CONTROLLABLE UNSUPERVISED EVENT-BASED VIDEO GENERATION

Yaping Zhao[1,2], *Pei Zhang*[1,2], *Chutian Wang*[1], *Edmund Y. Lam*[1,2,*]

[1] Department of Electrical and Electronic Engineering, The University of Hong Kong, Hong Kong SAR
[2]ACCESS — AI Chip Center for Emerging Smart Systems, Hong Kong SAR

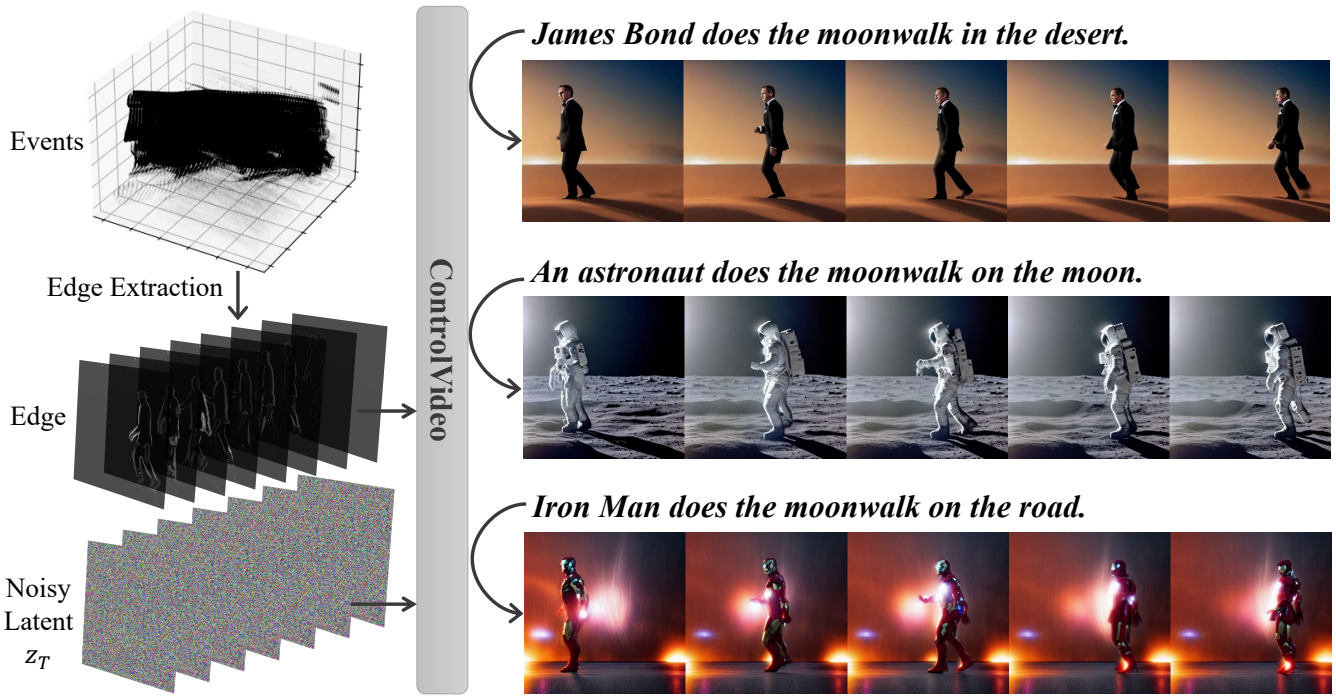

Fig. 1: We propose a video generation framework named **CUBE** (**C**ontrollable, **U**nsupervised, **B**ased on **E**vents). Left: CUBE is designed to generate videos conditioned on the edge information extracted from events using diffusion models. Right: CUBE could synthesize various photo-realistic videos given different textual descriptions.

ABSTRACT

The advent of event cameras, with their unique asynchronous sensing capabilities to capture the edge details of moving objects, has sparked new directions in video generation. So far, the challenge of integrating event-based data for controllable video generation remains largely unexplored. Addressing this gap, we introduce a framework that leverages the edge information from events and combines it with textual descriptions to synthesize videos without the requirement of extensive training. The framework marks a pioneering venture into event-based video generation using diffusion models. Comprehensive evaluations demonstrate the superior performance of our framework compared to existing methods. Code is available at: https://github.com/IndigoPurple/CUBE.

*Corresponding author.

This work is supported in part by the Research Grants Council (GRF 17201822), and by ACCESS — AI Chip Center for Emerging Smart Systems, Hong Kong SAR, China.

***Index Terms*—** neuromorphic imaging, computational imaging, event camera, video generation, diffusion model

1. INTRODUCTION

Event cameras, a groundbreaking type of asynchronous sensor, function distinctly from conventional cameras that capture images at consistent intervals. Drawing inspiration from biological systems, these cameras autonomously capture incremental brightness changes at each pixel, known as "events" [1]. The forefront of computational neuromorphic imaging (CNI) is focused on integrating the physical imaging process with the event-driven modality to enhance efficiency [2, 3, 4, 5, 6]. The capability of CNI to selectively capture edge information of moving objects while reducing bandwidth by discarding unnecessary visual data is noteworthy. CNI with event cameras are characterized by several

advantages including high dynamic range (HDR), superior temporal resolution, and low energy consumption. These attributes render CNI highly effective for specific applications in HDR environments and high-speed motion capture scenarios [7, 8, 9, 10].

