E-Motion: Future Motion Simulation via Event Sequence Diffusion

Song Wu [1], **Zhiyu Zhu** [2], **Junhui Hou** [2], **Guangming Shi** [1] , **Jinjian Wu** [1] *
[1] Xidian University, [2] City University of Hong Kong
swu_666@stu.xidian.edu.cn,zhiyuzhu2-c@my.cityu.edu.hk
jh.hou@cityu.edu.hk,gmshi@xidian.edu.cn,jinjian.wu@mail.xidian.edu.cn

Abstract

Forecasting a typical object's future motion is a critical task for interpreting and interacting with dynamic environments in computer vision. Event-based sensors, which could capture changes in the scene with exceptional temporal granularity, may potentially offer a unique opportunity to predict future motion with a level of detail and precision previously unachievable. Inspired by that, we propose to integrate the strong learning capacity of the video diffusion model with the rich motion information of an event camera as a motion simulation framework. Specifically, we initially employ pre-trained stable video diffusion models to adapt the event sequence dataset. This process facilitates the transfer of extensive knowledge from RGB videos to an event-centric domain. Moreover, we introduce an alignment mechanism that utilizes reinforcement learning techniques to enhance the reverse generation trajectory of the diffusion model, ensuring improved performance and accuracy. Through extensive testing and validation, we demonstrate the effectiveness of our method in various complex scenarios, showcasing its potential to revolutionize motion flow prediction in computer vision applications such as autonomous vehicle guidance, robotic navigation, and interactive media. Our findings suggest a promising direction for future research in enhancing the interpretative power and predictive accuracy of computer vision systems. The source code is publicly available at https://github.com/p4r4mount/E-Motion.

1 Introduction

Accurately capturing and interpreting dynamic scenes under fluctuating motion and illumination conditions remains an enduring challenge in computer vision [27, 33]. This challenge is particularly pronounced in real-world settings, where subtle variations can dramatically affect the perception and analysis of future motion [60, 11]. Traditional imaging modalities often struggle to capture these nuances, leading to a gap in accurately modeling and predicting motion flow in complex visual environments.

The rapid advancements in deep learning have catalyzed transformative developments in computer vision, particularly in the generative models [45, 16, 18, 25, 46, 48, 49]. Video diffusion models [1, 19, 32, 10, 2], which stand at the forefront of these innovations, leverage stochastic diffusion processes to generate, restore, and accurately manipulate video content. These models, emblematic of the state-of-the-art in temporal data processing, offer refined capabilities for complex video-based tasks [5, 39], underscoring the significant strides made in understanding and interpreting dynamic visual scenes.

*This work was supported in part by National Key Research and Development Program of China(2023YFA1008500), in part by NSFC Excellent Young Scientists Fund 62422118, and in part by Hong Kong Innovation and Technology Fund ITS/164/23. The first two authors contributed to this paper equally. Corresponding author: Jinjian Wu.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

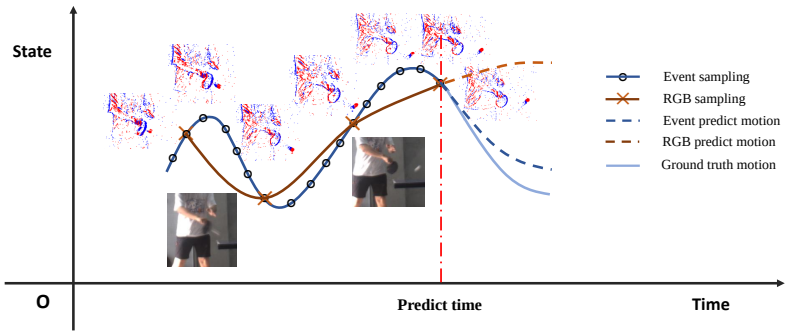


Figure 1: Illustration that the exceptional temporal resolution afforded by event cameras, alongside their distinctive event-driven sensing paradigm, presents a significant opportunity for advancing the precision in predicting future motion trajectories.

Coming into the high temporal resolution field, event data stands as a revolutionary sensing approach [37, 12, 53] that significantly mitigates this gap and consistently captures even very subtle fluctuations. Such strong capacity comes from the unique sensing pattern of the event camera, which asynchronously measures the intensity variation in high temporal resolution. It allows for the precise detection of miniature illumination changes, providing a rich, granular record that traditional cameras simply cannot offer [67, 68, 7].

To enable the video diffusion model to completely and correctly learn concise motion information for the estimation of potential future movement, integrating it with event data to realize an event sequence generation model is a potential solution that may inject high-frequency motion information into the video diffusion model. This paper delves into the symbiosis of video diffusion models and high-temporal-resolution event data, exploring its potential to redefine motion forecasting in computer vision. We commence by delineating the landscape of video diffusion models and the mechanics of event-based sensing, elucidating their complementary strengths. Subsequently, we introduce a novel framework that amalgamates these technologies, aiming to augment the precision of motion flow predictions. Through comprehensive experimentation, we validate the effectiveness of our methodology, demonstrating its superior performance across diverse scenarios. Our findings illuminate the path forward, showcasing the integration of video diffusion models with event data as a robust, innovative solution for capturing and interpreting the complexities of dynamic scenes with unparalleled detail and accuracy.

In summary, the contribution of this paper lies in the following three parts:

- we make the first attempt to integrate event-sequences with a video diffusion model, resulting in an event-sequence diffusion model, which could potentially estimate future object motion, given by a certain event prompt;
- we propose to align the pre-trained event-sequence diffusion model with the real-world motion via a reinforcement learning process, which stabilizes the results generated by the diffusion model and makes them more closely resemble real motion.
- we integrate a test-time prompt augmentation method to make use of high temporal resolution event sequence prompt to enhance the generation performance.

2 Related Work

2.1 Event-based Vision

Event-based vision represents a paradigm shift from traditional frame-based imaging, offering a dynamic and highly granular approach to capturing visual information. Unlike conventional cameras that record static frames at fixed intervals, the event-based sensor (developed by Lichtsteiner *et al.* [28] and further elaborated by Posch *et al.* [36]) asynchronously records intensity changes per pixel and generates a signal termed an ”event” whenever the intensity surpasses a threshold, thereby providing a continuous stream of data that reflects temporal changes with remarkable precision. This

method is particularly effective in environments with rapid motion or varying illumination, where traditional cameras suffer from motion blur and latency issues.

Recent advancements in this field have focused on leveraging the high temporal resolution of event data for various applications, including high-speed tracking [57, 30, 65], dynamic scene reconstruction [43, 23, 69], and optical flow estimation [44, 9, 22]. Works by Gallego *et al.* [31] and Rebecq *et al.* [38] have been instrumental in demonstrating the utility of event-based data in reconstructing high-speed phenomena and enhancing motion analysis, setting a solid foundation for our research.

2.2 Multimodal Diffusion Models

Generative diffusion models, introduced by Sohl-Dickstein *et al.* [45], represent a class of probabilistic generative models that simulate the gradual transformation of data from a complex distribution into a simpler, typically Gaussian distribution, and vice versa. [8, 25, 61, 4]. This process, characterized by a series of forward and reverse diffusion steps, has been applied successfully to a range of tasks, including image synthesis [20, 34], restoration [29, 21], and, more recently, temporal data manipulation [32, 15].

The application of diffusion models to video data, as explored by Ho *et al.* [19] and extended by others, marks a significant advancement in the field, offering new pathways for the generation and manipulation of dynamic scenes. These models have shown exceptional promise in capturing the temporal continuity and complexity inherent in video data, providing a robust framework for tasks such as video prediction and temporal interpolation.

In recent years, the development of multimodal diffusion technology has advanced rapidly. Researchers are dedicated to applying the powerful generative capabilities of diffusion to different modalities with unique advantages, such as optical flow and depth. Saxena et al. [40] was the first to apply diffusion models to optical flow and depth estimation. For the characteristics of training data, they introduced infilling, step-rolling, and L1 loss during training to mitigate distribution shifts between training and inference. To address the lack of ground truth in datasets, they also used a large amount of synthetic data for self-supervised pretraining, enabling the diffusion model to acquire reliable knowledge. Chen et al. [6] utilized the motion information embedded in control signals such as edges and depth maps to achieve more precise control over the text-to-video (T2V) process. They used pixel residuals and optical flow to extract motion-prior information to ensure continuity in video generation. Additionally, they proposed a first-frame generator to integrate semantic information from text and images.

Despite the extensive body of research within the domain of temporal analysis, it is important to acknowledge that the majority of these studies focus primarily on the domain of RGB images and videos. As previously discussed, the superior temporal resolution offered by event data holds significant potential for enhancing the alignment process. Consequently, it is imperative to undertake a thorough investigation into the application and adaptation of existing pre-trained diffusion models to the realm of event data.

3 Preliminary

Diffusion Models are a class of generative models that simulate the gradual transformation of data from a complex, high-dimensional distribution to a simpler, typically Gaussian distribution through a process known as forward diffusion [48, 18, 46, 47]. Conversely, the reverse diffusion process aims to reconstruct the original data distribution from the simpler one. This mechanism is inspired by thermodynamic processes and has been increasingly applied in the field of deep learning for generating high-quality, diverse samples from complex distributions.

The mathematical foundation of diffusion models is rooted in stochastic differential equations (SDEs), which describe the forward and reverse diffusion processes. By score-based formulation [48], the forward process of diffusion model acts as

$$d\mathbf{x} = \mathbf{f}(\mathbf{x}, t)dt + g(t)d\mathbf{w}, \tag{1}$$

where $\mathbf{x} \in \mathbb{R}^n$ indicates the state of reconstructed signal, $\mathbf{w} \in \mathbb{R}^n$ represents a standard Wiener process ($g(t)d\mathbf{w} \sim \mathcal{N}(0, g(t)^2\mathbf{I}d\mathbf{w})$), $\mathbf{f}(\cdot, t) : \mathbb{R}^n \to \mathbb{R}^n$ and $g(\cdot) : \mathbb{R} \to \mathbb{R}$ represent the drift and

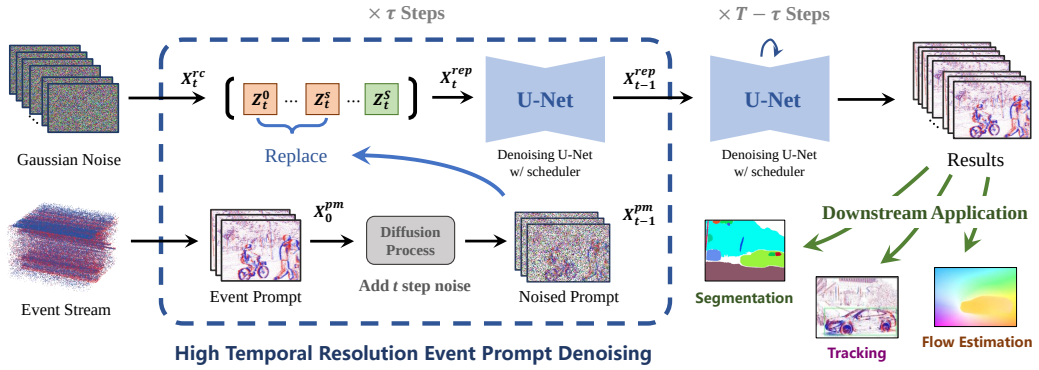


Figure 2: Inference workflow of the proposed method, where the left upper one indicates the random Gaussian noise, left lower one represents the prompted event sequence. We perform τ steps forward diffusion processing on the event prompt and substitute a portion of the diffusion input noise, followed by $T - \tau$ Steps of conventional denoising.

diffusion coefficients, respectively. Moreover, in the evaluating phase, the reverse (inference) process could be illustrated as iteratively performing the following ODE step:

$$d\mathbf{x} = [\mathbf{f}(\mathbf{x}, t) - \frac{1}{2} g(t)^2 \nabla_{\mathbf{x}} logp(\mathbf{x})]dt, \tag{2}$$

where $\nabla_{\mathbf{x}} logp(\mathbf{x})$ is usually approximated by a learnable score model $S_\theta(\mathbf{x}, t)$. Based on the theoretical formulation of stable video diffusion [1], e.g., variance exploding (VE) diffusion process [48], the reverse process is acted as

$$\mathbf{x}_{t-1} = \mathbf{x}_t - \frac{\mathbf{x}_t - \mu_\theta(\mathbf{x}_t, t)}{\sigma_t}(\sigma_t - \sigma_{t-1}), \tag{3}$$

where $\mu_\theta(\cdot)$ is one of parametrization method of $S_\theta(\cdot)$, which estimate the clean image from noise latent $\mathbf{x}_t \sim \mathcal{N}(\mathbf{x}_0, \sigma_t^2 \mathbf{I})$.

