ControlEvents: Controllable Synthesis of Event Camera Data with Foundational Prior from Image Diffusion Models

Yixuan Hu[1,*] Yuxuan Xue[2,3,*,†] Simon Klenk[1,5]
Daniel Cremers[1,5] Gerard Pons-Moll[2,3,4]

[1] Technical University of Munich [2] University of Tübingen [3] Tübingen AI Center
[4] Max Planck Institute for Informatics, Saarland Informatics Campus
[5] Munich Center for Machine Learning (MCML) [*] co-first author [†] corresponding author

https://yuxuan-xue.com/controlevents

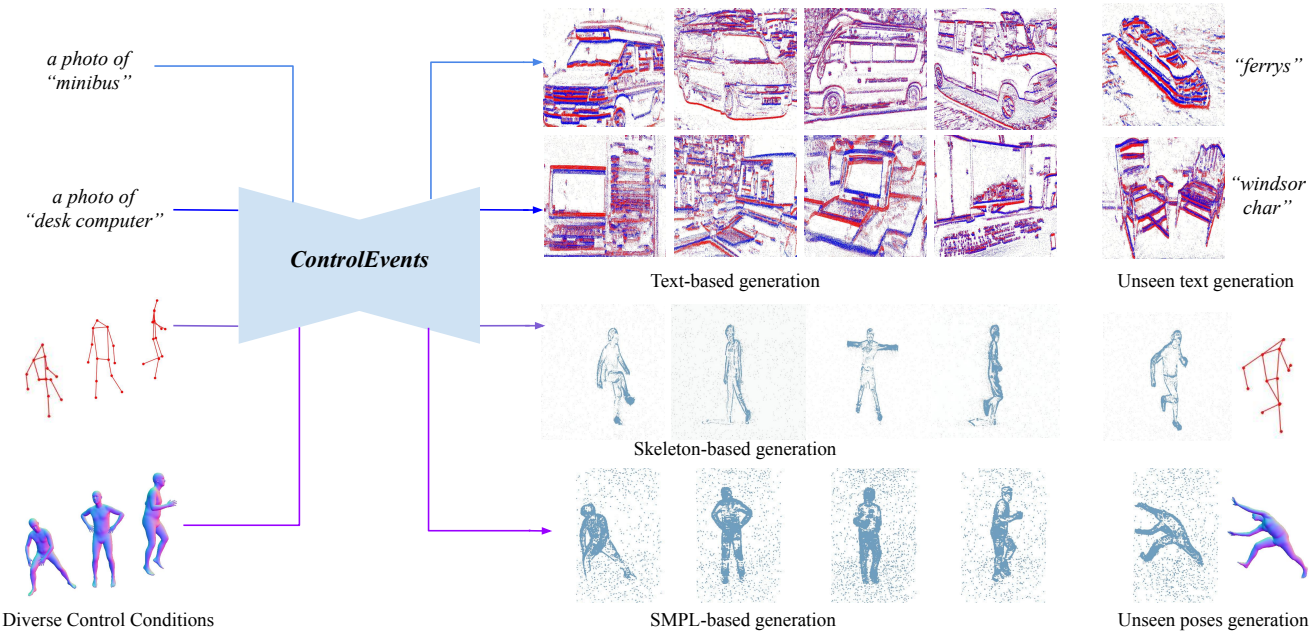


Figure 1. ***ControlEvents*** can synthesize realistic event camera data from diverse conditions, such as class text label, 2D human skeleton, 3D body poses. *ControlEvents* can generate large-scale realistic event data with pseudo ground-truth labels, which can enhance the deep learning model performance.

Abstract

In recent years, event cameras have gained significant attention due to their bio-inspired properties, such as high temporal resolution and high dynamic range. However, obtaining large-scale labeled ground-truth data for event-based vision tasks remains challenging and costly, which hinders the development of the event-based algorithm. In this paper, we present ***ControlEvents****, a diffusion-based generative model designed to synthesize unlimited high-quality event data guided by diverse control signals such as class text labels, 2D skeletons, and 3D body poses. Our key insight is to leverage the diffusion prior from foundation models, such as Stable Diffusion, enabling high-quality event data generation with minimal fine-tuning and limited labeled data. Our method streamlines the data generation process and significantly reduces the cost of producing labeled event datasets. We demonstrate the effectiveness of our approach by synthesizing event data for visual recognition, 2D skeleton estimation, and 3D body pose estimation. Our experiments show that the synthesized labeled event data enhances model performance in all tasks. Additionally, our approach can generate events based on unseen text labels during training, illustrating the powerful text-based generation capabilities inherited from foundation models. Our models and generated datasets is publicly available for*

future research.

1. Introduction

Visual recognition and body pose estimation in challenging scenarios, such as fast-motion and low-light conditions, are critical tasks in the field of computer vision. These applications are relevant in diverse fields, including autonomous driving, robotics, and human-computer interaction. However, conventional frame-based cameras often struggle in these environments due to limitations in temporal resolution and dynamic range, resulting in motion blur and reduced image quality which hinder performance.

In recent years, event cameras [24] have emerged as a promising alternative to frame-based cameras. Unlike traditional cameras, event cameras capture pixel-level changes in intensity rather than static frames, which enables them to operate effectively in high-speed and high-dynamic-range scenes [13]. These features make event cameras ideal for capturing visual data in challenging conditions.

