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.

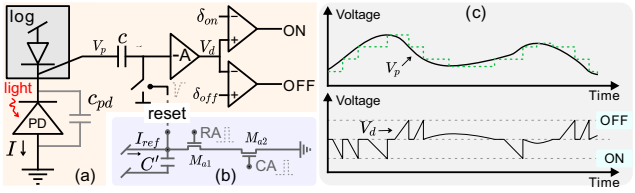

Figure 8. (a) Circuit of the event camera. (b) Reset and refractory period. (c) The principle of operation. Figure adapted from [8, 11].

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

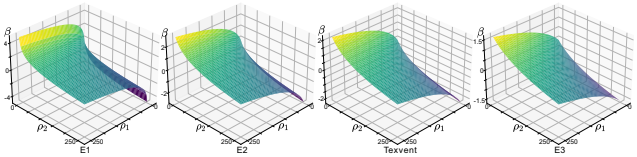

Figure 9. Visualization of equation $\beta = log(\alpha+\rho_1) - log(\alpha+\rho_2)$ with different balancing parameters (α). ρ_1 and ρ_2 denote the pixel value, subject to ($0 \sim 255$). This equation is derived from Eq. (1) in the main paper. E1: $\alpha = 0$. E2: $\alpha = 0.5\times$. E3: $\alpha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

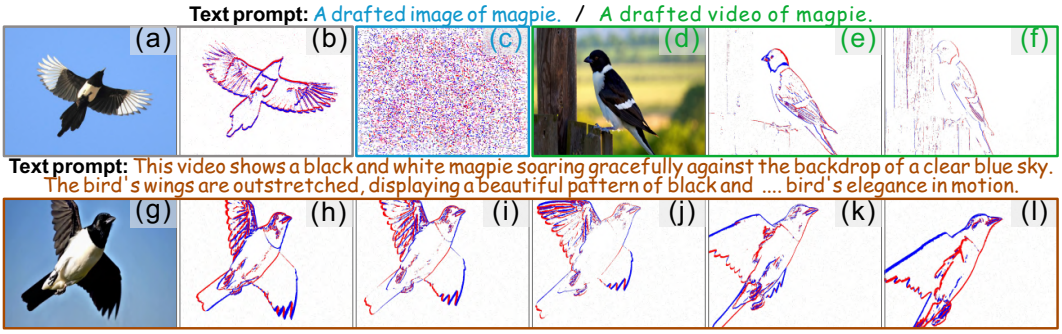

Figure 10. (a)-(b): Ground-truth image-event pair. (c): Event frame of Event decoder [14] + EventBind [21] using blue prompt. (d)-(f): Event frame of Texvent using green text prompt. (g)-(l): Event frame of Texvent using orange text prompt. The orange text prompt is generated by LLaVA-v1.5-13B [13].

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

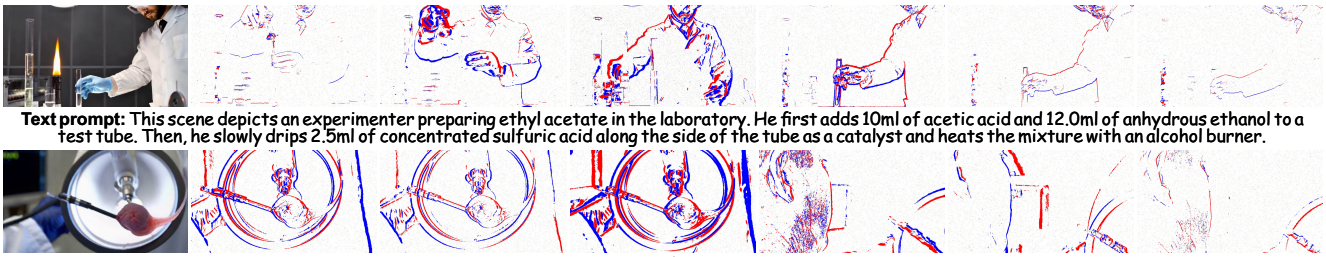

Figure 11. Failure cases of Texvent. Cosmos-1.0-Diffusion-7B-Text2World is employed as the video generator. Texvent struggles to simulate accurate event data in certain rare scenarios due to the limitations of current large video generators in specialized applications. This challenge can be effectively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