4 Methods

The endeavor of video diffusion presents multifaceted challenges, notably the simultaneous generation of sequences that not only manifest dynamic, convincing motion but also maintain authentic textures. The utilization of event data, characterized by its intrinsic high temporal resolution, emerges as a strategic solution for the precise modeling of object motion. In contrast to traditional stable video diffusion models, which are typically initialized using a singular image or textual prompt, our proposed methodology, designated as the Event-Sequence Diffusion Network, capitalizes on a succinct sequence of events as its conditioning input. This novel approach is illustrated in Fig. 2.

As elaborated in Sec. 4.1, we delineate the comprehensive pre-training regimen that facilitates the learning of object motion through the prediction of subsequent events. Nevertheless, due to the intrinsic diversity-generating property of diffusion models, they commonly yield several distinct samples from a single input prompt. Although these variations may each seem feasible, there is no assurance that they align consistently with the actual dynamics of real-world motion. Thus, in Sec. 4.2, we augment the quality of generation by incorporating an alignment process. It employs reinforcement learning techniques to impose a structured regularization on the generative process of the diffusion models, thereby enhancing the fidelity and coherence of the produced sequences.

4.1 Learning Motion Prior via Pretraining on Event Sequences

An initial formatting process is discussed to retain its high-temporal resolution properties effectively and to facilitate integration with pre-existing large-scale models. Furthermore, to mitigate the substantial computational demands associated with video diffusion models, we have implemented a prompt sampler designed to enhance the efficiency of information encapsulation derived from event

Algorithm 1 Motion Alignment Process

Input: Reference model θ', training model θ and reward model $\mathcal{R}(\mathbf{x})$.
For $i = 1, ..., I$ **do**
  Sample reference trajectory $\{x_O|\pi_{\theta'}\}, \mathcal{R}(\mathbf{x}_O)$ and $P_{\theta'}$.
  For $t = 1, ..., T$ **do**
    Calculate P_θ based on reference trajectory $\{x_O|\pi_{\theta'}\}$.
    Optimizing $\nabla_\theta \mathcal{J}'(\mathbf{x})$.
  End for
  $\theta' \leftarrow \theta$.
End for
Return θ

sequence diffusion frameworks. Subsequently, we will delineate these critical aspects in a detailed manner.

Event Representation. To leverage contemporary video diffusion models for event data generation, we adopt a strategy where both event information and corresponding images are concurrently inputted into the video diffusion framework, which then processes adaptively sampled outcomes. Specifically, as illustrated in Fig. 2, for a given event denoted as $\mathbf{E} = \{h, w, p, t\}$, we consolidate the event data into a voxel grid representation [66], symbolized as $\bar{\mathbf{E}} \in \mathbb{R}^{B\times H\times W}$, where B denotes the number of time bins. In order to utilize the rich pre-training information from RGB frames, we set $B = 3$. This approach ensures that the video diffusion model can effectively interpret and integrate the high-dimensional event data alongside conventional image inputs, facilitating a more comprehensive synthesis of dynamic visual content.

Pretraining. The training regimen for our proposed approach is similar with that of diffusion models [48, 24]. This methodology involves estimating the clean image from perturbed samples, as

$$\mathcal{L}_{pre} = \mathbb{E}_t\{\lambda(t)\mathbb{E}_{\mathbf{x}_0}\mathbb{E}_{\mathbf{x}_t|\mathbf{x}_0}[\|x_0 - \mu_\theta(\mathbf{x}_t, t)\|_2^2]\}, \tag{4}$$

where $\lambda(t)$ denotes a weighting function, parameterized as $\frac{\sigma_t^2+\sigma_{data}^2}{(\sigma_t+\sigma_{data})^2}$ During training, we randomly select a set of intermediate states $\mathbf{X}_t$ and apply regularization techniques to guide them towards the accurate estimation of the underlying noise component ϵ. Moreover, to make the diffusion network adaptively capture the object motion with different time windows, in the training process we randomly augment the diffusion network with the voxel from different time ranges. (*Please refer to the Appendix Sec. A for more details*)

High-temporal Resolution Guided Sampling. Owing to the unique operational mechanism of event cameras, which detect changes in intensity rather than capturing static images, the clarity and definition of recorded objects can significantly diminish if the chosen temporal window is either too brief or excessively prolonged, especially when we only feed one prompt, like a standard stable video diffusion network. This characteristic often results in blurred or indistinct imagery when the temporal resolution does not align optimally with the scene's dynamics.

Inspired by the recent advancement of guided sampling [35], we propose to aggregate multiple high-temporal resolution event frames as a test-time prompt. Specifically, as illustrated in Fig. 2, during the inference process, we feed $s > 1$ frames instead of a single frame as prompt $\mathbf{X}_0^{pm} \in \mathcal{R}^{s\times h\times w\times 3}$. For the step T, both $\mathbf{X}_T^{pm}$ and $\mathbf{X}_T^{rc}$ are initialized with the random Gaussian noise. However, for the step $t < \tau$, we set $\mathbf{X}_{t-1}^{pm} \leftarrow \mathbf{X}_0^{pm} + \sigma_t\epsilon$, Subsequently, the noised event prompt $\mathbf{X}_{t-1}^{pm}$ replaces the first s random noises in the noise tensor $\mathbf{X}_t^{rc}$, providing motion priors for the denoising process. (*Please refer to the Appendix Algorithm 2 for more details*)

4.2 Motion Alignment via Reinforcement Learning

Given the inherent challenges in precisely tailoring diffusion models to fit the entirety of the training data, particularly when considering the disparity between the model's size and the volume of the dataset, it is pragmatic to direct potential losses towards regions less perceptible to end-users or higher-level algorithms. Furthermore, the multi-step nature of the diffusion generation process renders the simultaneous training of the entire pipeline nearly unfeasible. To address this, we employ a strategy

of reinforced optimization, conceptualizing the generation process as a Markov chain. This approach underscores the importance of guiding the diffusion process to yield results of superior quality, thereby optimizing the model's performance while accommodating its structural and computational complexities. Optimizing such a diffusion model starts with the following equations:

$$\max_{\theta} \mathcal{J} = \int_{\mathbf{P}_\theta(\mathbf{x})} \mathcal{R}(\mathbf{x}) d\mathbf{x}, \tag{5}$$

where $\mathbf{P}_\theta(\mathbf{x})$ indicates the distribution of reconstructed samples under the model weight θ. Although the diffusion model is a probabilistic model, its weights are generally deterministic. [46] The randomness generally comes from Gaussian sampling. Thus, we adopt the same measurement as [64] to utilize the Gaussian density function $\mathcal{N}(\mathbf{x}_t|\mu_t, \sigma) \sim \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(\mathbf{x}_t - \mu(\mathbf{x}_t, t))^2}{2\sigma^2}}$ to measure the $\mathbf{P}_\theta(\mathbf{x})$ of a given sample $\mathbf{x}_t$, where $\mu(\cdot)$ indicates the estimated clean latent without adding random noise. We then adopt the policy gradient descent method, i.e., PPO [41, 42], to optimize the alignment process of these diffusion models. The policy gradient descent optimization process is formulated as follows:

$$\nabla_\theta \mathcal{J}'(\mathbf{x}) = -\sum \frac{\nabla_\theta P_\theta}{P_{\theta'}} \mathcal{R}(\mathbf{x}_O) + \lambda \nabla_\theta \mathcal{KL}(P_{\theta'}|P_\theta), \tag{6}$$

where $\mathcal{KL}(\cdot|\cdot)$ indicates the KL-divergence between two distributions. The training process is shown as Algorithm 1. Note that due to the huge GPU memory and time consumption of the video diffusion process, we distribute the data generation and network training on different GPUs. Please refer to Sec.5 **Settings** for more details.

Modeling of Reward. To measure the quality of the reconstructed event frame, we utilize the *FVD* and *SSIM* as the reward $\mathcal{R}(\cdot)$ to guide the training process of reinforcement alignment. Moreover, to remove the bias of rewards, we randomly generate M samples given one prompt, forming a pair of samples.

5 Experiments

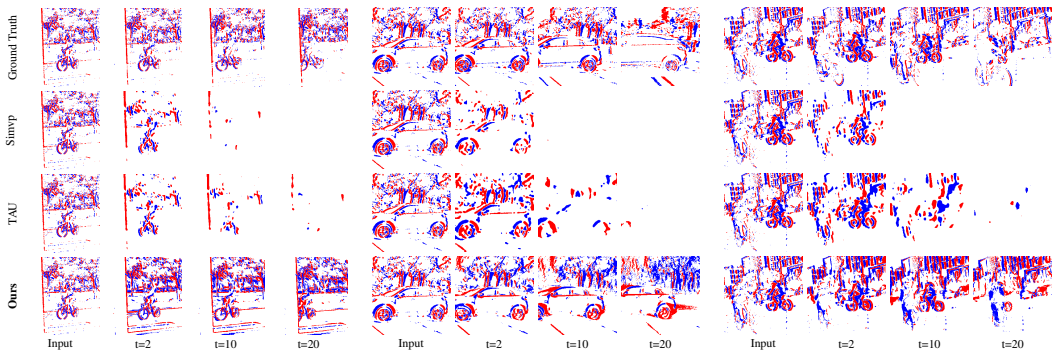
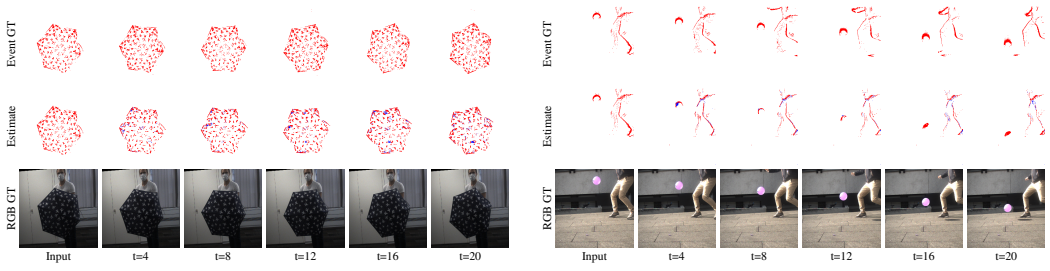


Figure 3: Qualitative comparison between SOTA methods. The first row of each sequence represents the ground truth of the event sequence. The second and third rows respectively depict the results of future event estimation by SimVP [14] and TAU [50]. The final row represents the results obtained by our method. The complete sequence is shown in Fig.9.

