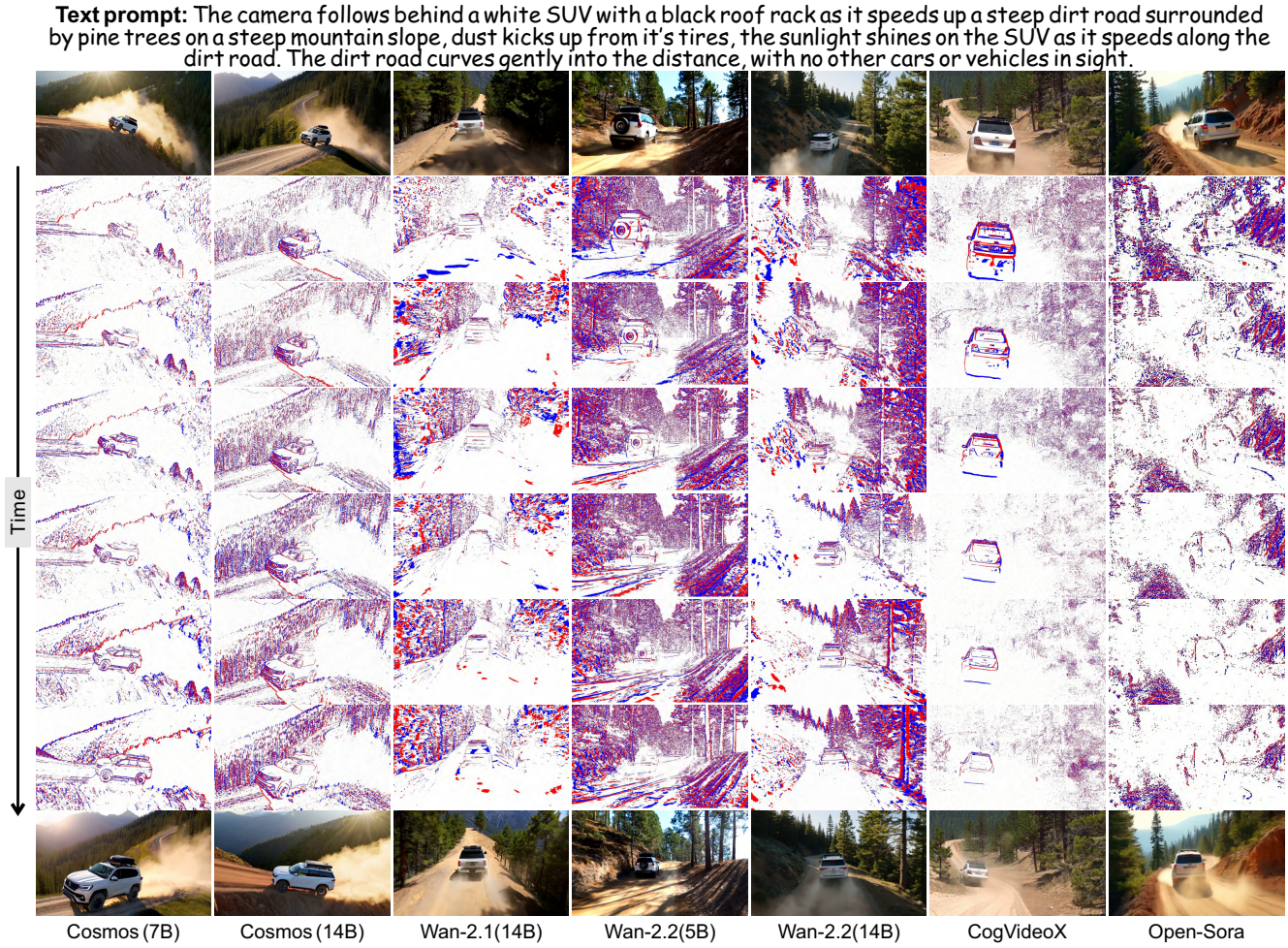

Figure 13. Comparison across different video generators, including Cosmos [1], Wan [16], CogVideoX [18], and Open-Sora [20]. From top to down, we show the text prompt, the first video frame, simulated event frames, and the last video frame, respectively.

efficient conversion of event data. Additionally, by focusing on high-level representations of text prompts, we can streamline prompt handling and improve the model's responsiveness to frequent requests. Ultimately, this improvement could lead to a faster simulation process and enhanced user experience across various applications.