However, the inherent sparsity and asynchronous nature of event streams present a challenge in recording absolute scene intensity, thus limiting their capacity for intuitive and natural visualization of detailed scene information. Consequently, events fall short in terms of perceptual realism. Fortunately, the event stream encapsulates a condensed form of visual data, furnishing essential elements for image or video reconstruction [11, 12, 13, 14]. A common practice involves reconstructing images from the event stream. Unfortunately, existing methods either exhibit limited performance [15, 16, 17, 18, 19, 20] or require extensive ground truth frames for neural network training [21, 22, 23, 24]. Recent studies [24] have delved into the application of diffusion models [25, 26, 27, 28, 29] for image generation. Despite these advancements, the reconstruction quality substantially lags behind the standards of photo-realistic videos, particularly in synthesizing individual frames independently, and suffers in training requirements. Additionally, the outcomes generated by previous methods lack controllability and cannot be guided by high-level semantic information provided by users to create specific scene content.

To address these issues, we introduce a video generation framework named **CUBE** (**C**ontrollable, **U**nsupervised, **B**ased on **E**vents). Our approach is designed to generate videos conditioned on both the edge information derived from events and a given textual description. A key insight of our work is that while events capture motion information, we can artificially endow these moving objects with specific appearances, textures, and scene backgrounds. As illustrated in Fig. 1, instead of training from scratch, our approach efficiently utilizes the generative capabilities of pre-trained text-to-image models [30, 31], coupled with the temporal consistency inherent in event streams, to produce vivid videos.

Our main contributions are as follows:

- To the best of our knowledge, this is the **first** work for event-based video reconstruction with diffusion model.
- We introduce a **controllable, training-free** framework that combines an edge extraction module with an existing diffusion model. This combination facilitates the reconstruction of video from events, leveraging on the controllability of ControlVideo [32] while circumventing the extensive training requirements.
- Quantitative and qualitative evaluations demonstrate the **superior performance** of our framework in video quality, temporal consistency, and textual alignment compared to existing methods.

2. RELATED WORK

2.1. Event-based Video Reconstruction

CNI encodes logarithmic intensity changes into a low-redundancy event stream [33, 34, 35], enabling efficient image or video reconstruction [11, 12, 13]. However, existing methods either perform poorly [15, 16, 17, 18, 19, 20] or require auxiliary ground truth frames for neural network training [21, 22, 23, 24]. Recent research [24] uses the diffusion model for event-based image generation but reconstruction quality still falls short of photo-realistic videos and necessitates training. Additionally, previous approach [24] lacks control over generated results, complicating the guidance of video generation based on semantic information.

2.2. Diffusion Model

Denoising diffusion probability models (DDPMs) [25, 26, 27, 28, 29, 36, 37] have emerged as popular research models in computer vision, demonstrating impressive capabilities in image generation. The latent diffusion model (LDM) [26] is an efficient variant of diffusion models that applies the diffusion process in the latent space instead of the image space. LDM consists of two main components. First, it employs an encoder $\mathcal{E}$ to compress an image $\boldsymbol{x}$ into a latent code $\boldsymbol{z} = \mathcal{E}(\boldsymbol{x})$ and a decoder to reconstruct the image $\boldsymbol{x} \approx D(\boldsymbol{z})$. Second, it learns the distribution of image latent codes $\boldsymbol{z}_0 \sim p_{data}(\boldsymbol{z}_0)$ using a DDPM formulation [25], which includes a forward and a backward process. The forward process gradually adds Gaussian noise at each timestep t to obtain $\boldsymbol{z}_t$:

$$q(\boldsymbol{z}_t|\boldsymbol{z}_{t-1}) = \mathcal{N}(\boldsymbol{z}_t; \sqrt{1-\beta_t}\boldsymbol{z}_{t-1}, \beta_t I), \tag{1}$$

where ${\beta_t}^T t = 1$ is the scale of noises, and T denotes the number of diffusion timesteps. The backward denoising process reverses the diffusion process to predict less noisy $\boldsymbol{z}_{t-1}$:

$$p_\theta(\boldsymbol{z}_{t-1}|\boldsymbol{z}_t) = \mathcal{N}(\boldsymbol{z}_{t-1}; \mu_\theta(\boldsymbol{z}_t, t), \Sigma_\theta(\boldsymbol{z}_t, t)). \tag{2}$$

The μ_θ and Σ_θ are implemented using a denoising model ϵ_θ with learnable parameters θ, which is trained with a simple objective:

$$\mathcal{L}_{simple} := \mathbb{E}_{\mathcal{E}(\boldsymbol{z}),\epsilon\sim\mathcal{N}(0,1),t}\left[||\epsilon = \epsilon_\theta(\boldsymbol{z}, t)||_2^2\right]. \tag{3}$$