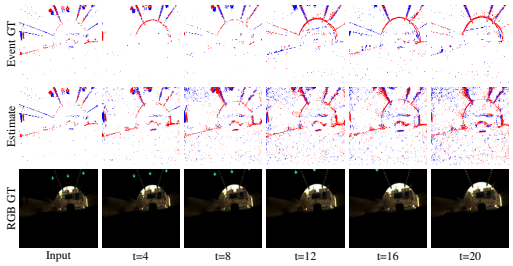
Dataset. In our study, we utilize two large-scale event datasets, i.e., VisEvent [55] and EventVOT dataset [56]. The VisEvent dataset encompasses a wide variety of scenes, including 820 RGB-Event video pairs and 371,128 RGB frames. Moreover, it was captured by the DAVIS346 camera [3], with resolutions of 346×260 for RGB and events. It addresses diverse environmental conditions and includes 17 distinct attributes such as camera motion, low illumination, and background clutter, facilitating detailed performance analysis under various challenging scenarios.

The EventVOT dataset provides 1,141 high-definition videos with 569,359 frames, making it the largest dataset in this domain. It features a resolution of 1280×720, encompassing 19 diverse classes of target objects. Compared to earlier datasets such as VOT-DVS, TD-DVS, and Ulster from 2016, and more recent ones like FE108 and COESOT, EventVOT offers an unprecedented scale and variety,

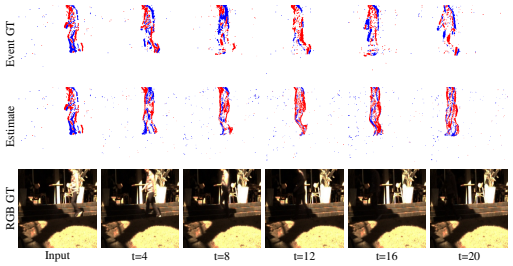


(a) Rotating Umbrella.

(b) Falling Balloon.



(c) Poor exposure: The car driving out of the tunnel.



(d) Occlusion: A person walking to shadow.

Figure 4: More visualization of our method's prediction in various scenarios. The results of the complete sequence along with other methods are presented in Fig. 12 and Fig. 13.

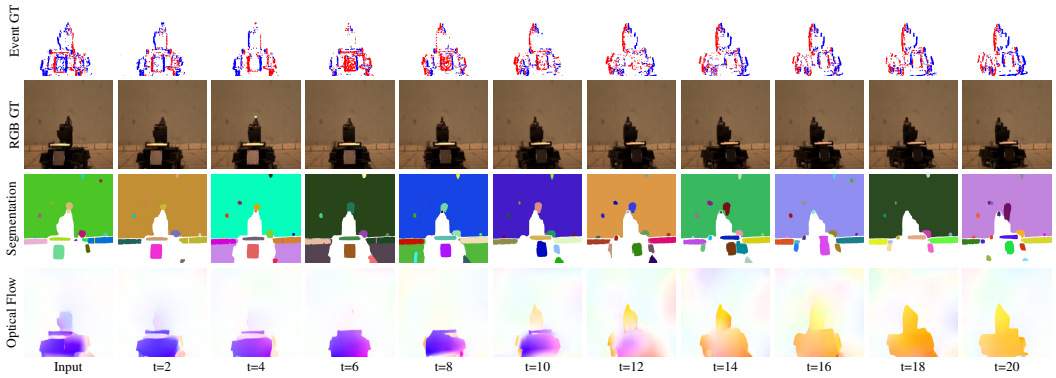
Table 1: Quantitative comparison between SOTA methods, where SVD denotes the standard stable video diffusion network. "VID‘' represents the video and "EVT" indicates the event data. ↑ (resp.↓) represents the bigger (resp. lower) the better.

Methods	Modal	*FVD* ↓	*MSE* ↓	*SSIM* ↑	*LPIPS* ↓	*mIoU* ↑	*aIoU* ↑
PhydNet [17]	VID	1602.84	0.0295	0.4299	0.6048	0.077	0.348
SimVP [13]	VID	1347.57	0.0216	0.6261	0.3051	0.264	0.520
TAU [50]	VID	1371.65	0.0240	0.6381	0.3026	0.264	0.529
PredRNNv2 [59]	VID	1266.83	0.0176	0.5550	0.2407	0.275	0.540
SVD [1]	VID	1122.54	0.0246	0.6451	0.3299	0.233	0.506
PredRNNv2 [59]	EVT	1339.05	0.0306	0.6598	0.3388	0.166	0.504
SimVP [13]	EVT	1242.25	<u>0.0210</u>	0.7961	0.3371	0.213	**0.532**
TAU [50]	EVT	<u>1218.03</u>	0.0231	0.7972	<u>0.3354</u>	<u>0.228</u>	0.514
Ours	EVT	**1055.25**	**0.0170**	**0.7998**	**0.3123**	**0.302**	<u>0.522</u>

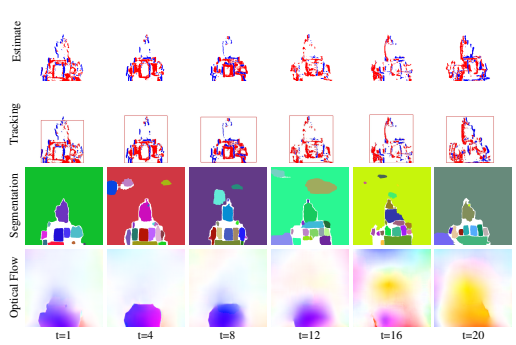
including a substantial number of videos and a wide array of environmental conditions, aimed at improving the development and evaluation of event-based visual tracking algorithms. The dataset is meticulously annotated and divided into training (841 videos), validation (18 videos), and testing (282 videos) subsets, ensuring a comprehensive framework for robust algorithm testing and benchmarking.

Settings. All experiments are conducted on machines with $8\times$ GeForce RTX 3090 GPUs, Intel(R) Core(TM) i7-10700 CPU of 2.90GHz, and 64-GB RAM. In the pre-training stage, we employed the ADAM optimizer with the exponential decay rates $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The total training process was 20000 iterations for both kinds of noise experiments. We initialized the learning rate as 1e-5. We set the batch size to 128 (with 8 gradient accumulation steps).

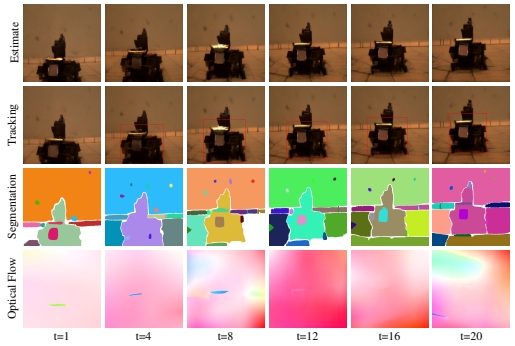
For the alignment process, due to the fact that it takes quite a lot of GPU memory to generate a reference trajectory during the training process, we have to achieve the trajectory generation and reinforcement alignment in a parallel and distributed manner. Specifically, we utilize $4\times$ RTX3090s to train reinforcement learning alignment processes. Moreover, $4\times$ RTX3090s is utilized to generate training trajectory data. The updating episode of the reinforcement learning process is set at 100 optimization steps. (*Please refer to the Appendix for the detailed illustration of*) We also employed the ADAM optimizer with the exponential decay rates $\beta_1 = 0.9$ and $\beta_2 = 0.999$. We initialized



(a) Segmentation and flow estimation results using GT event and RGB frame.



(b) Downstream results using our estimated events.



(c) Downstream results using SVD estimated frames.

Figure 5: Visualization results on downstream tasks, where we show the tasks of tracking, segmentation, and flow estimation. (a) denotes the ceiling performance of settings (b) and (c).

the learning rate as $2e^{-6}$ and set the batch size to 16(with 2 gradient accumulation steps) in all experiments.

Comparison Methods. To comprehensively evaluate the performance of the proposed method, we compared different methods, which could be briefly divided into two categories: image-based future frame prediction and event-based future motion prediction. For the image-based future frame prediction, we adopt the original stable video diffusion [1] with some popular time serious forecasting methods, e.g., PredRNN [58], SimVP [14], TAU [51]. Moreover, for future event forecasting, we simply use stable video diffusion [1] as our baseline since other methods make it hard to achieve desirable performance on those datasets.

Metrics. In the evaluation of our proposed model, we employ a comprehensive suite of metrics designed to assess the quality. We incorporate the *Fréchet Video Distance (FVD)* [52] to evaluate the temporal coherence and visual quality of generated video sequences against real video content. The *Mean Squared Error (MSE)* serves as a fundamental metric, quantifying the average squared difference between the estimated values and the actual values. The *Structural Similarity Index Measure (SSIM)* is employed to assess the visual impact of structural information, brightness, and contrast differences between the generated images and the ground truth. Lastly, the *Learned Perceptual Image Patch Similarity (LPIPS)* [63] metric is adopted to evaluate the perceptual similarity between the generated and real images. In addition, we apply *mIoU* and *aIoU* as metrics for the object segmentation task to validate the feasibility of the generated events in downstream tasks.

5.1 Quantitative and Qualitative Comparison Results

The experimental outcomes are illustrated in Table 1. It is evident that by applying regularization to event data, our methodology attains performance levels comparable to state-of-the-art (SOTA) methods. Notably, in metrics specific to the event domain, our approach outperforms the comparative

Table 2: Quantitative comparison between SOTA methods on object tracking. SVD denotes the standard stable video diffusion network. "VID‘' represents the video and "EVT" indicates the event data. ↑ (resp.↓) represents the bigger (resp. lower) the better. R_{upper} denotes a comparison of task performance with GT Event, where we average the ratio of different metrics (please refer to the Appendix for the detailed comparisons).

Methods	Modal	*AUC* ↑	*OP50* ↑	*OP75* ↑	*PR* ↑	R_{upper} ↑
PredRNN [59]	EVT	28.54	22.24	17.24	14.96	0.496
SimVP [13]	EVT	35.10	32.41	18.87	21.74	0.640
TAU [50]	EVT	34.95	31.85	20.31	24.74	0.669
Ours	EVT	43.77	47.77	27.31	30.99	**0.891**
GT Event	EVT	47.87	53.30	31.93	34.62	1.000

Table 3: Ablation Study of Pre-training Phase. All models are tested with only feeding single event voxel frame, where "#Prompt" indicates the number of training prompt event data, $\mathcal{U}(1,3)$ indicates to randomly select 1 to 3 frames as training prompt. 'Fine-tuning' refers to the parameters that are fine-tuned, 'T' indicates the fine-tuning of all temporal attention parameters, and 'S+T' indicates the simultaneous fine-tuning of both spatial and temporal attention parameters. 'CLIP' indicates whether the CLIP features extracted have been fine-tuned for events. 'RGB' refers to the CLIP model pre-trained on RGB data, while 'Event' indicates the CLIP model fine-tuned with event data.