Despite the advantages of event cameras, progress toward large-scale deep learning for event-based vision remains constrained by the scarcity of labeled ground-truth. Data annotation is time-intensive and costly, impeding robust model development. Existing simulators, such as ESIM [39] for rigid motion; EventHands [42] and SMPL-ESIM [49, 50] for human movement, help but fail to capture key sensor idiosyncrasies—such as background activity noise and spatially varying brightness thresholds—leading to a pronounced domain gap between simulated and real event streams (see Fig. 2). They also suffer from slow simulation, heavy reliance on curated 3D assets, and limited labeling flexibility. Moreover, most pipelines are constrained to paired RGB images or videos as inputs and cannot generate events under diverse conditioning modalities (e.g., class, text labels), which further restricts their utility for synthesizing labeled data for event-based visual recognition.

To address this gap, we introduce *ControlEvents*, a diffusion-based framework which can synthesize high-quality event data guided by diverse control signals. *ControlEvents* can generate labeled event data conditioned on various inputs, such as class text labels [21, 33], 2D skeletons [9, 44], and 3D body poses [62]. To address the challenge of limited labeled event data, *ControlEvents* leverages the diffusion prior and generation capabilities of visual foundation models, such as Stable Diffusion [41] or ControlNet [57]. By fine-tuning pre-trained foundation models, our method can synthesize realistic, high-quality event data with minimal training and limited labeled data. By learning directly on real captured events, our model can generate realistic noisy event data, making it more applicable for training machine learning models that generalize well to real-world data. Compared to traditional event simulators [39, 42, 49, 50], *ControlEvents* simplifies data generation and significantly reduces the costs of creating large-scale, realistic, and labeled event datasets. In summary, our contributions are as follows:

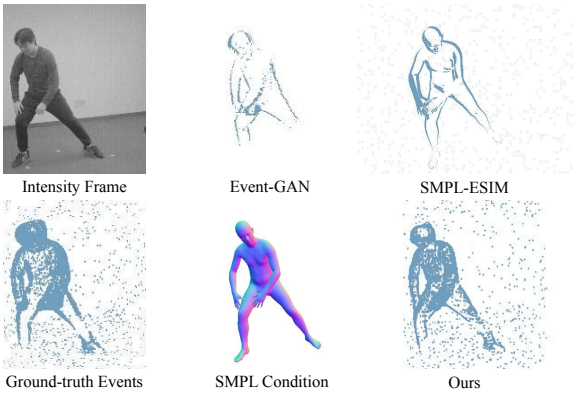


Figure 2. Comparison with Evant-GAN [60] and SMPL-ESIM [50] on 3D human-based events generation tasks. Our ***ControlEvents*** can effectively synthesize both events and noise, minimizing domain gap and closely matching the characteristics of real event data.

- We propose *ControlEvents*, a diffusion-based generative model that synthesizes realistic event camera data guided by diverse control signals, which can address the scarcity of labeled data for event-based vision tasks.
- We demonstrate the effectiveness of *ControlEvents* by synthesizing labeled event data for downstream tasks, achieving improved performance in visual recognition, 2D skeleton estimation, and 3D body pose estimation.
- We show *ControlEvents* can generate event data based on unseen text labels during training, showcasing its text-based generation capabilities and potential for zero-shot visual recognition tasks.
- We make both our model and large-scale synthesized datasets publicly available, which could foster research in event-based vision tasks.

2. Related Works

Event-based Vision Unlike frame-based cameras, which sample a scene at fixed intervals as a sequence of intensity frames, event cameras operate independently at the pixel level, detecting changes in brightness and generating event streams rather than conventional images. Event cameras capture asynchronous pixel-wise brightness changes, mimicking how the retina processes visual information [13, 24]. Due to high temporal resolution, low latency, high dynamic range, and efficient power usage, event cameras are particularly suited to dynamic applications. However, their distinct outputs challenge traditional computer vision algorithms, which are designed for frame-based data. Recent advancements have adapted event-based cameras to various computer vision tasks [13]. Applications in optical flow estimation [4, 14, 59], SLAM [7, 20, 38, 48], and 3D body recovery [42, 49, 50], demonstrate the potential of event cameras.

Despite above advantages, labeling ground truth for event data is challenging, which hinders the training of deep learning models for event-based vision tasks. To address this, event simulators [39, 42, 49, 50] generate synthetic events by subtracting consecutive frames from RGB renderings. Unfortunately, these simulators often results in a significant domain gap from real-world events. In this work, we leverage pre-trained foundation models [41, 57] to synthesize realistic event data, enabling producing infinite labeled events data from a limited real-world labeled events.

Synthetic Dataset Generation Synthetic datasets are increasingly used to address tasks with limited ground-truth data. This approach is prominent in optical flow estimation [47], point tracking [58], depth estimation [40, 43], and 3D body estimation [5, 8, 47, 55]. Most synthetic datasets are generated via computer graphics engines. For example, Infinigen [36, 37] uses Blender to render photorealistic scenes, and BEDLAM [5] leverages the Unreal Engine to create human movements with realistic clothing dynamics, providing accurate ground-truth for body pose and shape in SMPL [25, 34] format. These works underscore the potential of graphics engines for generating detailed synthetic data. However, high-quality 3D assets are required, which can be costly and time-intensive.