- Current event simulators employ various strategies [12, 15, 19] to reconstruct the dense time stamps. However, the estimated time stamps still diverge from the original values, even when evaluated using the brightness variation rate (our solution), as shown in Fig. 12. The optical flow indicates both the direction and magnitude of motion between two consecutive events, which is closely linked to the fidelity of the reconstructed time stamps. The chaotic optical flows show that more effective time stamp reconstruction methods should be proposed. This

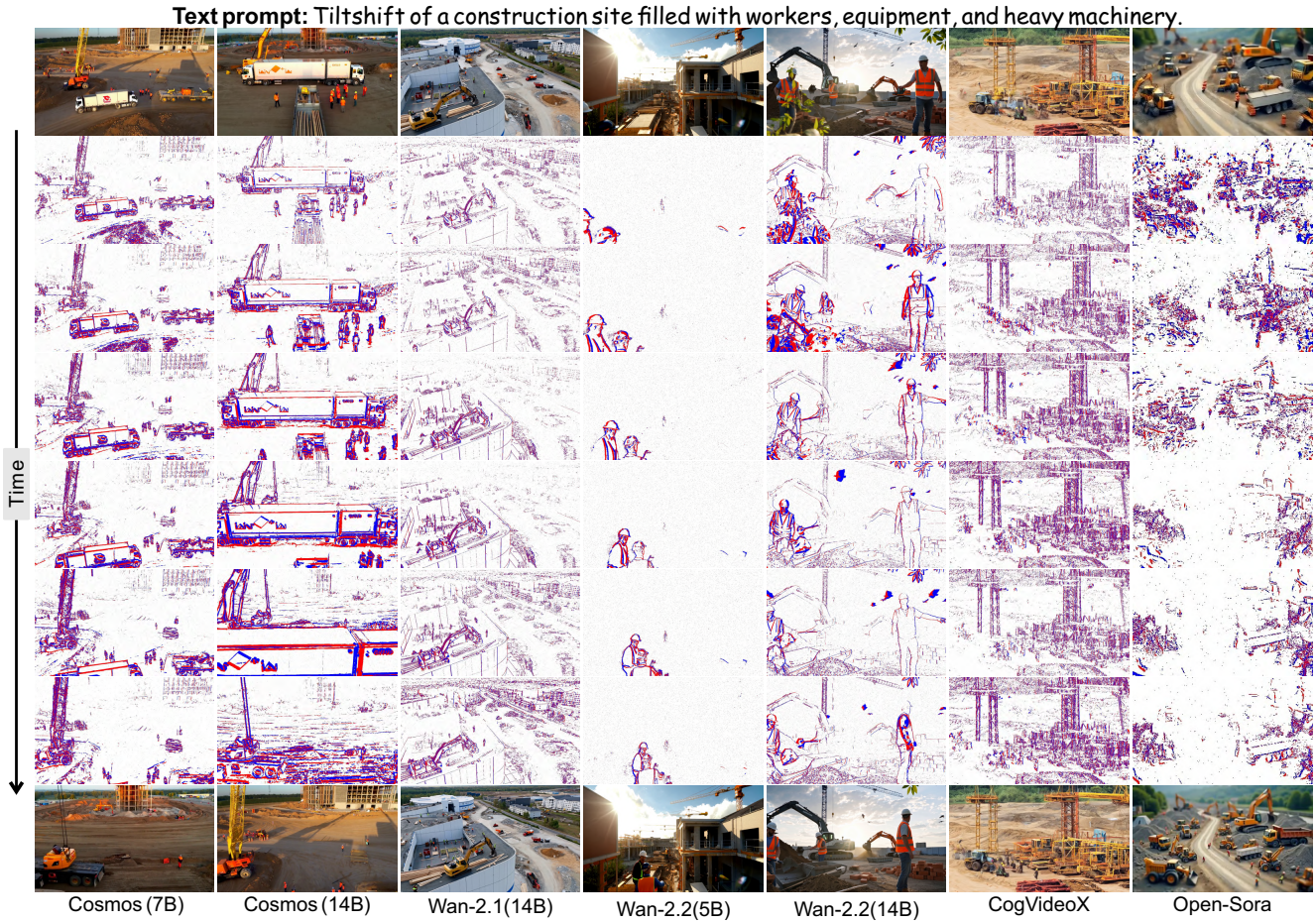

Figure 14. Comparison across different video generators, including Cosmos [1], Wan [16], CogVideoX [18], and Open-Sora [20]. From top to bottom, we show the text prompt, the first video frame, simulated event frames, and the last video frame, respectively.

advancement not only improves the fidelity of the simulated event data but also broadens current simulation methods in event regression tasks.

- In our experimental section, we have conducted extensive experiments on image reconstruction, object recognition, depth and optical flow estimation, *etc*. The limited augmented data may not adequately demonstrate the generalizability of our method in visual odometry, high-speed counting (*e.g*., cytometry), and other application-specific scenarios. In the future, we will propose a comprehensive plan that includes diversifying tasks to encompass scene understanding, conducting cross-domain evaluations in areas like autonomous driving, and benchmarking our method in specific scenes where it is difficult to record actual data. We believe that the expanded evaluation will help assess the versatility of our method across various applications.
- To directly evaluate the quality of the simulated event data, the EQS [2], a newly proposed event metric, is employed in our experiments. However, it has two limitations: 1) High model dependence. EQS employs an object detection model [4] to extract event features for calculating the cosine similarity. This means the EQS is dependent on the selected event-based model, leading to different scores when employing different models. 2) High computational cost. Compared with image metrics, EQS shows a higher computational cost since it extracts deep features. Therefore, a novel metric designed for the event simulation task is needed. We aim to propose a spatial-temporal pooling operation (like SPP [7]) that can map irregular event data into a regular representation for similarity calculation, showing less sensitivity to specific model architectures and reducing computational cost.