#Prompt	Fine-tuning	CLIP	*FVD* ↓	*SSIM* ↑	*LPIPS* ↓	*mIoU* ↑	*aIoU* ↑
$\mathcal{U}(1,3)$	T	RGB	1972.91	0.69513	0.3651	0.266	0.507
$\mathcal{U}(1,3)$	S+T	RGB	**1378.92**	**0.78496**	**0.3076**	0.252	**0.525**
1	S+T	RGB	1406.24	0.78374	0.3219	**0.268**	0.524
1	S+T	Event	1646.78	0.72779	0.3464	-	-

method of SVD. The visual results, as shown in Fig. 3, demonstrate that compared to TAU, SimVP, and other methods, our approach can predict longer durations, more precise motion, and generate events that are more stable and closer to the ground truth.

In addition to the standard tests, we further validated our method on the BS-ERGB [54] dataset, which was captured using a completely different event camera. As shown in Fig. 4a and Fig. 4b, our method achieves satisfactory prediction results on this dataset, demonstrating its effectiveness. Furthermore, we conducted experiments in extreme scenarios, with the visualization results shown in Fig. 4c and Fig. 4d. Even under stringent lighting conditions, our method is still able to predict human and viewpoint motion with high accuracy, indicating strong robustness.

For downstream task results, we conducted both qualitative and quantitative evaluations on the tasks of target segmentation and object tracking, and performed qualitative evaluations on optical flow estimation. Fig. 5 shows the qualitative results on downstream tasks. Due to the motion information priors provided by the **high temporal resolution event prompt**, our method can accurately predict the tank's turning direction, whereas using only a single input RGB frame makes this difficult. Since directly comparing cross-modal results does not provide meaningful insights, we defined an upper bound ratio R_{upper} based on the upper bound of target tracking performance for comparison. Table 2 presents the quantitative comparison results, showing that our method significantly outperforms other methods. *For more detailed definitions and results, please refer to the Appendix materials.*

5.2 Ablation Study

Fine-tuning Layers. As delineated in Table 3, we conduct empirical validations across various fine-tuning configurations, including temporal-only and joint spatial-temporal adjustments. Note that, during the refinement process, we keep the neural network with a similar amount of training weights. Moreover, all methods are trained under the same configuration, excluding the training parameters. The results from these experiments provide critical insights into the optimal configuration settings that enhance the accuracy and efficiency of event-based video generation. Besides, the authors also want to note that it's ineffective to only change some parts of a large generative model, since the diffusion U-Net is trained with the perception of original VAE, CLIP models generation distributions. We also

Table 4: Ablation Study of motion alignment and multi prompt. All models are tested with only feeding single event voxel frame. 'EP' denotes denoising using the high temporal resolution event prompt, and 'MA' denotes motion alignment based on reinforcement learning.

Method	**EP**	**MA**	*FVD* ↓	*SSIM* ↑	*LPIPS* ↓	*mIoU* ↑	*aIoU* ↑
A	×	×	1378.92	0.78496	**0.3076**	0.252	0.505
B	✓	×	1227.56	0.79077	0.3101	0.295	**0.528**
C	×	✓	1119.71	0.79597	0.3246	0.277	0.516
D	✓	✓	**1055.25**	**0.79981**	0.3123	**0.302**	0.522

have experimentally validated that after changing those modules. The experimental results are shown in Table 3, where we feed features from different clip models (Event-trained or RGB-trained) to the SVD U-Net. Note that all CLIPs are fed with event voxels. Even further fine-tuning the SVD with plenty of data, the resulting diffusion model with Event-trained is still underperformed.

Testing-time Prompt Augmentation. As detailed in Sec. 4.1, we enhance the test-time prompt by incorporating multiple event frames of high temporal resolution. To thoroughly examine the impact of these hyperparameters and ascertain the efficacy of our test-time augmentation strategy, we compare various approaches employing different test-time prompts, as delineated in Table 6, titled 'Testing-time'. From this comparative analysis, it is evident that there is a gradual improvement in the neural network's performance as the extent of prompt augmentation increases.

5.3 Discussion

While the proposed method has exhibited preliminary capabilities, it is imperative to address its existing limitations. One significant challenge is the inherent nature of event data; while it boasts high temporal resolution, this type of data typically lacks texture, thereby impeding the effective capture of detailed semantic information. This limitation highlights the difficulty in accurately representing complex visual scenes solely based on event data. Furthermore, there is a pressing need for the development of a lossless representation technique that can fully preserve the unique high-temporal attributes of event data, ensuring no critical information is lost during processing. *Specific constrained scenarios are visualized and discussed in detail in Appendix Section D.*

To address these challenges and enhance the efficacy of the proposed method, future work should focus on several key areas. Firstly, advancing the method's ability to interpret texture would mark a significant improvement. This could involve integrating additional sensory inputs, e.g., RGB images, or employing more sophisticated data fusion techniques. Secondly, **the creation of a more comprehensive event sequence dataset is essential**. Such a dataset should encompass a wider variety of scenarios and conditions, thereby providing a robust platform for training and testing the improved models. By addressing these aspects, future research can pave the way for more accurate, reliable, and versatile event-based vision systems for object motion forcasting.

6 Conclusion

In this study, we have introduced the Event-Sequence Diffusion Network, a novel approach poised to redefine the landscape of video diffusion. By leveraging event-based data, characterized by its high temporal resolution, our methodology advances the frontiers of motion modeling, enabling the generation of video sequences that are not only rich in detail but also grounded in the realistic dynamics of object motion. Our approach stands in stark contrast to traditional video diffusion models that rely on single images or textual prompts for initialization. By employing sequences of events as the conditioning input, we ensure a more nuanced and temporally coherent synthesis of video content.

Future work will focus on refining the Event-Sequence Diffusion Network, exploring its applicability across a broader spectrum of computer vision tasks, and enhancing its efficiency for real-time applications. Moreover, we aim to delve deeper into the interplay between event-based data and diffusion models, seeking to unlock new potentials and applications in areas such as autonomous navigation, interactive gaming, and dynamic scene reconstruction.

References

[1] A. Blattmann, T. Dockhorn, S. Kulal, D. Mendelevitch, M. Kilian, D. Lorenz, Y. Levi, Z. English, V. Voleti, A. Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023.

[2] A. Blattmann, R. Rombach, H. Ling, T. Dockhorn, S. W. Kim, S. Fidler, and K. Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[3] C. Brandli, R. Berner, M. Yang, S.-C. Liu, and T. Delbruck. A 240$\times$ 180 130 db 3 μs latency global shutter spatiotemporal vision sensor. *IEEE Journal of Solid-State Circuits*, 49(10): 2333–2341, 2014.

[4] H. Cao, C. Tan, Z. Gao, Y. Xu, G. Chen, P.-A. Heng, and S. Z. Li. A survey on generative diffusion models. *IEEE Transactions on Knowledge and Data Engineering*, 2024.

[5] W. Chai, X. Guo, G. Wang, and Y. Lu. Stablevideo: Text-driven consistency-aware diffusion video editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23040–23050, 2023.

[6] W. Chen, J. Wu, P. Xie, H. Wu, J. Li, X. Xia, X. Xiao, and L. Lin. Control-a-video: Controllable text-to-video generation with diffusion models, 2023.

[7] Z. Chen, Z. Zhu, Y. Zhang, J. Hou, G. Shi, and J. Wu. Segment any events via weighted adaptation of pivotal tokens. *arXiv preprint arXiv:2312.16222*, 2023.

[8] F.-A. Croitoru, V. Hondru, R. T. Ionescu, and M. Shah. Diffusion models in vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[9] Z. Ding, R. Zhao, J. Zhang, T. Gao, R. Xiong, Z. Yu, and T. Huang. Spatio-temporal recurrent networks for event-based optical flow estimation. In *Proceedings of the AAAI conference on artificial intelligence*, pages 525–533, 2022.

[10] P. Esser, J. Chiu, P. Atighehchian, J. Granskog, and A. Germanidis. Structure and content-guided video synthesis with diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7346–7356, 2023.

[11] S. Ettinger, S. Cheng, B. Caine, C. Liu, H. Zhao, S. Pradhan, Y. Chai, B. Sapp, C. R. Qi, Y. Zhou, et al. Large scale interactive motion forecasting for autonomous driving: The waymo open motion dataset. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9710–9719, 2021.

[12] G. Gallego, T. Delbrück, G. Orchard, C. Bartolozzi, B. Taba, A. Censi, S. Leutenegger, A. J. Davison, J. Conradt, K. Daniilidis, et al. Event-based vision: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 44(1):154–180, 2020.

[13] Z. Gao, C. Tan, L. Wu, and S. Z. Li. Simvp: Simpler yet better video prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3170–3180, June 2022.

[14] Z. Gao, C. Tan, L. Wu, and S. Z. Li. Simvp: Simpler yet better video prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3170–3180, 2022.

[15] S. Ge, S. Nah, G. Liu, T. Poon, A. Tao, B. Catanzaro, D. Jacobs, J.-B. Huang, M.-Y. Liu, and Y. Balaji. Preserve your own correlation: A noise prior for video diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22930–22941, 2023.

[16] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

[17] V. L. Guen and N. Thome. Disentangling physical dynamics from unknown factors for unsupervised video prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[18] J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[19] J. Ho, W. Chan, C. Saharia, J. Whang, R. Gao, A. Gritsenko, D. P. Kingma, B. Poole, M. Norouzi, D. J. Fleet, et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022.

[20] J. Ho, C. Saharia, W. Chan, D. J. Fleet, M. Norouzi, and T. Salimans. Cascaded diffusion models for high fidelity image generation. *The Journal of Machine Learning Research*, 23(1): 2249–2281, 2022.

[21] J. Hou, Z. Zhu, J. Hou, H. Liu, H. Zeng, and H. Yuan. Global structure-aware diffusion process for low-light image enhancement. *Advances in Neural Information Processing Systems*, 36, 2024.

[22] L. Hu, R. Zhao, Z. Ding, L. Ma, B. Shi, R. Xiong, and T. Huang. Optical flow estimation for spiking camera. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17844–17853, 2022.

[23] Y. Jiang, Y. Wang, S. Li, Y. Zhang, M. Zhao, and Y. Gao. Event-based low-illumination image enhancement. *IEEE Transactions on Multimedia*, 2023.

[24] T. Karras, M. Aittala, T. Aila, and S. Laine. Elucidating the design space of diffusion-based generative models. *Advances in Neural Information Processing Systems*, 35:26565–26577, 2022.

[25] D. Kingma, T. Salimans, B. Poole, and J. Ho. Variational diffusion models. *Advances in neural information processing systems*, 34:21696–21707, 2021.

[26] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollár, and R. Girshick. Segment anything. *arXiv:2304.02643*, 2023.

[27] M. Liang, B. Yang, R. Hu, Y. Chen, R. Liao, S. Feng, and R. Urtasun. Learning lane graph representations for motion forecasting. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 541–556. Springer, 2020.