Recent progress in deep generative models has facilitated image synthesis directly from control signals, bypassing the need for extensive 3D assets. Methods using pre-trained models like Stable Diffusion [41] have enabled the generation of controlled images with minimal fine-tuning for depth and normal maps using ControlNet [57]. STAGE [22] synthesizes human images from skeletons and depth cues, while 3D-DST [26] uses 3D CAD models for controlled generation. These studies illustrate that generative models can produce realistic images under explicit control, and generated data could further enhancing the performance of machine learning models. However, no existing work explores controllable generation for events data, which exhibits a considerable domain gap from RGB images.

For event data generation, simulators like ESIM [39] generate events from rigid object motion, supporting applications in SLAM and object tracking. More recent simulators, such as EventHands [42] and E-LnR [49, 50], focus on non-rigid deformations like human body movement. However, these simulators struggle to simulate the noise and artifacts of real-world event cameras, leading to domain discrepancies between synthetic and actual event data. By contrast, our *ControlEvents* leverages the diffusion prior from [41, 57] and simulates characteristics of real-world captured events. Experiments in Fig. 2 and Sec. 4 show that our generated events outperform previous simulators and can be used as training data to improve the visual recognition and 2D & 3D pose estimation performance.

3. Method

3.1. Preliminaries

Event Camera model Event cameras record logarithmic pixel-level brightness change asynchronously. An event $e_i = (\mathbf{u}_i, t_i, p_i)$ is triggered at pixel $\mathbf{u}_i$ at time t_i when the logarithmic brightness change $\Delta\mathcal{L}$ reaches a threshold. The polarity of event e_i is

$$p(\mathbf{u}_i, t_i) = \begin{cases} +1 & \text{if } \mathcal{L}(\mathbf{u}_i, t_i) - \mathcal{L}(\mathbf{u}_i, t_{i-1}) \geq C\,, \\ -1 & \text{if } \mathcal{L}(\mathbf{u}_i, t_{i-1}) - \mathcal{L}(\mathbf{u}_i, t_i) \geq C\,, \end{cases} \quad (1)$$

where C is the contrastive threshold, a hardware-specific property of event cameras [24].

To fully leverage the capabilities of pre-trained image diffusion models [41, 57], we convert asynchronous events into 2D histograms, which can be represented as a 3-channel RGB image (H×W×3) with positive events shown in red and negative events in blue. We accumulate events into frames based on the strategy used in the adopted datasets [9, 21, 44, 62] for our downstream tasks.

In N-ImageNet [21], events are generated by using an event camera to capture original ImageNet images displayed on a screen. Hence, we accumulate all events for each image. For DHP-19 [9], we accumulate a fixed number (7,500) of events per frame. For CDEHP [44], events are accumulated over a fixed time interval (8.333 ms) per frame. In Event-HPE [62], events are aggregated over a fixed temporal length (15 ms) per frame. Since DHP-19, CDEHP, and Event-HPE do not provide polarity information, we represent event locations in white within the 2D histogram. In the paper, we visualize event images with a white background and blue color for enhanced appearance and reduced printing costs. Please refer to Fig. 1 and the supplementary material for visualizations of the event frames.

Diffusion Models under Control DDPM [18] is a generative model which learns a data distribution by iteratively adding and removing noise. Formally, the forward process in DDPM iteratively adds noise to a sample $\mathbf{x}_0$ drawn from a distribution $p_{\text{data}}(\mathbf{x})$:

$$\mathbf{x}_t \sim \mathcal{N}(\mathbf{x}_t; \sqrt{\alpha_t}\mathbf{x}_{t-1}, (1-\alpha_t)\mathbf{I}), \quad (2)$$

and α_t schedules the amount of noise added at each step t [18]. To sample data from the learned distribution, the reverse process starts from Gaussian noise $\mathbf{x}_T \sim \mathcal{N}(0, \mathbf{I})$ and iteratively denoises it until $t = 0$:

$$\mathbf{x}_{t-1} \sim \mathcal{N}(\mathbf{x}_{t-1}; \mu_\theta(\mathbf{x}_t, t), \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}(1-\alpha_t)\mathbf{I}), \quad (3)$$

where $\mu_\theta(\mathbf{x}_t, t)$ is the estimated posterior mean, predicted by a neural network ϵ_θ trained to minimize the difference