References

[1] Niket Agarwal, Arslan Ali, Maciej Bala, Yogesh Balaji, Erik Barker, Tiffany Cai, Prithvijit Chattopadhyay, Yongxin Chen, Yin Cui, Yifan Ding, et al. Cosmos world foundation model platform for physical AI. *arXiv preprint arXiv:2501.03575*, 2025. 4, 5, 6

[2] Kaustav Chanda, Aayush Verma, Arpitsinh Vaghela, Yezhou Yang, and Bharatesh Chakravarthi. Event quality score (eqs): Assessing the realism of simulated event camera streams via distance in latent space. In *Proc. CVPRW*, pages 5105–5113, 2025. 1, 6

[3] Daniel Gehrig, Mathias Gehrig, Javier Hidalgo-Carrió, and Davide Scaramuzza. Video to events: Recycling video datasets for event cameras. In *Proc. CVPR*, pages 3586–3595, 2020. 2, 3, 4

[4] Mathias Gehrig and Davide Scaramuzza. Recurrent vision transformers for object detection with event cameras. In *Proc. CVPR*, pages 13884–13893, 2023. 6

[5] Mathias Gehrig, Willem Aarents, Daniel Gehrig, and Davide Scaramuzza. DSEC: A stereo event camera dataset for driving scenarios. *RA-L*, 6(3):4947–4954, 2021. 4

[6] Joseph L Greene, Adrish Kar, Ignacio Galindo, Elijah Quiles, Elliott Chen, and Matthew Anderson. A pytorch-enabled tool for synthetic event camera data generation and algorithm development. In *Synthetic Data for Artificial Intelligence and Machine Learning: Tools, Techniques, and Applications III*, pages 117–137. SPIE, 2025. 4

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. *TPAMI*, 37(9):1904–1916, 2015. 6

[8] Yuhuang Hu, Shih-Chii Liu, and Tobi Delbruck. V2E: From video frames to realistic dvs events. In *Proc. CVPR*, pages 1312–1321, 2021. 1, 4

[9] Huaizu Jiang, Deqing Sun, Varun Jampani, Ming-Hsuan Yang, Erik Learned-Miller, and Jan Kautz. Super slomo: High quality estimation of multiple intermediate frames for video interpolation. In *Proc. CVPR*, pages 9000–9008, 2018. 1, 2

[10] Junho Kim, Jaehyeok Bae, Gangin Park, Dongsu Zhang, and Young Min Kim. N-imagenet: Towards robust, fine-grained object recognition with event cameras. In *Proc. ICCV*, pages 2146–2156, 2021. 2, 3, 4

[11] Patrick Lichtsteiner, Christoph Posch, and Tobi Delbruck. A 128×128 120db 15μs latency asynchronous temporal contrast vision sensor. *JSSC*, 43(2):566–576, 2008. 1

[12] Songnan Lin, Ye Ma, Zhenhua Guo, and Bihan Wen. Dvs-voltmeter: Stochastic process-based event simulator for dynamic vision sensors. In *Proc. ECCV*, pages 578–593, 2022. 5

[13] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Proc. NeurIPS*, 2023. 2

[14] Joachim Ott, Zuowen Wang, and Shih-Chii Liu. Text-to-Events: Synthetic event camera streams from conditional text input. In *NICE*, pages 1–10, 2024. 2

[15] Henri Rebecq, Daniel Gehrig, and Davide Scaramuzza. ESIM: an open event camera simulator. In *CoRL*, pages 969–982, 2018. 5

[16] Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025. 4, 5, 6

[17] Wenming Weng, Yueyi Zhang, and Zhiwei Xiong. Event-based video reconstruction using transformer. In *Proc. ICCV*, pages 2563–2572, 2021. 3

[18] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, Da Yin, Yuxuan.Zhang, Weihan Wang, Yean Cheng, Bin Xu, Xiaotao Gu, Yuxiao Dong, and Jie Tang. CogVideoX: Text-to-video diffusion models with an expert transformer. In *Proc. ICLR*, 2025. 4, 5, 6

[19] Zhongyang Zhang, Shuyang Cui, Kaidong Chai, Haowen Yu, Subhasis Dasgupta, Upal Mahbub, and Tauhidur Rahman. V2CE: Video to continuous events simulator. In *ICRA*, pages 12455–12461, 2024. 4, 5

[20] Zangwei Zheng, Xiangyu Peng, Tianji Yang, Chenhui Shen, Shenggui Li, Hongxin Liu, Yukun Zhou, Tianyi Li, and Yang You. Open-Sora: Democratizing efficient video production for all. *arXiv preprint arXiv:2412.20404*, 2024. 4, 5, 6

[21] Jiazhou Zhou, Xu Zheng, Yuanhuiyi Lyu, and Lin Wang. Eventbind: Learning a unified representation to bind them all for event-based open-world understanding. In *Proc. ECCV*, pages 477–494, 2024. 2