[28] P. Lichtsteiner, C. Posch, and T. Delbruck. A 128$\times$ 128 120 db 15 μs latency asynchronous temporal contrast vision sensor. *IEEE Journal of Solid-State Circuits*, 43(2):566–576, 2008. doi: 10.1109/JSSC.2007.914337.

[29] X. Lin, J. He, Z. Chen, Z. Lyu, B. Fei, B. Dai, W. Ouyang, Y. Qiao, and C. Dong. Diffbir: Towards blind image restoration with generative diffusion prior. *arXiv preprint arXiv:2308.15070*, 2023.

[30] F. Mahlknecht, D. Gehrig, J. Nash, F. M. Rockenbauer, B. Morrell, J. Delaune, and D. Scaramuzza. Exploring event camera-based odometry for planetary robots. *IEEE Robotics and Automation Letters*, 7(4):8651–8658, 2022.

[31] A. I. Maqueda, A. Loquercio, G. Gallego, N. García, and D. Scaramuzza. Event-based vision meets deep learning on steering prediction for self-driving cars. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5419–5427, 2018. doi: 10.1109/CVPR.2018.00568.

[32] K. Mei and V. Patel. Vidm: Video implicit diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 9117–9125, 2023.

[33] N. Nayakanti, R. Al-Rfou, A. Zhou, K. Goel, K. S. Refaat, and B. Sapp. Wayformer: Motion forecasting via simple & efficient attention networks. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2980–2987. IEEE, 2023.

[34] A. Nichol, P. Dhariwal, A. Ramesh, P. Shyam, P. Mishkin, B. McGrew, I. Sutskever, and M. Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021.

[35] J. Pan, H. Yan, J. H. Liew, J. Feng, and V. Y. F. Tan. Towards accurate guided diffusion sampling through symplectic adjoint method, 2023.

[36] C. Posch, D. Matolin, and R. Wohlgenannt. A qvga 143 db dynamic range frame-free pwm image sensor with lossless pixel-level video compression and time-domain cds. *IEEE Journal of Solid-State Circuits*, 46(1):259–275, 2011. doi: 10.1109/JSSC.2010.2085952.

[37] H. Rebecq, R. Ranftl, V. Koltun, and D. Scaramuzza. High speed and high dynamic range video with an event camera. *IEEE transactions on pattern analysis and machine intelligence*, 43(6): 1964–1980, 2019.

[38] H. Rebecq, R. Ranftl, V. Koltun, and D. Scaramuzza. High speed and high dynamic range video with an event camera. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(6): 1964–1980, 2021. doi: 10.1109/TPAMI.2019.2963386.

[39] L. Ruan, Y. Ma, H. Yang, H. He, B. Liu, J. Fu, N. J. Yuan, Q. Jin, and B. Guo. Mm-diffusion: Learning multi-modal diffusion models for joint audio and video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10219–10228, 2023.

[40] S. Saxena, C. Herrmann, J. Hur, A. Kar, M. Norouzi, D. Sun, and D. J. Fleet. The surprising effectiveness of diffusion models for optical flow and monocular depth estimation. *Advances in Neural Information Processing Systems*, 36, 2024.

[41] J. Schulman, S. Levine, P. Abbeel, M. Jordan, and P. Moritz. Trust region policy optimization. In *International conference on machine learning*, pages 1889–1897. PMLR, 2015.

[42] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

[43] R. Shaw, S. Catley-Chandar, A. Leonardis, and E. Pérez-Pellitero. Hdr reconstruction from bracketed exposures and events. *arXiv preprint arXiv:2203.14825*, 2022.

[44] S. Shiba, Y. Aoki, and G. Gallego. Secrets of event-based optical flow. In *European Conference on Computer Vision*, pages 628–645. Springer, 2022.

[45] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. PMLR, 2015.

[46] J. Song, C. Meng, and S. Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

[47] Y. Song and S. Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32, 2019.

[48] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

[49] Y. Song, P. Dhariwal, M. Chen, and I. Sutskever. Consistency models. *arXiv preprint arXiv:2303.01469*, 2023.

[50] C. Tan, Z. Gao, L. Wu, Y. Xu, J. Xia, S. Li, and S. Z. Li. Temporal attention unit: Towards efficient spatiotemporal predictive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18770–18782, June 2023.

[51] C. Tan, Z. Gao, L. Wu, Y. Xu, J. Xia, S. Li, and S. Z. Li. Temporal attention unit: Towards efficient spatiotemporal predictive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18770–18782, 2023.

[52] Thomas, S. van Steenkiste, K. Kurach, R. Marinier, M. Michalski, and S. Gelly. Fvd: A new metric for video generation. *CoRR*, 2019.

[53] S. Tulyakov, D. Gehrig, S. Georgoulis, J. Erbach, M. Gehrig, Y. Li, and D. Scaramuzza. Time lens: Event-based video frame interpolation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16155–16164, 2021.

[54] S. Tulyakov, A. Bochicchio, D. Gehrig, S. Georgoulis, Y. Li, and D. Scaramuzza. Time lens++: Event-based frame interpolation with parametric non-linear flow and multi-scale fusion. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[55] X. Wang, J. Li, L. Zhu, Z. Zhang, Z. Chen, X. Li, Y. Wang, Y. Tian, and F. Wu. Visevent: Reliable object tracking via collaboration of frame and event flows, 2023.

[56] X. Wang, S. Wang, C. Tang, L. Zhu, B. Jiang, Y. Tian, and J. Tang. Event stream-based visual object tracking: A high-resolution benchmark dataset and a novel baseline. *arXiv preprint arXiv:2309.14611*, 2023.

[57] X. Wang, J. Li, L. Zhu, Z. Zhang, Z. Chen, X. Li, Y. Wang, Y. Tian, and F. Wu. Visevent: Reliable object tracking via collaboration of frame and event flows. *IEEE Transactions on Cybernetics*, 54(3):1997–2010, 2024. doi: 10.1109/TCYB.2023.3318601.

[58] Y. Wang, M. Long, J. Wang, Z. Gao, and P. S. Yu. Predrnn: Recurrent neural networks for predictive learning using spatiotemporal lstms. *Advances in neural information processing systems*, 30, 2017.

[59] Y. Wang, H. Wu, J. Zhang, Z. Gao, J. Wang, P. S. Yu, and M. Long. Predrnn: A recurrent neural network for spatiotemporal predictive learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(2):2208–2225, 2023. doi: 10.1109/TPAMI.2022.3165153.

[60] Y. Xu, L. Chambon, M. Chen, A. Alahi, M. Cord, P. Perez, et al. Towards motion forecasting with real-world perception inputs: Are end-to-end approaches competitive? In *International Conference on Robotics and Automation (ICRA)*, 2024.

[61] L. Yang, Z. Zhang, Y. Song, S. Hong, R. Xu, Y. Zhao, W. Zhang, B. Cui, and M.-H. Yang. Diffusion models: A comprehensive survey of methods and applications. *ACM Computing Surveys*, 56(4):1–39, 2023.

[62] B. Ye, H. Chang, B. Ma, S. Shan, and X. Chen. Joint feature learning and relation modeling for tracking: A one-stream framework. In *ECCV*, 2022.

[63] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018.

[64] Y. Zhang, E. Tzeng, Y. Du, and D. Kislyuk. Large-scale reinforcement learning for diffusion models. *arXiv preprint arXiv:2401.12244*, 2024.

[65] J. Zhao, S. Ji, Z. Cai, Y. Zeng, and Y. Wang. Moving object detection and tracking by event frame from neuromorphic vision sensors. *Biomimetics*, 7(1):31, 2022.

[66] A. Z. Zhu, L. Yuan, K. Chaney, and K. Daniilidis. Ev-flownet: Self-supervised optical flow estimation for event-based cameras. *arXiv preprint arXiv:1802.06898*, 2018.

[67] Z. Zhu, J. Hou, and X. Lyu. Learning graph-embedded key-event back-tracing for object tracking in event clouds. *Advances in Neural Information Processing Systems*, 35:7462–7476, 2022.

[68] Z. Zhu, J. Hou, and D. O. Wu. Cross-modal orthogonal high-rank augmentation for rgb-event transformer-trackers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22045–22055, 2023.

[69] Y. Zou, Y. Zheng, T. Takatani, and Y. Fu. Learning To Reconstruct High Speed and High Dynamic Range Videos From Events. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2024–2033, 2021.

Appendix Overview

In this Appendix material, we provide additional details to complement the content of the paper, including the training and inference details of our method(**Sec.A**), an analysis of the results from comparative methods, more experiments on downstream tasks(**Sec.B**), details related to reinforcement learning(**Sec.C**), and limitation of proposed method(**Sec.D**).

A Training and Inference Details

We provide a comprehensive training pipeline along with detailed parameter specifications, encompassing the model architecture and fine-tuning parameters. For voxelized event data, we perform preprocessing and data augmentation to enhance the performance and generalization capability of the model.

A.1 Architecture

As discussed in Section. 4.1, We employ the temporal U-Net architecture based on *Stable Video Diffusion*. During training, we preprocess the event sequence into multiple frames of voxel data, which serve as conditional inputs to the network. Fig. 6 illustrates the overall training procedure, which bears resemblance to SVD. The initial segment of the event sequence serves as the conditional input to the U-Net model. For the parameters trained in the U-Net, we only fine-tuned the cross-attention layer that incorporates the event condition. We conducted ablation experiments on different attention layers, as shown in Table 3. Simultaneously fine-tuning both spatial and temporal attention layers yielded the best results.

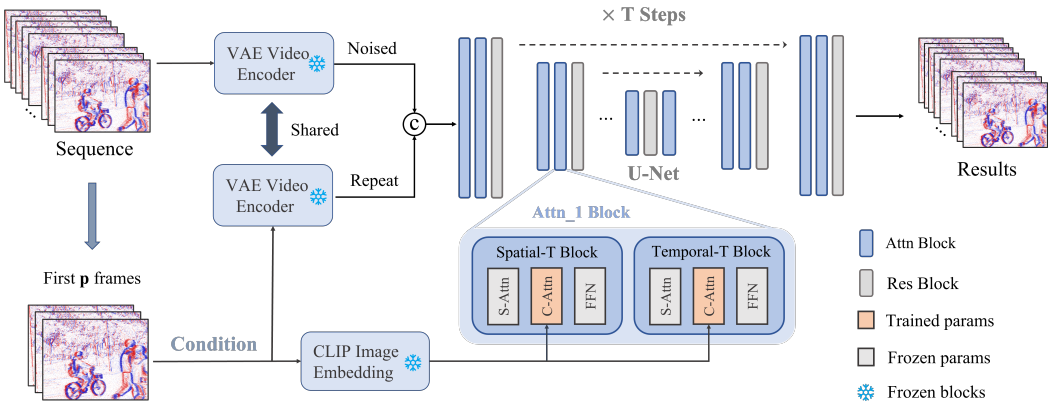


Figure 6: Training workflow of the proposed method, where the left upper one indicates the target noised latent and the left lower one represents the prompted event sequence. We concatenate the prompt information with denoising latent for noise learning. Moreover, the feature from the CLIP model is also injected into the diffusion U-Net.