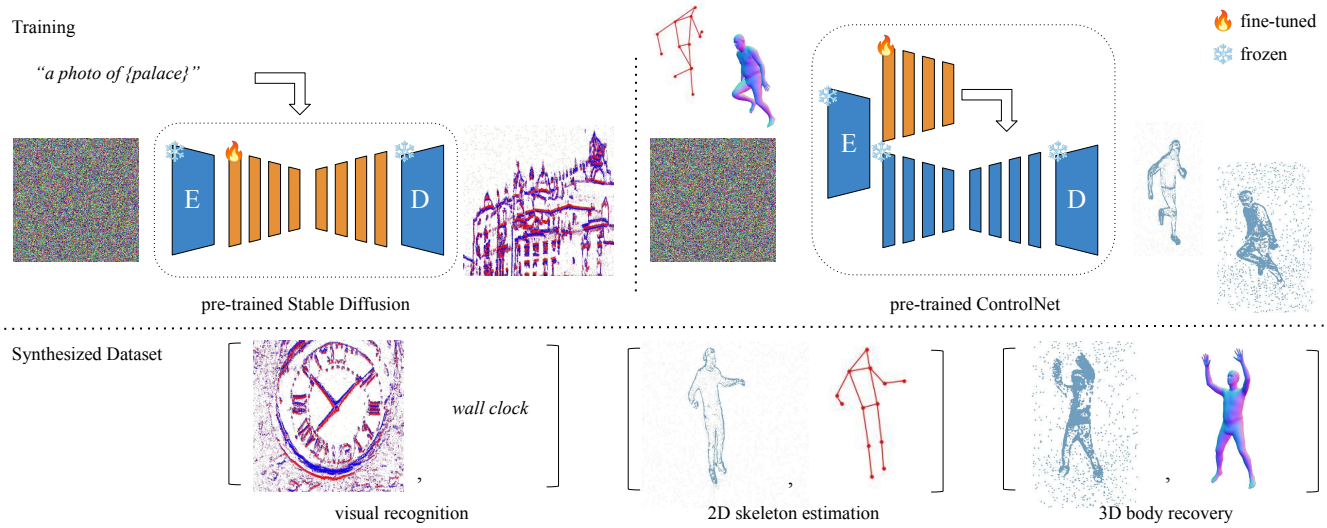


Figure 3. Overview of ***ControlEvents***. For text-conditioned event data synthesis, we fine-tune Stable Diffusion [41]. For 2D & 3D pose-conditioned event data synthesis, we fine-tune ControlNet [57] using skeleton map or normal map. Our *ControlEvents* can synthesize large-scale dataset for various tasks.

between true and estimated noise. One can also model conditional distribution $p_{\text{data}}(\mathbf{x}|\mathbf{c})$ by adding the condition to the network input [11, 17]. In practice, the network ϵ_θ predicts the amount of noise to remove based on the noisy state $\mathbf{x}_t$, condition x, and diffusion timestep t, optimizing an objective function:

$$L_{\text{objective}} = \sum_{t=1}^{T} \mathbb{E}_{\mathbf{x}_0,\epsilon,t}\left[\|\epsilon - \epsilon_\theta\left(\mathbf{x}_t, t, \mathbf{c}\right)\|^2\right]. \quad (4)$$

3.2. ControlEvents

Due to the scarcity of event data its ground-truth labels, training large-scale deep learning models for event-based tasks is impossible. Our goal is to improve model performance by training on synthetic event data generated by *ControlEvents*. This procedure can be viewed as an analysis-by-synthesis loop: given a small amount of labeled real-world event data, we train a generative model to synthesize event data with pseudo ground-truth labels. This synthetic data, paired with pseudo labels, serves as additional training data to enhance the model's analytical capabilities. To achieve this alignment, it is essential to match the distribution of synthetic events to the real event data. However, the limited availability of event data presents a scientific challenge: *how can a generative model effectively learn a robust data distribution when only a small amount of training data from the target distribution is available?*

Inspired by recent works [19, 51–54, 56, 57], we observe that the diffusion prior of foundational models (e.g., Stable Diffusion [41]) contains rich information that enhances the generalization capabilities of downstream applications. Thus, we design *ControlEvents* on top of foundational generative models, allowing us to leverage their generative power to synthesize realistic event data by minimal fine-tuning on small-scale real event datasets.

We denote the available small-scale real event datasets $\mathcal{D}_{\text{real}} = \{(\mathbf{e}^i, \mathbf{c}^i)\}_{i=1}^N$, where $\mathbf{e}^i$ is an event image obtained by accumulating the event stream within a fixed temporal window or a fixed number of events. The corresponding ground-truth label, $\mathbf{c}^i$, could be a class label in natural language [21, 33], a 2D human skeleton [9, 44], or a 3D human body [44] represented in SMPL [25, 34], depending on the generation task. To model the conditional data distribution $p_{\text{data}}(\mathbf{e}|\mathbf{c})$, we adopt the DDPM framework described in Sec. 3.1 and formulate the objective function as

$$L_{\textit{ControlEvents}} = \sum_{i=1}^{N}\sum_{t=1}^{T} \mathbb{E}_{\mathbf{e},\epsilon,t}\left[\left\|\epsilon - \epsilon_\theta\left(\mathbf{e}_t^i, t, \mathbf{c}^i\right)\right\|^2\right]. \quad (5)$$

Here, $\mathbf{e}_t^i$ is a noisy image generated from the clear event image $\mathbf{e}^i$ using DDPM scheduler as outlined in Eq. (2). The term $\epsilon_\theta(\cdot)$ represents the diffusion network, parametrized by parameters θ, which predicts the noise level to subtract during inference. At inference time, given an unseen control signal $\mathbf{c}^{\text{novel}}$, we synthesize the corresponding event image $\mathbf{e}^{\text{novel}}$ by sampling from the learned distribution $p_{\text{data}}(\mathbf{e}|\mathbf{c})$. Specifically, we perform the iterative denoising starting from randomly sampled Gaussian noise using Eq. (3) until a clear event image is generated.

Generation from Text Label To leverage the foundational text-to-image generative capabilities of image diffusion models [41], we fine-tune a pre-trained Stable Diffusion [41]. Originally, it models the text-conditioned image distribution $p_{\text{data}}(\mathbf{x}|\mathbf{c}^{\text{text}})$. Here, we fine-tune and adapt it to the target event distribution $p_{\text{data}}(\mathbf{e}|\mathbf{c}^{\text{class}})$.