During the generation of new samples, we start from $\boldsymbol{z}_T \sim (0, 1)$ and employ DDPM sampling to predict $\boldsymbol{z}_{t-1}$ at the previous timestep:

$$\begin{aligned} \boldsymbol{z}_{t-1} &= \sqrt{\alpha_{t-1}}z' + \sqrt{1-\alpha_{t-1}} \cdot \epsilon_\theta(\boldsymbol{z}_t, t), \\ z' &= \frac{\boldsymbol{z}_t - \sqrt{1-\alpha_t}\epsilon_\theta(\boldsymbol{z}_t, t)}{\sqrt{\alpha_t}}, \end{aligned} \tag{4}$$

where $\alpha_t = \prod_{i=1}^{t}(1-\beta_i)$. We use $\boldsymbol{z}_{t\to 0}$ to represent the "predicted $\boldsymbol{z}_0$" at timestep t for simplicity. We employ Stable Diffusion (SD) $\epsilon_\theta(\boldsymbol{z}_t, t, \tau)$ as our base model, which is an instantiation of text-guided LDMs pre-trained on billions of image-text pairs. Here, τ represents the text prompt.

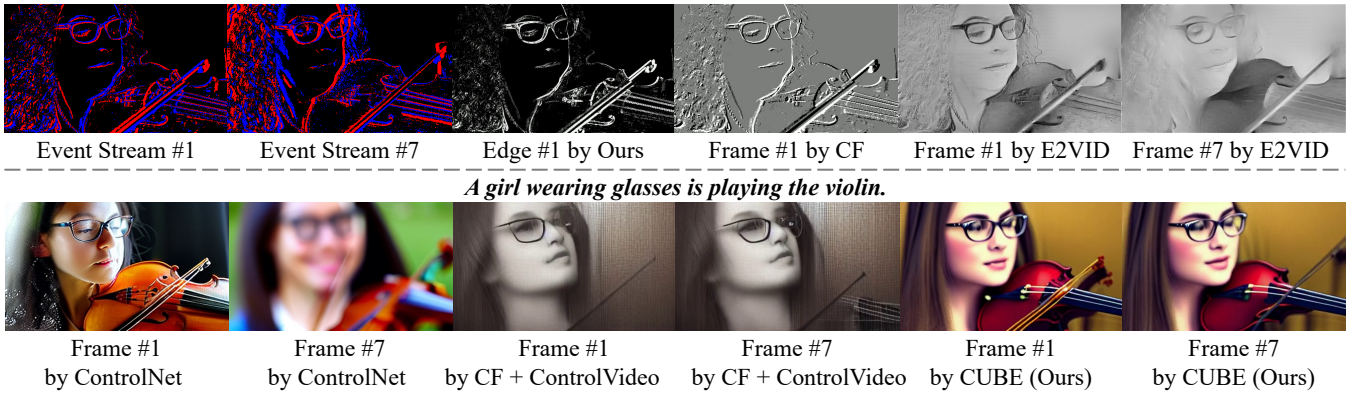

Fig. 2: Qualitative comparisons show CUBE outperforms others in video quality, temporal consistency, and textual alignment.

Method	Structure Condition	Frame Consistency (%)	Prompt Consistency (%)
ControlNet	Edge by Ours	84.52	21.47
ControlVideo	Edge by CF	90.03	23.62
CUBE (Ours)	Edge by Ours	**92.27**	**27.74**

Table 1: Quantitative comparisons of CUBE with other methods.

3. METHOD

ControlNet [38] and ControlVideo [32] have expanded the scope of text-to-image and text-to-video generation to include varied input conditions like depth maps, poses, scribbles, and edges. Despite these advancements, the incorporation of events as input conditions remains unexplored. Our framework integrates an edge extraction module with ControlVideo, enabling the reconstruction of videos from events.

3.1. Edge Extraction

The event stream obtained from an event camera can be denoted as $\varepsilon = \{e_i\}_{i=1}^{N}$, where N is the number of events. Here, each event $e_i \in \varepsilon$ is represented by a tuple (x_i, y_i, s_i, p_i), where x and y represent the spatial position, s represents the timestamp, and $p = \pm 1$ represents the polarity of the event.

To facilitate the integration of event stream ε with ControlVideo, we design an edge extraction module to convert events into edges. For synthesizing V video frames, ε is segmented into V bins $\varepsilon_{j\in[1,V]}$, each holding n events. Then, the edge map is extracted using the following equation:

$$\mathbf{I}_{j\in[1,V]}(x,y) = \sum_{i,e_i\in\varepsilon_j} \frac{|p_i|\delta(x-x_i)\delta(y-y_i)}{N}, \quad (5)$$

resulting in an intensity image $\mathbf{I} \in [0,1]^{H\times W\times 1}$, with H and W representing height and width, respectively. Here, $\delta(\cdot)$ is defined as the Kronecker delta function. This method ensures that each edge in the video is directly traceable to the specific events that occurred at that spatial location, capturing crucial details of motion and change in the scene. Additionally, the method's reliance on the number of events rather than continuous intensity values allows for a more robust edge detection, particularly effective in dynamic and challenging lighting conditions.