A.2 Datasets and augmentation

During the pre-training phase of the proposed diffusion model, events of poor quality can adversely affect model training. Therefore, we emulate the approach taken in EventSAM [7] by filtering out sequences with severe degradation in VisEvent. Finally, we integrated 291 event sequences from the VisEvent dataset and 841 sequences from the EventVOT dataset. We omitted specific sequences from the CeleX-V sensors due to their absence of polarity information. For the alignment phase, we uniformly sourced all data, the same as the training dataset, aligning them in accordance with the pre-training prompts. In the evaluation stage, a random selection of 100 sequences from each dataset was employed to rigorously assess the efficacy of our proposed methodology.

During training, we randomly crop the Visevent sequences using a 256×256 kernel and then resize them to 128×128 before feeding them into the network. For the higher-resolution EventVOT dataset, in order to preserve as much motion information as possible, we randomly select crop kernels ranging

from 512 to 720, crop the images accordingly, and then resize them to 128×128 as well. We divide the events into intervals of 20ms.

After events are converted into voxels, normalization is required. Simple normalization in the range of 0 to 1 may lead to flickering in the sequence. Therefore, when obtaining the sequence, we perform a smoother normalization using maximum value normalization. Specifically, dividing each frame of the sequence by the maximum absolute value of the sequence ensures that the voxel values are between 0 and 1 and remain stable.

B Downstream tasks

To validate the reliability of the generated event voxels, we assessed our approach's performance on several classic downstream tasks such as object tracking, segmentation, and optical flow estimation.

B.1 Evaluation Methods and Metrics

For the object tracking task, we utilized the OSTrack [62] library to evaluate the event voxels estimated by our method and the RGB frames estimated by SVD [1], focusing on the selected VisEvent test set to ensure alignment. Due to the significant modality differences between event data and RGB data, directly comparing tracking metrics would be unfair. Therefore, we tested on the ground truth of the selected test sequences separately for each modality, using these as performance upper bounds. The model performance was assessed by analyzing the discrepancies between the obtained results and these upper bounds. We utilized AUC (Area Under the Curve),OP50, OP75 and PR(Precision and Recall) as our base evaluation metrics,upon which we calculated the mean ratio of each metric to its upper bound, obtaining a cross-modality tracking evaluation metric R_{upper}:

$$R_{upper} = Mean(\frac{AUC}{AUC_{upper}} + \frac{OP50}{OP50_{upper}} + \frac{OP75}{OP75_{upper}} + \frac{PR}{PR_{upper}}) \tag{7}$$

For the object segmentation task, we employed SAM [26] to segment objects in the RGB domain and fine-tuned EventSAM [7] on event sequences for object segmentation in the event domain. Given the absence of segmentation ground truth in the selected dataset and the strong generalization capabilities of the segmentation models in their respective domains, we used the sequence ground truth segmentation results as the ground truth for our evaluations. we employ mIOU (mean Intersection over Union) and aIOU (average Intersection over Union) as evaluation metrics.

Optical flow estimation is particularly challenging due to the limited and narrowly distributed datasets with optical flow ground truth. Although event-based optical flow estimation has been explored, it often suffers from poor generalization performance. Consequently, we restricted our qualitative experiments to simple motion scenarios involving a single object to better understand the model's performance under controlled conditions. The DDVM [40] was employed to estimate the optical flow for both event and RGB sequences.

B.2 Results Analysis

The quantitative results of the tracking are presented in Table 2. Due to the modality differences, the overall tracking performance of RGB surpasses that of event tracking, as evident from their respective ground truth sequences. However, our method exhibits a higher ratio of upper bounds compared to SVD, indicating that our estimated results are closer to the ground truth, making the motion estimation more reliable.

Fig. 5 and Fig. 7 depict the visual results of object tracking. In Fig. 5c, due to erroneous motion trajectory prediction by SVD, a significant discrepancy between tracking and labeling is observed. Conversely, our method demonstrates better trajectory estimation, attributed to the high temporal resolution of event prompts providing prior information on motion during denoising.

Table 1 presents the quantitative results of object segmentation, where our method achieves the best performance. Additional visual results are provided in Fig. 10, illustrating that the event segmentation generated by our method closely resembles the ground truth.

Optical flow is typically employed to describe object motion. In our work, the emphasis of optical flow estimation lies in the accuracy of motion prediction. Fig. 5 and Fig. 7 illustrate that our method can predict motion more precisely, leading to better optical flow estimation results.

C Reinforcement learning for SVD

C.1 Event voxel normalization based on standard deviation.

In reinforcement learning, the choice of the reward function is a crucial factor in determining training outcomes. We selected the FVD and SSIM metrics, which are closest to perceptual results, as our reward functions.Specifically, the final reward function is defined as $R(x) = SSIM(x) + \lambda(FVD(x))$, where λ equals 2. Upon each model update, reward scores of the data generated from the previous batch are normalized to a standard normal distribution.

Further, To quantitatively demonstrate the rationale behind our selection of these two metrics, we actually have experimentally tried to utilize a mixture of all metrics to model the reconstruction reward. However, the performance results are much worse than the method we ultimately used, as shown in Table. 5. It may be due to that sometimes the optimization of pixel-level metrics of MSE and PSNR may be contradictory to perceptual metrics. Thus, such a mixture of metrics makes the objective hard to optimize. Meanwhile, due to the target data distribution, it's more plausible to optimize the process on the data manifold (perceptual space) than the raw space.

Table 5: Ablation Study of reward metrics

Reward Metrics	*MSE* ↓	*PSNR* ↑	*FVD* ↓	*FID* ↓	*SSIM* ↑	*LPIPS* ↓
mixture metrics	0.0240	16.198	1562.66	265.32	0.6674	0.3463
FVD & SSIM	0.0170	17.696	1055.25	243.45	0.7998	0.3123

Directly utilizing spatial metrics as evaluation criteria in the RGB domain is typically feasible because RGB data contains dense spatial information, and the distribution of motion and static scenes is relatively similar. However, this approach is not viable for the event domain, which only records motion information. In Fig. 8b, the samples illustrate this point. As depicted in the rightmost column, when capturing a static scene, the event camera does not generate events, leading to a high original reward score. However, the samples in the first two columns occur in scenarios with significant motion, resulting in relatively high-quality, natural outcomes. Nevertheless, when normalized alongside the results from the first column, even these samples yield negative scores.

Fig. 8a illustrates the distribution of reward scores with respect to sample standard deviation. Typically, higher standard deviations in samples correspond to lower scores, indicating poorer performance. Conversely, in relatively static scenes, higher scores are typically obtained. Such outcomes may inadvertently induce an overall trend toward static motion, which is not desirable. As illustrated by the green curve in Fig. 8a, after several model parameter updates, the reward score begins to decline, ultimately resulting in unsatisfactory outcomes.

To address this issue, we perform normalization of samples based on their standard deviation each time parameters are updated. Specifically, the standardized $SCORE_{std}$ denoted as $SCORE_{std} = SCORE + \beta\,(std(x) - std_{min})$, where std_{min} represents the minimum value of the total standard deviation of the previous batch of generated samples and β equals 30.

C.2 Training Setting and Results Analysis

During the training process, we initialize the training with the pre-trained model obtained from Section 4.1. The batch size for training is set to 64, and the model is updated every 100 iterations, simultaneously updating the sample pool. Gaussian normalization and standard deviation normalization are sequentially applied to the sample pool. After standardizing the samples, the training process returns to a positive trajectory. The purple curve in Fig. 8a represents the score curve during training, with convergence observed around the 1000th iteration.

Fig. 11 illustrates the visual results before and after reinforcement learning. Due to the inherent unpredictability and randomness in the generation of results by diffusion, utilizing only the results

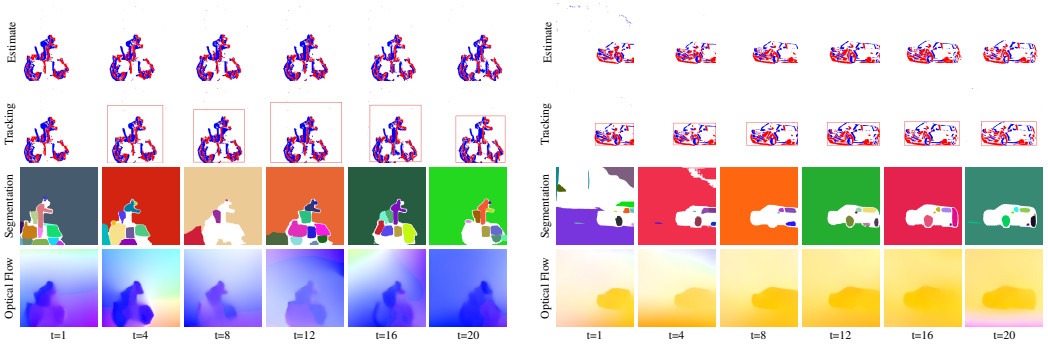
Algorithm 2 Multi Prompt Reverse Process

$\mathbf{X}_T^{pm} \sim \mathcal{N}(0, \mathbf{I})$,$\mathbf{X}_T^{rc} \sim \mathcal{N}(0, \mathbf{I})$
For $t = T, ..., 1$ **do**
  if $t \geq \tau$ **then**
  $\mathbf{X}_{t-1}^{pm} \leftarrow \sqrt{\bar{\alpha}_t}\mathbf{X}_0^{pm} + \sqrt{1-\bar{\alpha}_t}\mathbf{X}_T^{pm}$.
  $\mathbf{X}_t^{rep} \leftarrow$ **replace** $(\mathbf{X}_t^{pm}, \mathbf{X}_t^{rc})$.
  Sampling $\mathbf{X}_{t-1}^{rc}$ from $\mathbf{X}_t^{rep}$ via Eq. 3.
  else Sampling $\mathbf{X}_{t-1}^{rc}$ from $\mathbf{X}_t^{rc}$ via Eq. 3.
End for
Return $\mathbf{X}_0$

Table 6: Ablation study of different number of testing-time prompts. Here we only compare with the nearest 10 future event voxels. All experimental settings share the same model weights.

Metric	1	2	4	8	12	15
MSE ↓	0.0177	0.0195	0.0170	0.0196	0.0178	0.0182
FVD ↓	1170.48	1196.43	1055.25	1137.69	1142.15	1188.01
FID ↓	242.318	247.803	243.451	243.079	238.369	239.297
SSIM ↑	0.79349	0.78988	0.79981	0.80718	0.82553	0.83545
LPIPS ↓	0.3272	0.3228	0.3123	0.3228	0.3189	0.3308
mIoU ↑	0.287	0.296	0.302	0.302	0.309	0.302
aIoU ↑	0.518	0.527	0.522	0.529	0.525	0.518

generated by the pre-trained model (Fig. 11b) often leads to uncontrollable distortions and deformations. However, the results after reinforcement learning (Fig. 11c) tend to be more stable, with motion trends closely resembling real-world scenarios.



(a) A tank moving towards the right.

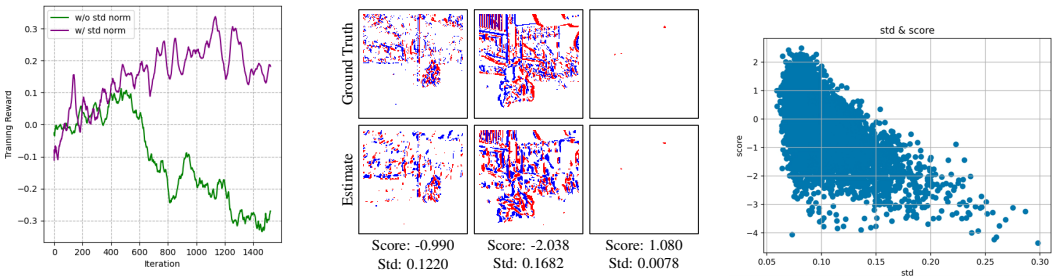
(b) A car moving towards the left.