In an event-based classsification dataset $\mathcal{D}_{\text{real}}^{\text{class}} = \{(\mathbf{e}^i, \mathbf{c}^{\text{class},i})\}_{i=1}^N$, such as N-ImageNet [21] or N-Caltech101 [33], each event image $\mathbf{e}^i$ is associated with a

class label $\mathbf{c}^{\text{class},i}$. We enhance the class label $\mathbf{c}^{\text{class},i}$ by augmenting it with the phrase: *"A photo of* $\{\mathbf{c}^{class,i}\}$*"*, making it resemble a natural language description and more aligned with the text conditioning used in Stable Diffusion [41]. The event image $\mathbf{e}^i$ is a 3-channel image, with positive events encoded in the first channel (red) and negative events in the third channel (blue). We utilize the frozen VAE encoder [41] to obtain the latent representation of $\mathbf{e}^i$ and fine-tune only the U-Net diffusion model using the objective function defined in Eq. (5). We denote the fine-tuned text-based event generator as *ControlEvents*$^{\text{class}}$.

After adapting the *ControlEvents*$^{\text{class}}$ with the event-based classification dataset $\mathcal{D}_{\text{real}}^{\text{class}}$, we generate events for each class with the adapted model conditioned on the same text prompt, forming a synthetic dataset $\mathcal{D}_{\text{synth}}^{\text{class}}$. We show in Sec. 4.1 that the synthetic data generated by our *ControlEvents*$^{\text{class}}$ can improve the event-based visual recognition performance.

Generation from 2D Skeleton Estimating 2D human pose in skeleton from event data is a valuable application, as human motion is often too fast for frame-based cameras and can cause motion blur. As a widely-used generative framework, ControlNet [57] enables the incorporation of 2D skeleton signals into the realistic image diffusion process of Stable Diffusion [41]. Therefore, we use ControlNet$^{\text{skeleton}}$ [57] as our starting point, as it already models the image distribution $p_{\text{data}}(\mathbf{x}|\mathbf{c}^{\text{skeleton}})$ effectively.

Here, we use an event-based pose estimation dataset $\mathcal{D}_{\text{real}}^{\text{skeleton}} = \{(\mathbf{e}^i, \mathbf{c}^{\text{skeleton},i})\}_{i=1}^{N}$, such as DHP-19 [9] or CDEHP [44]. Since ControlNet also supports text descriptions as additional conditioning, we apply the description *"A moving human"* for all event images $\mathbf{e}^i$. Following the training procedure of ControlNet [57], we fine-tune only the trainable copy of the U-Net, keeping the original Stable Diffusion U-Net and VAE frozen. We denote the fine-tuned skeleton-based event generator as *ControlEvents*$^{\text{skeleton}}$. Using this model, we generate event images from unseen 2D skeletons during training and construct a synthetic dataset $\mathcal{D}_{\text{synth}}^{\text{skeleton}}$, which we then use to train an event-based pose estimation model. Further details about the generation process and training of the 2D pose estimator are provided in Sec. 4.2.

Generation from 3D Body Estimating the parameters of a 3D body model such as SMPL [25, 34] from event data is more challenging than estimating a 2D skeleton due to self-occlusion and monocular ambiguity. Training large-scale deep learning models is a common approach, which motivates us to synthesize high-quality event data. However, no existing image foundation models take numerical SMPL parameters (e.g., pose and shape) as direct conditioning, which makes our fine-tuning difficult.

To fully leverage the diffusion prior of foundation models, we propose representing the 3D SMPL model using 2D images that can serve as conditioning inputs. Specifically, we choose normal maps as the 2D conditioning images for the 3D SMPL body, as they provide surface direction information at each pixel [61]. This design choice allows us to utilize the capabilities of pre-trained ControlNet$^{\text{normal}}$, which generates corresponding realistic images from normal map by sampling from learned distribution $p_{\text{data}}(\mathbf{x}|\mathbf{c}^{\text{normal}})$.

Given a real-world dataset which contains event data and corresponding SMPL parameters $\mathcal{D}_{\text{real}}^{\text{SMPL}} = \{(\mathbf{e}^i, \mathbf{c}^{\text{SMPL},i})\}_{i=1}^{N}$, such as Event-HPE [62], we render each SMPL parameters $\mathbf{c}^{\text{SMPL},i}$ into a 2D normal map $\mathbf{c}^{\text{normal},i}$ in Blender [10] with the camera intrinsics and extrinsics from [62]. Similar to the 2D pose task, we apply the description "A moving human" to all event images $\mathbf{e}^i$ and fine-tune only the trainable copy of the U-Net. We denote the SMPL normal-based event generator as *ControlEvents*$^{\text{SMPL}}$. Given such a model, we can easily generate event images from arbitrary SMPL parameters. We demonstrate that our *ControlEvents*$^{\text{SMPL}}$ generalizes well to diverse and challenging human poses from SMPL datasets like AMASS [32] and AIST [46]. This capability allows us to obtain large-scale challenging event data with pseudo ground-truth, which is not available in existing real-world event datasets. Please refer to Sec. 4.3 for details.