3.2. Video Generation

Our approach to controllable event-based video generation aims to produce a V-length video, leveraging both the extracted edge information $\mathbf{I}$ and a textual prompt τ. As depicted in Fig. 1, we introduce CUBE, a training-free framework adapted from ControlVideo[32], augmented with our edge extraction module for consistent and efficient video generation. In alignment with ControlVideo, we first estimate the clean video latent $\boldsymbol{z}_{t\to 0}$ from $\boldsymbol{z}_t$ using the formula:

$$\boldsymbol{z}_{t\to 0} = \frac{\boldsymbol{a}_t - \sqrt{1-\alpha_t}\epsilon_\theta(\boldsymbol{z}_t, t, \mathbf{I}, \tau)}{\sqrt{\alpha_t}}. \quad (6)$$

Following ControlVideo [32], after mapping $\boldsymbol{z}_{t\to 0}$ to an RGB video $\boldsymbol{x}_{t\to 0} = \mathcal{D}(\boldsymbol{z}_{t\to 0})$, we refine it to a smoother version $\tilde{\boldsymbol{x}}_{t\to 0}$ by employing the interleaved-frame technique from RIFE [39]. This technique helps in maintaining temporal consistency by interpolating intermediate frames that reduce visual discontinuities between successive frames, thus enhancing the fluidity of motion in the generated video. The smoother video latent $\tilde{\boldsymbol{z}}_{t\to 0} = \mathcal{E}(\tilde{\boldsymbol{x}}_{t\to 0})$ is then used to deduce a less noisy latent $\boldsymbol{z}_{t-1}$, following the DDPM denoising process as outlined in Eq. 4:

$$\boldsymbol{z}_{t-1} = \sqrt{\alpha_{t-1}}\boldsymbol{z}_{t\to 0} + \sqrt{1-\alpha_{t-1}} \cdot \epsilon_\theta(\boldsymbol{z}_t, t, \mathbf{I}, \tau). \quad (7)$$

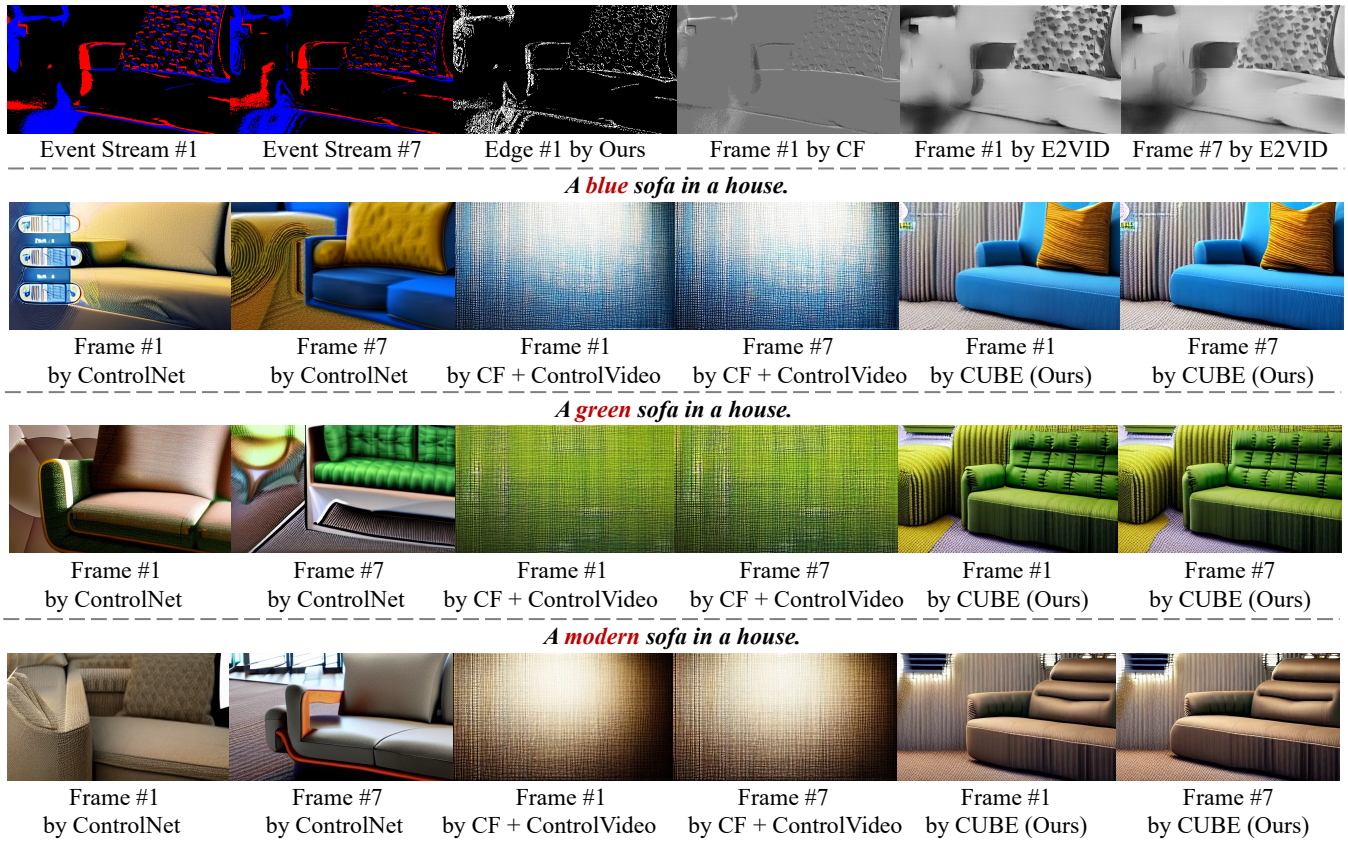

Fig. 3: Qualitative comparisons show CUBE outperforms others in video quality, temporal consistency, and textual alignment.