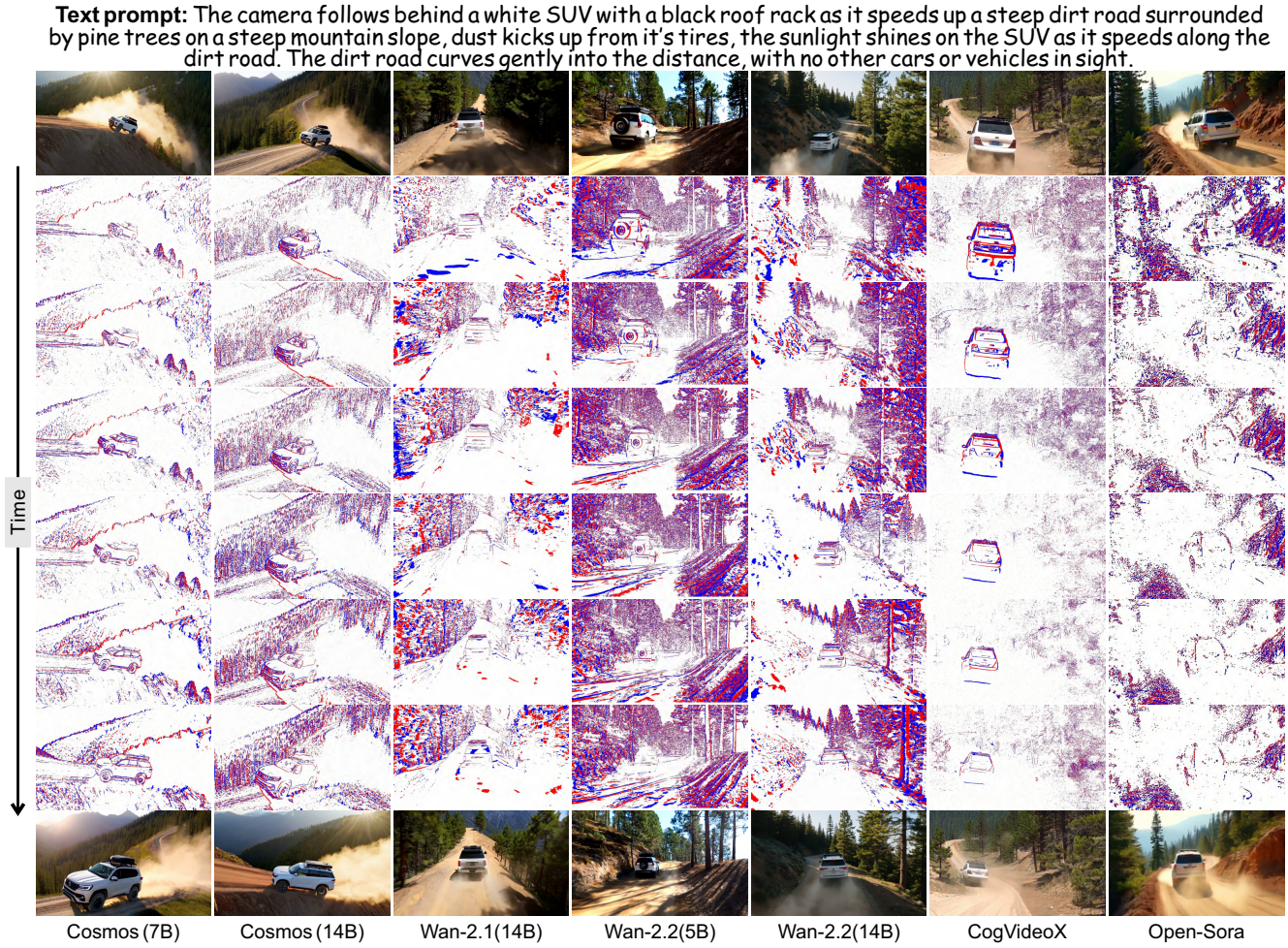

Figure 13. Comparison across different video generators, including Cosmos [1], Wan [16], CogVideoX [18], and Open-Sora [20]. From top to down, we show the text prompt, the first video frame, simulated event frames, and the last video frame, respectively.