4. Experiments

In this section, we empirically demonstrate the effectiveness of our *ControlEvents* and synthesized event data by evaluating it on discriminative downstream tasks: object recognition in Sec. 4.1, 2D skeleton estimation in Sec. 4.2, and 3D SMPL estimation in Sec. 4.3. We further validate our synthetic event data using standard generative model metrics, namely Fréchet Inception Distance (FID) [16] in Tab. 4

Baselines Event simulators can be classified into two categories: learning-based approaches, such as Event-GAN [60], which synthesize events from two consecutive intensity frames; and graphics-based approaches, such as SMPL-ESIM [49, 50], which simulate events from a sequence of SMPL motions. Consequently, we compare our method with Event-GAN in Sec. 4.2 and Sec. 4.3, and with SMPL-ESIM Sec. 4.3, because these methods support similar control signals for those tasks. However, neither can synthesize event data from text-based class labels, which is a unique advantage of our proposed *ControlEvents*.

4.1. Event-based Object Recognition

Implementation Details We evaluate our method on the NImageNet dataset [21], which contains 1,000 classes. For each class in NImageNet, multiple (typically 2-5) similar class names are available. We use only 100 event images per class for learning $p_{\text{data}}(\mathbf{e}|\mathbf{c}^{\text{class}})$. In each training iteration,

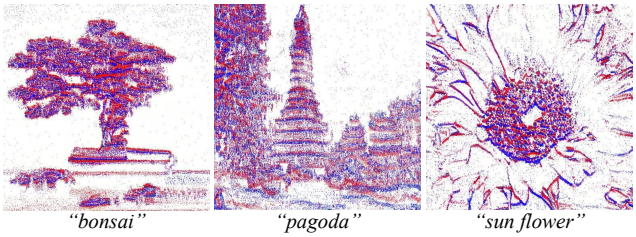
we randomly sample a class name and augment it as *"A photo of* $\{\mathbf{c}^{class,i}\}$*"* to serve as the text description. Please refer to supp. mat. for training details of *ControlEvents*$^{\text{class}}$.

Classification with Generated Events We infer *ControlEvents*$^{\text{class}}$ to synthesize 650 event images for each class. We train a classifier with a pre-trained ResNet-34 [15] and randomly initialize MLP prediction head with different setting, including original training data of our generative model and our synthesized event data. It shows that by combining our synthesized event images with original training data, we obtain better classification accuracy than only using the raw event data.

N-ImageNet evaluation set [9]		
Data setting	Acc. ↑	Prec. ↑
original training data	39.33	40.02
Ours	27.00	23.32
Ours + original training data	**45.33**	**46.52**

Table 1. **Object recognition** with different training data. Our synthetic event data can enhance classification performance.

Zero-Shot Generation A key insight in *ControlEvents* is leveraging the powerful diffusion prior in pre-trained foundation models. In our *ControlEvents*$^{\text{class}}$ experiments , we utilize the robust text-based generation capabilities inherited from Stable Diffusion [41]. Despite training on only 1,000 classes from N-ImageNet [21], our model generalizes to unseen text labels from the N-Caltech [33] dataset and can generate corresponding event images. We present qualitative results of zero-shot event data generation alongside their closest training class labels (determined by cosine similarity in CLIP [35] feature space) in Fig. 4. Please refer to the supplementary material for quantitative zero-shot event-based classification results on N-Caltech dataset.



"bonsai" *("broccoli", 0.65)* | *"pagoda"* *("palace", 0.70)* | *"sun flower"* *("daisy", 0.71)*

Figure 4. **Zero-shot generation of unseen text label**. We also present the closest seen text label during training based on the CLIP cosine similarity.

4.2. Event-based 2D Skeleton Estimation

Implementation Details We evaluate our method on the DHP-19 [9] and CDEHP [44] datasets because they provide the ground-truth 2D skeleton in the same format. We randomly sample 10000 event images with skeleton across different subjects and motion to train *ControlEvents*$^{\text{skeleton}}$. Please refer to supp. mat. for training details.

Pose Estimation with Generated Events We adopt the pose estimation network architecture from [9] and randomly initialize the network parameters during training. For our *ControlEvents*$^{\text{skeleton}}$, we randomly sample 5000 unseen skeletons from DHP-19 [9] and 5000 from CDEHP [44]. Since Event-GAN [60] requires intensity frames and DHP-19 [9] does not provide them, we sample 10000 frames from CDEHP [44] for generating events using Event-GAN.

We report the quantitative evaluation of 2 unseen subjects using mean per joint position error (MPJPE) and average precision (AP@25%, AP@50%) in Tab. 2. The results indicate that our synthetic event data is more effective for training 2D pose estimators and can further enhance performance when combined with the original training data of *ControlEvents*$^{\text{skeleton}}$.

4.3. Event-based 3D Body Recovery

Implementation Details We evaluate our method and data on the Event-HPE [62] dataset. We randomly sample 10000 event images with SMPL annotations to train *ControlEvents*$^{\text{SMPL}}$. Please refer to supp. mat. for details.

SMPL Recovery from Generated Events In this experiment, we use the randomly initialized event-based SMPL regressor from [62] as the SMPL recovery model. For our *ControlEvents*$^{\text{SMPL}}$, we randomly sample 10,000 unseen SMPL annotations from the Event-HPE dataset [62]. We compare our generated event images with those from Event-GAN [60] and the SMPL-ESIM simulator [50]. To minimize variations due to pose differences during generation, we use the same sampled SMPL annotations to synthesize events for both Event-GAN and SMPL-ESIM.