Method Comparison	Video Quality	Temporal Consistency (%)	Textual Alignment (%)
CUBE (Ours) vs. ControlNet	85.9	100	83.1
CUBE (Ours) vs. CF + ControlVideo	78.2	59.6	76.2

Table 2: User preference study shows the percentage of raters who favor the videos synthesized by CUBE over other method.

4. EXPERIMENT

4.1. Experimental Settings

Implementation Details. In our experiments, short videos are synthesized with lengths of either 7 or 15 frames, whereas longer videos can comprise approximately 100 frames, all rendered at a spatial resolution of 256×448. We utilize DDPM sampling techniques [29] with 50 timesteps, for this process. Thanks to the efficient architecture of xFormers [40], our CUBE framework efficiently generates videos of both 7-frame and 100-frame lengths in about 0.5 and 5 minutes, respectively, using a single NVIDIA RTX 4090.

Dataset. For a comprehensive evaluation of CUBE, we collect 35 object-centric videos from the Vimeo90K dataset [41], and V2E [42] is utilized to generate events. Then, we wrote three textual prompts for each event, resulting in a dataset of 105 event-prompt pairs for testing.

Metrics. Following [43, 44, 32], we adopt CLIP [45] to evaluate the video quality from two perspectives: (a) frame consistency, measured by the average cosine similarity across consecutive frame pairs, and (b) prompt consistency, measured through the average cosine similarity between the input prompt and all video frames.

Baselines. CUBE is benchmarked against two event-based reconstruction methods, CF [46] and E2VID [15, 47], and compared with recent generative methods, ControlNet [38] and ControlVideo [32]. We adapted ControlNet and ControlVideo to support event input, and details are in Sec. 4.3.

4.2. Qualitative and Quantitative Evaluations

Qualitative Comparisons. Figures 2, 3, 4, and 5 illustrate the visual comparisons of synthesized videos by various methods. (a) As observed in Fig. 2, ControlNet lacks tem-