Figure 7: More results of downstream tasks.

D Limitation

D.1 Limitation scenarios

Although our method achieves good results in most scenarios, it is still significantly limited in the following cases due to the sampling characteristics of event data: (1) **Complex background scenarios.** Event cameras may capture incomplete textures in certain situations, leading to poorer prediction performance, as shown in Fig. 14a. Complex backgrounds can reduce the clarity of the target object, resulting in worse outcomes, especially in cases where the camera lens is shaking. (2) **Heavily overlapped object scenarios.** When objects overlap, their motion becomes quite complex, and due to the edge-focused characteristics of event cameras, understanding such motion is



(a) (b) (c)

Figure 8: Analysis of the Reinforcement Learning Process. Fig. 8a illustrates the reward curves during training, with the purple curve representing the scenario with standard deviation normalization applied, and the green curve representing the scenario without it.Fig. 8b displays partial visualization results during the training process, where the first row represents the ground truth, and the second row depicts the results estimated by the pre-trained model. Fig. 8c illustrates the distribution of reward scores with respect to standard deviation normalization for all training samples.

Table 7: Comparison of methods in terms of parameters and FLOPs.

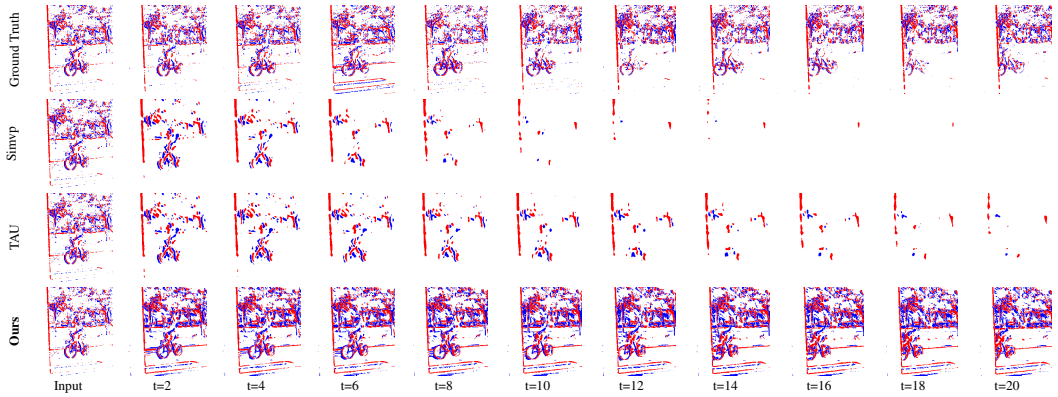
Methods	**Params (M)**	**FLOPs (G)**
PredRNNv2	23.9	48.92
SimVP	58.0	60.61
TAU	44.7	92.50
ours	1521.0	693.92

challenging, as shown in Fig. 14b and Fig. 14c. When people overlap, their footsteps often become chaotic, leading to less accurate predictions.

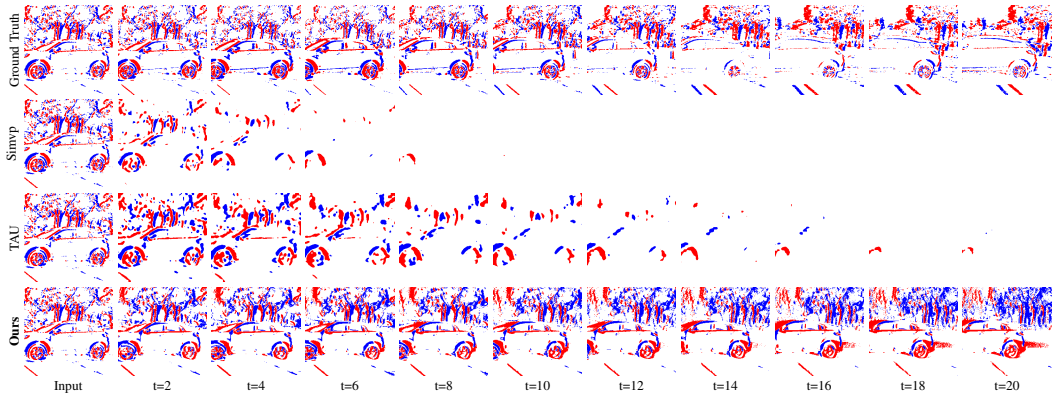
D.2 Computational Resource & Scalability

The following Table. 7 compares the computational resources of our method with SOTA methods. The powerful generative capability and high fidelity of diffusion models lead to the cost of substantial computational resource consumption. As shown in the following table, our parameter count and FLOPs significantly exceed those of traditional models. However, we believe this trade-off is necessary because of the powerful learning capability of large models in the real world. Taking the future motion estimation task as an example, our diffusion-based method significantly surpasses traditional methods in understanding and learning motion.

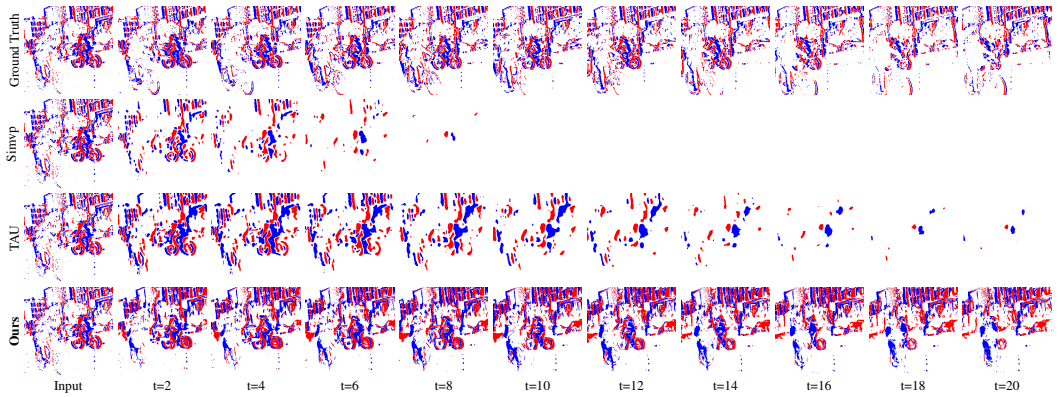
For a salable model size for different inference environments, there indeed are many works indicating that the diffusion model can be applied to quantization or other acceleration techniques for speeding up the inference process. We believe that with the advancement of hardware and acceleration techniques, the inference speed of diffusion models will be significantly improved in the near future.



(a) A bicycle riding into an obstacle.



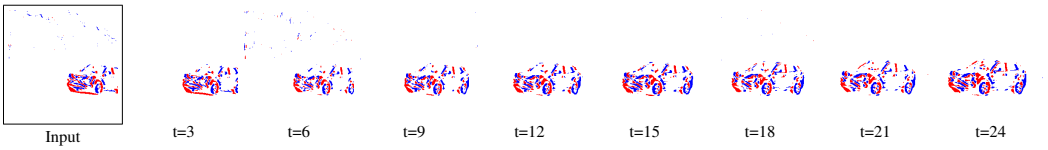
(b) A car driving to the left, with the camera lens rotating to follow.



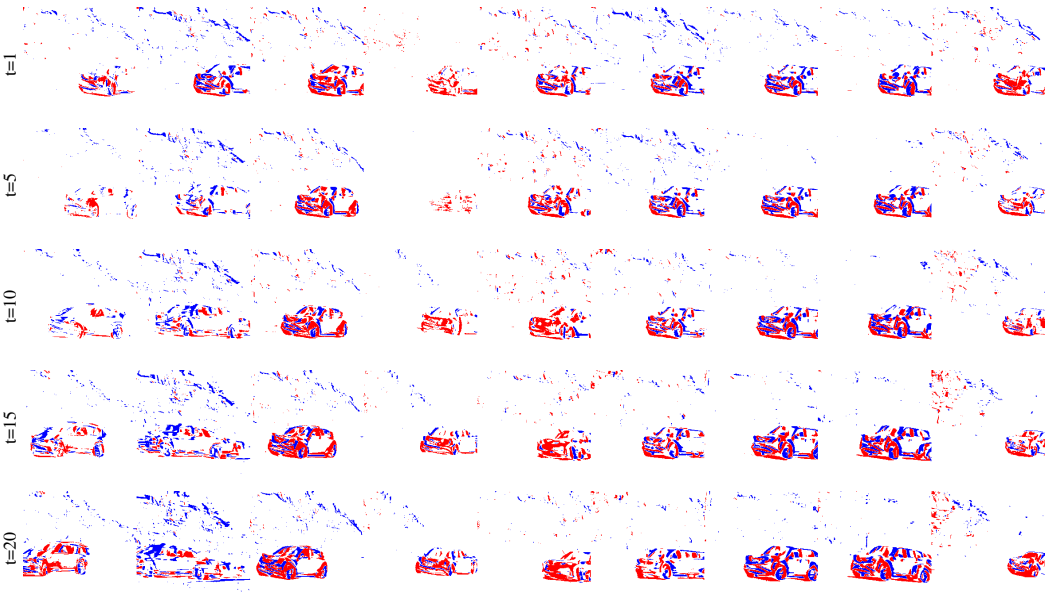
(c) Two bicycles traveling towards each other and intersecting.

Figure 9: Qualitative comparison between SOTA methods. The first row of each sequence represents the ground truth of the event sequence. The second and third rows respectively depict the results of future event estimation by SimVP [14] and TAU [50]. The final row represents the results obtained by our method.

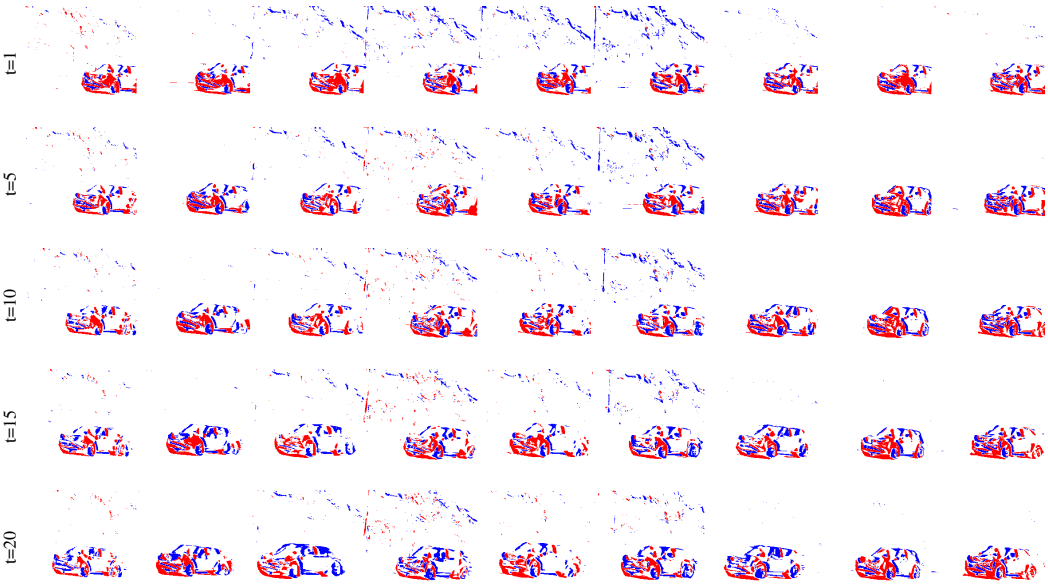
Figure 10: Additional results of Segmentation. The first row of each sequence represents the ground truth of the events, and the second row shows the segmentation results. The last two rows respectively display the results estimated by our method and the corresponding segmentation results.