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

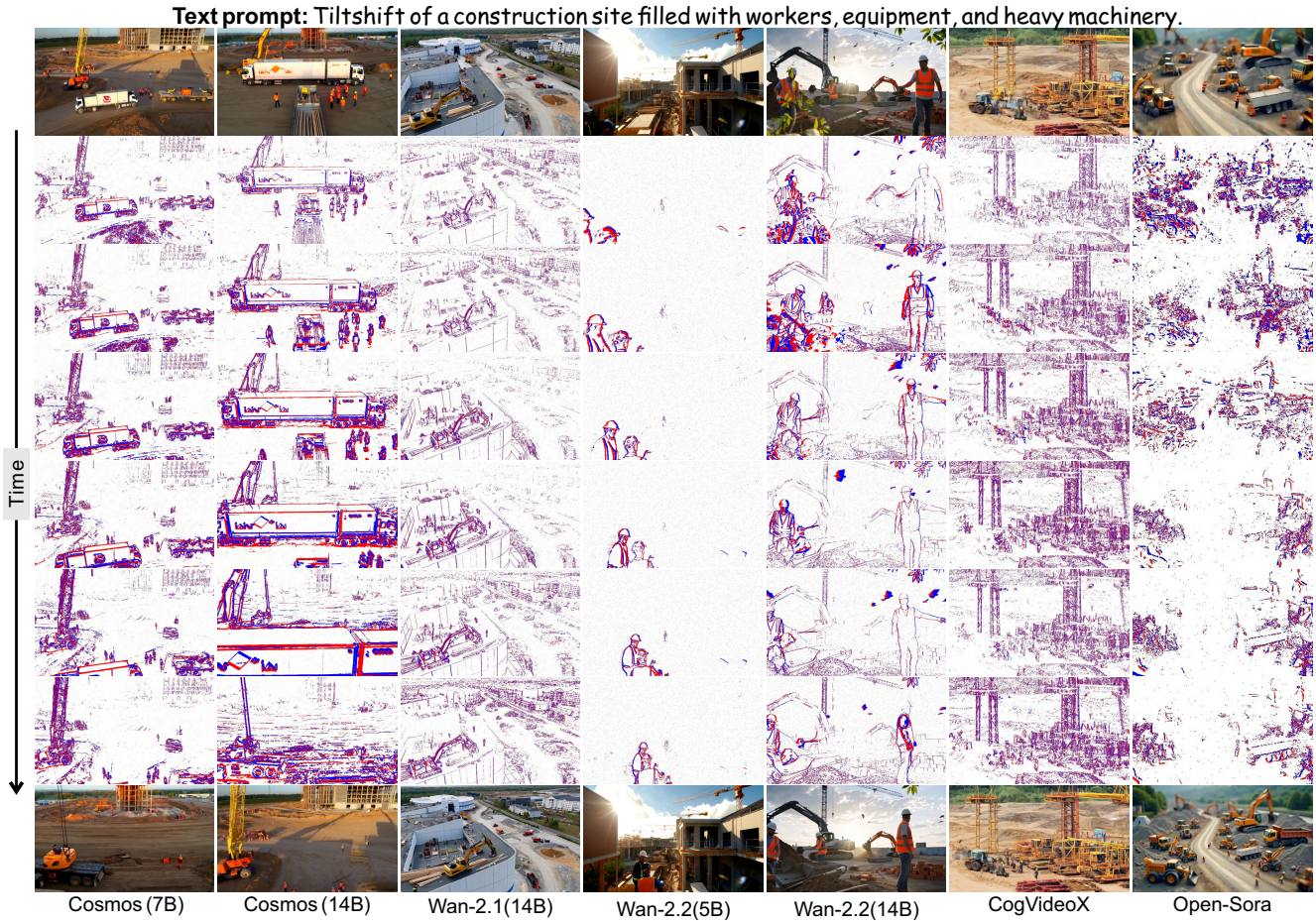

Figure 14. Comparison across different video generators, including Cosmos [1], Wan [16], CogVideoX [18], and Open-Sora [20]. From top to bottom, we show the text prompt, the first video frame, simulated event frames, and the last video frame, respectively.

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