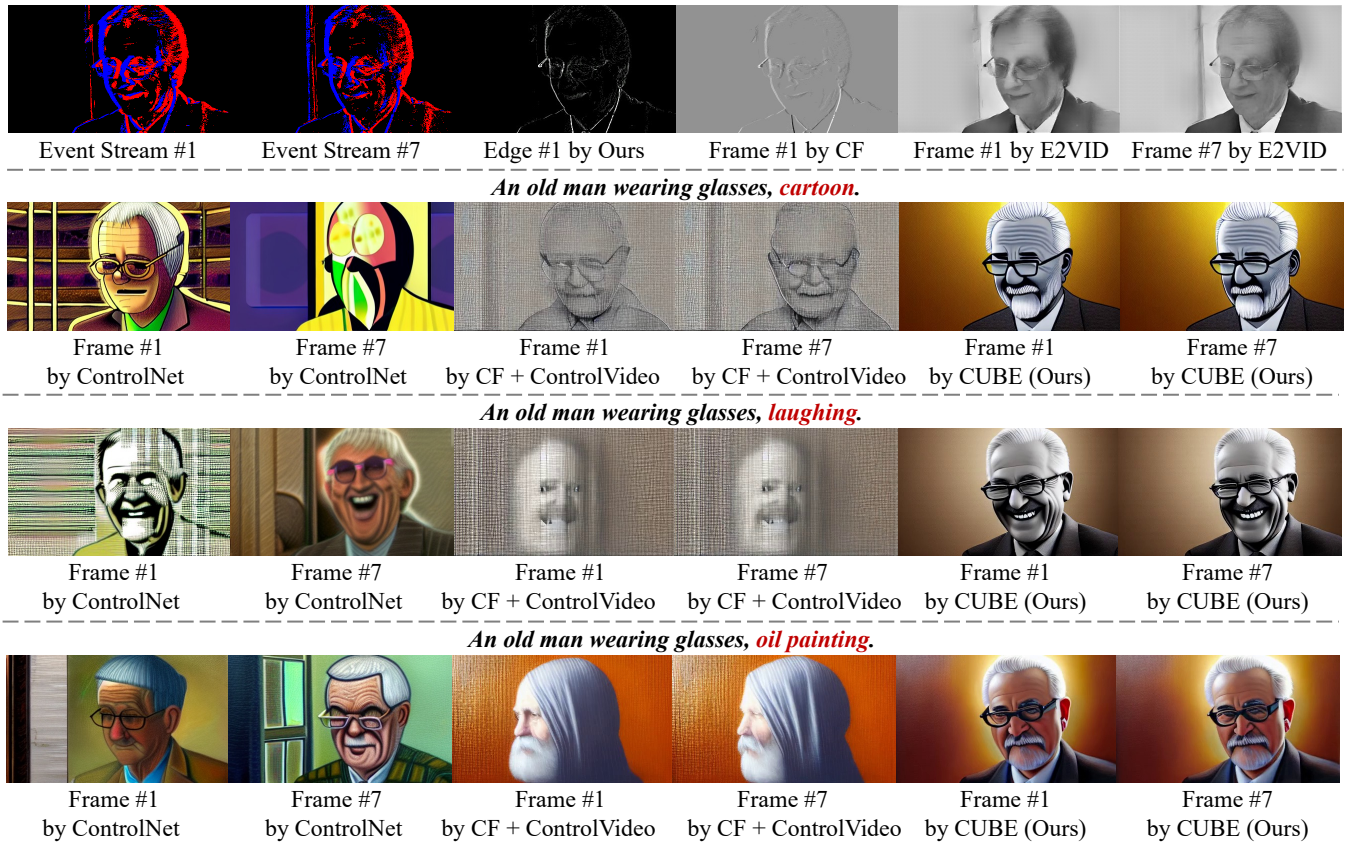

Fig. 4: Qualitative comparisons show CUBE outperforms others in video quality, temporal consistency, and textual alignment.

poral consistency, while ControlVideo, although maintaining temporal coherence, fails to generate a violin. (b) In Fig. 3, ControlNet shows temporal inconsistency and also fails to produce the correct color (green) in Frame #1 of the second row, while ControlVideo generates no meaningful content. (c) Fig. 4 shows unnatural image quality from ControlNet and multiple issues with ControlVideo, including non-compliance with the prompt (cartoon) in the first row, indiscernible images in the second row, and structural discrepancies with the event data in the third row (differing facial orientations). (d) Fig. 5 shows unnatural and inconsistent frame output from ControlNet, with ControlVideo results not aligning with the event data. In contrast, CUBE generates videos with better video quality, temporal consistency and textual alignment.

Quantitative Comparisons. We also compare our CUBE with other methods quantitatively on 105 video-prompt pairs. From Table 1, our CUBE consistently outperforms the baselines in terms of frame and prompt consistency, aligning with our qualitative findings. Despite utilizing the same edges, ControlNet demonstrated worse frame consistency than ours.

User Study. To further validate our CUBE framework, we conduct a user study. Participants are presented with visualizations of event streams, associated text prompts, and videos synthesized by distinct methods, in a random order. They judge the videos based on three criteria: (i) overall video quality, (ii) temporal consistency across all frames, and (iii) alignment between the text prompts and the synthesized videos, using an evaluation set of 105 event-prompt pairs assessed by 5 raters each. From Table 2, our generated videos are preferred across all metrics. In contrast, ControlNet and ControlVideo generally produced lower quality and less consistent videos.

4.3. Ablation Study

Effect of Edge Extraction Module. To demonstrate the effectiveness of the edge extraction module, we conduct a comparison with the variant of ControlVideo. For this variant, frames reconstructed by CF are used as input edge conditions for ControlVideo. However, as depicted in Figures 2, 3, 4, and 5, our edge extraction module demonstrates superior integration with ControlVideo, resulting in improved outcomes.

Effect of Video Generation. The efficacy of our video generation process was evaluated against a variant of ControlNet. It is evident from Figures 2, 3, 4, and 5 that ControlNet struggles to maintain temporal consistency. This observation validates our choice of ControlVideo as the base model for video generation as an effective strategy.