(a) Ground Truth of the Car Sequence

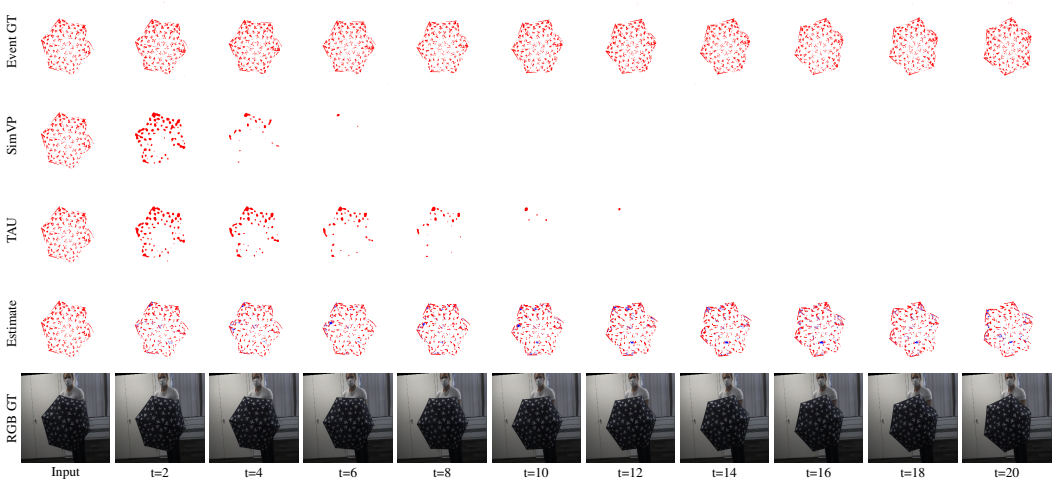


(b) Visualization of the Pre-train model's Estimation Results(without Motion Alignment)

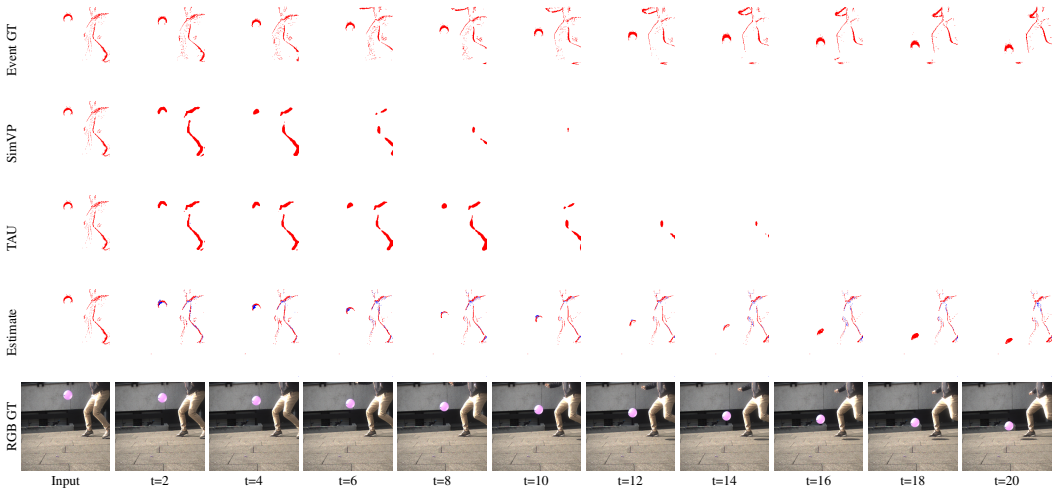


(c) Visualization of the Results after Motion Alignment

Figure 11: Visualization results of Motion Alignment. Fig. 11a shows the ground truth of the car sequence; Fig. 11b presents the estimated results using a pre-trained model with 9 different random seeds, where several instances resulted in failures; Fig.11c illustrates the motion alignment results generated by applying reinforcement learning, yielding more stable estimation.

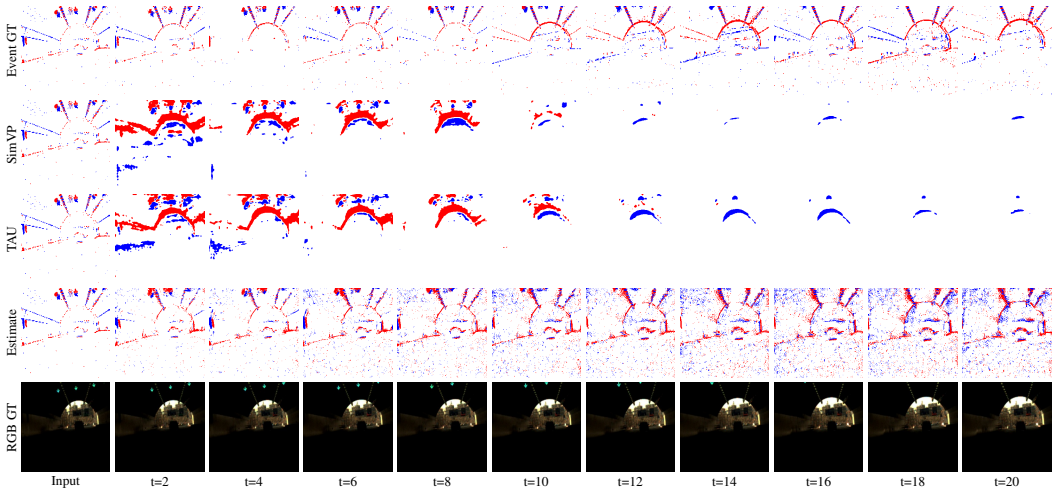


(a) Rotating Umbrella.

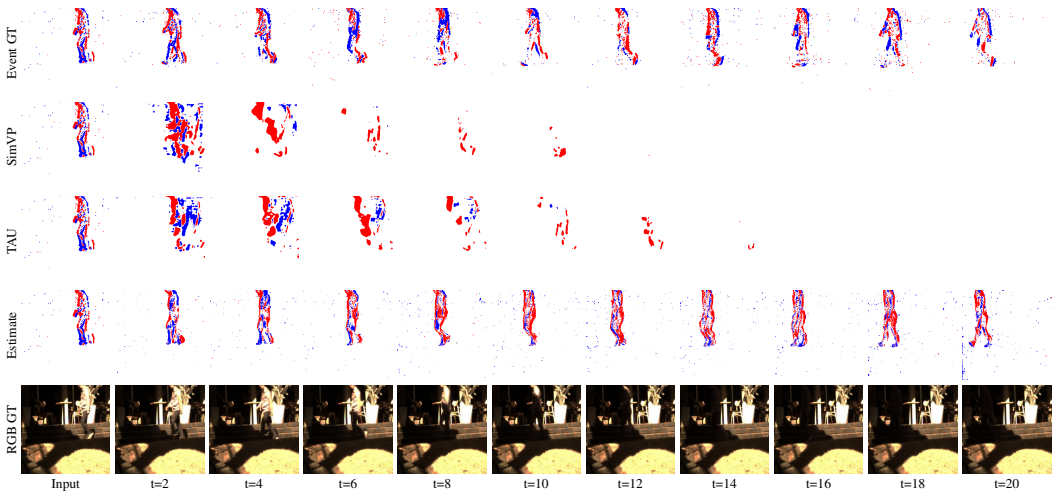


(b) Falling Balloon.

Figure 12: Visualization of our method's prediction in hs-ergb dataset.

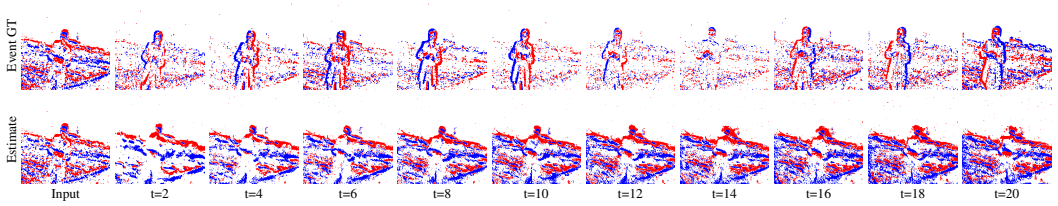


(a) Poor exposure: The car driving out of the tunnel.

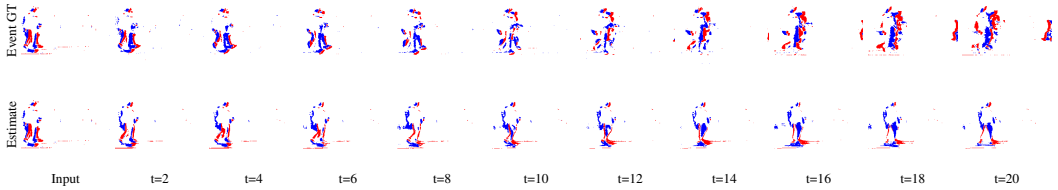


(b) Occlusion: A person walking from a bright place to shadow.

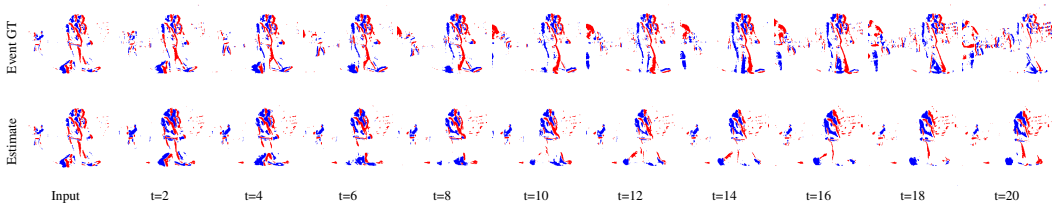
Figure 13: Visualization of our method's prediction in various scenarios.



(a) Shaky camera capturing a stationary person.



(b) A person running out behind another person.



(c) Two people walking side by side.

Figure 14: Visualization of our method's prediction in severely degraded scenarios.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The claims indeed reflect the paper's contributions and scope.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We have described the limitations of our work in Sec 6.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

Justification: No theoretical results

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provide detailed settings and parameters for the experimental training and testing.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We provide the source code and information on how to obtain the datasets used in the Appendix materials.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We provide all experimental settings and details in Sec. A.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [NA]

   Justification: We fixed the random seed to reproduce the experimental results.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).

- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We provide the experimental platform details in Sec. 5.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: We meticulously reviewed the ethical guidelines.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We discuss the potential societal impacts in Sec. 1 and Sec. 6.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: We do not have such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We have provided proper attribution for every asset used in the paper.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We have included the source code in the Appendix material. And we will also release the Pre-train model.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: We do not engage in these subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing or research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.