We report quantitative evaluation of 2 unseen subjects on MPJPE, mean per vertex error (PVE), percent of correct keypoints (PCK@25mm, PCK@50mm), and area under PCK curve (AUC). We compare 3D SMPL estimators with different training data, from original training data of *ControlEvents*$^{\text{SMPL}}$ to synthetic event data from baselines [50, 60] and our generated data. Quantitative results in Tab. 3 show that our generated data outperforms synthetic data from Event-GAN [60] and from SMPL-ESIM [50]. It also shows that our generated data can be combined with original training data to further enhance the 3D SMPL estimation performance.

Synthesizing Events for OOD Poses Generalizing to challenging out-of-distribution (OOD) poses remains a persistent issue in the pose recovery field, largely due to the scarcity of labeled training data for such complex poses, making this task inherently difficult. As illustrated in Fig. 5, we demonstrate that our *ControlEvents*$^{\text{SMPL}}$ generalizes effectively to challenging poses from the AMASS dataset [32]. Moreover, we quantitatively validate in Tab. 3 the benefits of training a SMPL estimator with our synthesized event data using pseudo ground-truth derived from

	DHP-19 evaluation set [9]			CDEHP evaluation set [44]		
Data Setting	2D-MPJPE ↓	AP@25 ↑	AP@50 ↑	2D-MPJPE ↓	AP@25 ↑	AP@50 ↑
original training data	7.180	99.5	98.0	12.149	98.2	95.9
Event-GAN [60]	7.986	98.4	98.0	16.264	97.5	94.1
Ours	6.223	**99.7**	99.1	10.936	98.8	96.5
Ours + original training data	**6.006**	**99.7**	**99.4**	**10.792**	**98.9**	**96.7**

Table 2. **2D pose estimation** with different training data. Our synthetic event data outperforms [60] and enhances estimation performance.

	Event-HPE evaluation set [9]				
Data Setting	3D-MPJPE$_{mm}$ ↓	3D-PCK@25 ↑	3D-PCK@50 ↑	PVE$_{mm}$ ↓	AUC ↑
original training data	27.30	46.05	95.02	34.24	81.79
Event-GAN [60]	29.43	35.61	91.79	35.27	80.37
SMPL-ESIM [50]	53.53	5.84	42.54	55.35	36.44
Ours w/ Event-HPE pose	25.13	54.97	96.45	34.22	83.24
Ours w/ AMASS pose	28.32	45.11	92.59	34.36	81.12
Ours both + original training data	**18.65**	**79.37**	**99.66**	**32.12**	**87.57**

Table 3. **3D SMPL Body Recovery** with different training data. Our synthetic event data outperforms both learning-based synthetic data [60] and graphics-based synthetic data [49, 50]. Additionally, we demonstrate the capability to generate synthetic events for challenging poses from AMASS [32], which significantly enhances SMPL estimation performance.

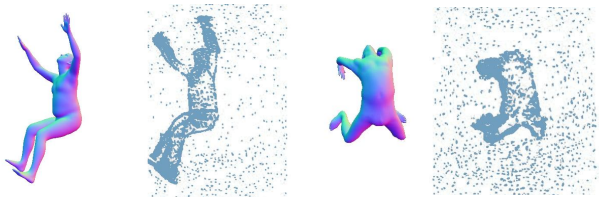


Figure 5. **SMPL-based generation on challenging unseen poses**. Given challenging AMASS [32] poses, we can generate realistic event data.

AMASS. Training solely with AMASS synthetic data yields slightly less accurate results, primarily due to differing pose distributions between the challenging AMASS dataset [32] and the Event-HPE evaluation set [62]. However, training on a combination of real and challenging synthetic event data significantly improves SMPL recovery performance with a 31.7% reduction in 3D-MPJPE.

Text-to-Events in Motion Our *ControlEvents*$^{\text{SMPL}}$ can synthesize realistic event data from arbitrary SMPL annotations, providing the potential to leverage powerful text-to-motion models [23, 45] for generating human motion events based on text descriptions. In contrast, other learning-based methods, such as [60], which generate events from intensity frames, cannot directly condition on SMPL for generation. Please refer to the supplementary video for animations.

4.4. Event Generation Quality

We quantitatively evaluate the quality of generated event data from our method compared to baselines under diverse conditions, as shown in Tab. 4. As a reference, we provide the FID [16] scores of random unseen real event data relative to the training data distribution of generative models from the same datasets. As shown in Fig. 2 and Tab. 4, the event data generated by *ControlEvents* is closest in distribution to the training data and closely approximates real captured event data, demonstrating the effectiveness of our method in generating high-quality synthetic events.

Method	FID$_{SMPL}$ ↓	FID$_{skeleton}$ ↓	FID$_{class}$ ↓
Unseen real event data	3.64	23.22	0.36
Event-GAN [60]	196.41	165.93	–
SMPL-ESIM [50]	206.23	–	–
Ours	**15.55**	**28.97**	**27.43**

Table 4. **FID score between synthesized event and real-world event data.** We evaluate the quality of conditional distribution modelling w.r.t. SMPL, skeleton, and class label. Our synthetic event is more close to real event data compared to [49, 50, 60].