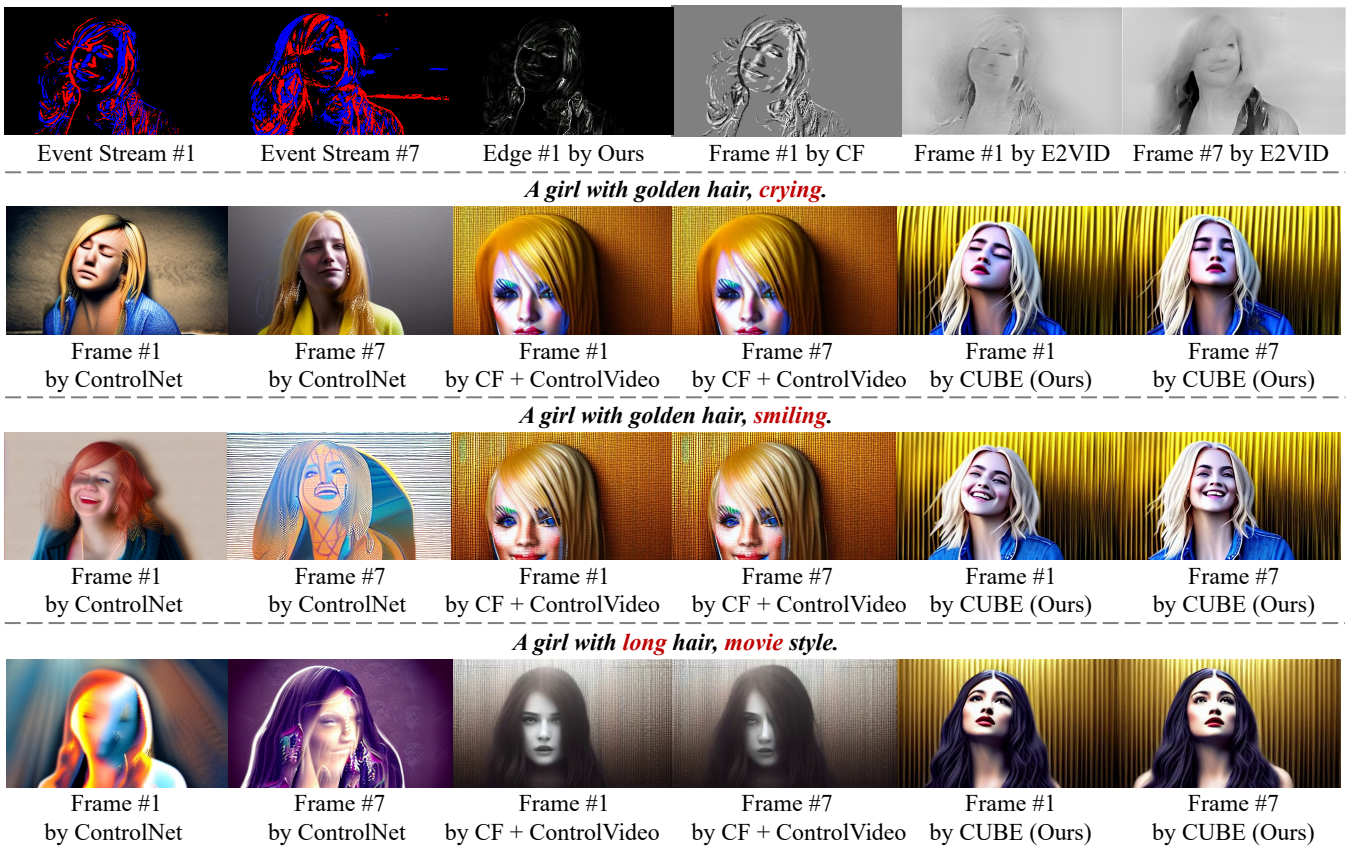

Fig. 5: Qualitative comparisons show CUBE outperforms others in video quality, temporal consistency, and textual alignment.

5. DISCUSSION

The reliance on extracted edge maps may restrict capturing subtle textures and complex patterns. To address this, we plan to explore hybrid models that integrate edge and texture information, enhancing the visual fidelity of generated videos. Furthermore, using pre-trained diffusion models without extensive diverse dataset training might limit adaptation to new or out-of-distribution scenarios. Future work will consider domain adaptation [48] and test-time tuning [49] strategies to improve robustness. Additionally, while CUBE leverages efficient architectures, it is not yet fully optimized to reduce computational demands, further efforts will aim to reduce processing times, paving the way for real-time deployment.

6. CONCLUSION

We introduce CUBE, a framework for controllable unsupervised event-based video generation, which effectively bridges the gap between event cameras and the need for perceptually realistic video synthesis. Combining event-derived edges with textual descriptions, CUBE transcends the limitations of existing methods, offering controllability and superior performance without the requirement of extensive training.

7. REFERENCES

[1] Christian Brandli et al., "A 240× 180 130 db 3 μs latency global shutter spatiotemporal vision sensor," *IEEE Journal of Solid-State Circuits*, vol. 49, no. 10, pp. 2333–2341, 2014.

[2] Shuo Zhu et al., "Computational neuromorphic imaging: principles and applications," in *Computational Optical Imaging and Artificial Intelligence in Biomedical Sciences*, 2024.

[3] Chutian Wang et al., "Neuromorphic shack-hartmann wave normal sensing," *arXiv preprint arXiv:2404.15619*, 2024.

[4] Chutian Wang et al., "Tracking the shack-hartmann spots using neuromorphic motion compensation," in *Computational Optical Sensing and Imaging*, 2023, pp. CTu2B–5.

[5] Shuo Zhu et al., "Removing wall redundancy in non-line-of-sight object-tracking using neuromorphic imaging," in *Computational Optical Sensing and Imaging*, 2023, pp. CTu2B–6.

[6] Pei Zhang et al., "Event encryption: Rethinking privacy exposure for neuromorphic imaging," *Neuromorphic Computing and Engineering*, vol. 4, no. 1, pp. 014002(1–8), January 2024.

[7] Guillermo Gallego et al., "Event-based vision: A survey," *IEEE transactions on pattern analysis and machine intelligence*, vol. 44, no. 1, pp. 154–180, 2020.

[8] Yaping Zhao et al., "Cross-camera human motion transfer by time series analysis," *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2024.

[9] Zhou Ge et al., "Event-based laser speckle correlation for micro motion estimation," *Optics Letters*, 2021.

[10] Zhou Ge et al., "Lens-free motion analysis via neuromorphic laser speckle imaging," *Optics Express*, 2022.

[11] Patrick Bardow et al., "Simultaneous optical flow and intensity estimation from an event camera," in *CVPR*, 2016.

[12] Gottfried Munda et al., "Real-time intensity-image reconstruction for event cameras using manifold regularisation," *International Journal of Computer Vision*, 2018.

[13] Henri Rebecq et al., "Evo: A geometric approach to event-based 6-dof parallel tracking and mapping in real time," *IEEE Robotics and Automation Letters*, 2016.

[14] Yaping Zhao et al., "Adaptive compressed sensing for real-time video compression, transmission, and reconstruction," in *IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, 2023.

[15] Henri Rebecq et al., "High speed and high dynamic range video with an event camera," *IEEE transactions on pattern analysis and machine intelligence*, 2021.

[16] Cedric Scheerlinck et al., "Fast image reconstruction with an event camera," in *the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2020, pp. 156–163.

[17] Timo Stoffregen et al., "Reducing the sim-to-real gap for event cameras," in *European Conference on Computer Vision*, 2020.

[18] Lin Wang et al., "Event-based high dynamic range image and very high frame rate video generation using conditional generative adversarial networks," in *CVPR*, 2019.

[19] Wenming Weng et al., "Event-based video reconstruction using transformer," in *ICCV*, 2021, pp. 2563–2572.

[20] Yunhao Zou et al., "Learning to reconstruct high speed and high dynamic range videos from events," in *CVPR*, 2021.

[21] Jonghyun Choi et al., "Learning to super resolve intensity images from events," in *CVPR*, 2020, pp. 2768–2776.

[22] Bishan Wang et al., "Event enhanced high-quality image recovery," in *European Conference on Computer Vision*, 2020.

[23] Lin Wang et al., "Eventsr: From asynchronous events to image reconstruction, restoration, and super-resolution via end-to-end adversarial learning," in *CVPR*, 2020.

[24] Quanmin Liang et al., "Event-diffusion: Event-based image reconstruction and restoration with diffusion models," in *the 31st ACM International Conference on Multimedia*, 2023.

[25] Jonathan Ho et al., "Denoising diffusion probabilistic models," *Advances in neural information processing systems*, 2020.

[26] Robin Rombach et al., "High-resolution image synthesis with latent diffusion models," in *CVPR*, 2022, pp. 10684–10695.

[27] Jascha Sohl-Dickstein et al., "Deep unsupervised learning using nonequilibrium thermodynamics," in *International conference on machine learning*, 2015, pp. 2256–2265.

[28] Yang Song et al., "Generative modeling by estimating gradients of the data distribution," *Advances in neural information processing systems*, vol. 32, 2019.

[29] Yang Song et al., "Score-based generative modeling through stochastic differential equations," *arXiv preprint arXiv:2011.13456*, 2020.

[30] Hengyuan Ma et al., "Accelerating score-based generative models with preconditioned diffusion sampling," in *European Conference on Computer Vision*, 2022, pp. 1–16.

[31] Elias Mueggler et al., "The event-camera dataset and simulator: Event-based data for pose estimation, visual odometry, and slam," *The International Journal of Robotics Research*, vol. 36, no. 2, pp. 142–149, 2017.

[32] Yabo Zhang et al., "ControlVideo: Training-free controllable text-to-video generation," *International Conference on Learning Representations (ICLR)*, 2024.

[33] Pei Zhang et al., "Neuromorphic imaging with density-based spatiotemporal denoising," *IEEE Transactions on Computational Imaging*, vol. 9, pp. 530–541, May 2023.

[34] Pei Zhang et al., "Neuromorphic imaging and classification with graph learning," *Neurocomputing*, 2024.

[35] Pei Zhang et al., "Neuromorphic imaging with joint image deblurring and event denoising," *IEEE TIP*, 2024.

[36] Shansi Zhang et al., "Light field image restoration via latent diffusion and multi-view attention," *IEEE Signal Processing Letters*, vol. 31, pp. 1094–1098, 2024.

[37] Zhen Yuen Chong et al., "Solving inverse problems in compressive imaging with score-based generative models," in *IEEE DSAA*, 2023.

[38] Lvmin Zhang et al., "Adding conditional control to text-to-image diffusion models," in *ICCV*, 2023, pp. 3836–3847.

[39] Zhewei Huang et al., "Real-time intermediate flow estimation for video frame interpolation," in *ECCV*, 2022, pp. 624–642.

[40] Benjamin Lefaudeux et al., "xformers: A modular and hackable transformer modelling library," 2021.

[41] Tianfan Xue et al., "Video enhancement with task-oriented flow," *International Journal of Computer Vision*, 2019.

[42] Yuhuang Hu et al., "V2E: From video frames to realistic dvs events," in *CVPR*, 2021, pp. 1312–1321.

[43] Patrick Esser et al., "Structure and content-guided video synthesis with diffusion models," in *ICCV*, 2023, pp. 7346–7356.

[44] Jay Zhangjie Wu et al., "Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation," in *ICCV*, 2023, pp. 7623–7633.

[45] Alec Radford et al., "Learning transferable visual models from natural language supervision," in *ICML*, 2021.

[46] Cedric Scheerlinck et al., "Continuous-time intensity estimation using event cameras," in *Asian Conference on Computer Vision*, 2018, pp. 308–324.

[47] Henri Rebecq et al., "Events-to-video: Bringing modern computer vision to event cameras," in *CVPR*, 2019.

[48] Shai Ben-David, John Blitzer, Koby Crammer, and Fernando Pereira, "Analysis of representations for domain adaptation," *Advances in neural information processing systems*, 2006.

[49] Yaping Zhao et al., "Improving video colorization by test-time tuning," in *IEEE International Conference on Image Processing*, 2023.