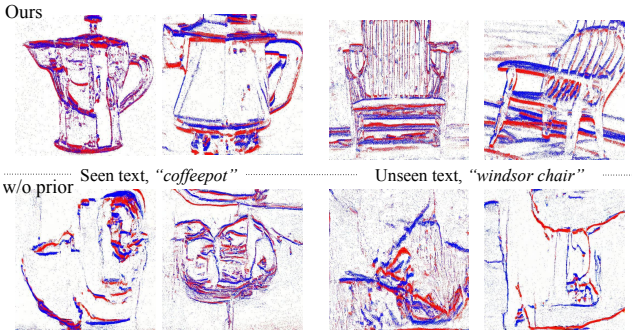


Figure 6. **Ablation on foundational prior from Stable Diffusion [41].** Without the prior, our method cannot generate meaningful on unseen text.

4.5. Ablation Study

Diffusion Prior from Foundation Models A key insight in *ControlEvents* is leveraging the powerful diffusion prior in pre-trained foundation models, enabling fine-tuning with

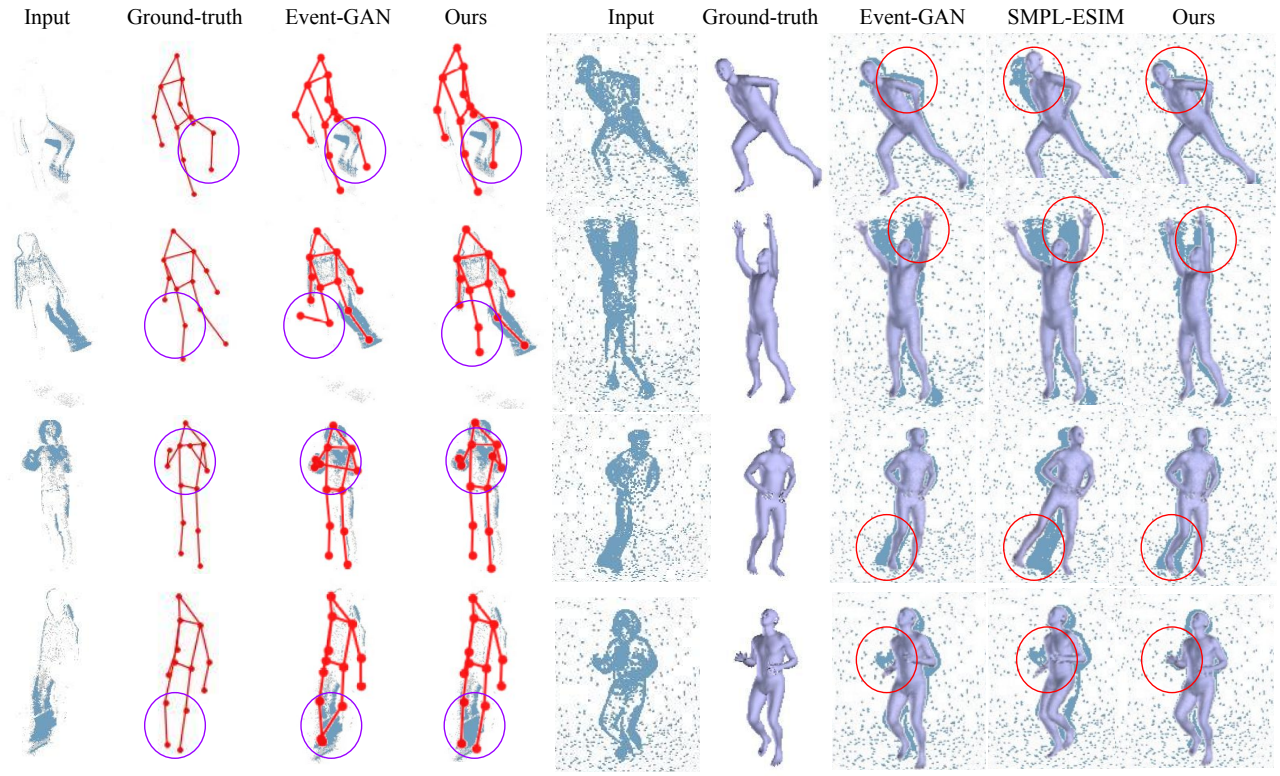


Figure 7. Comparison synthetic data between ***ControlEvents*** and other event simulators (Event-GAN [60], SMPL-ESIM [50]) for 2D skeletal estimation and 3D body recovery. Estimators trained on *ControlEvents* generated data shows better performance due to high quality data, especially on unseen parts.

minimal data while inheriting the capabilities of the original models. To demonstrate the effectiveness of this approach, we present *Zero-Shot Generation* experiments in Sec. 4.1, showing our model's impressive text-to-image capabilities by generalizing to unseen text labels. Here, we further examine the benefits of the diffusion prior by conducting an ablation study on the diffusion prior from pre-trained foundation models.

Using the same training data $\mathcal{D}_{\text{real}}^{\text{class}}$ as experiments in Sec. 4.1, we directly train a *ControlEvents*$^{\text{class}}$ model without a foundational diffusion prior by randomly initializing the U-Net weights of Stable Diffusion [41]. In Fig. 6, we visualize the generation results at intermediate steps during training. The results intuitively show that, without the diffusion prior, the generative model struggles to learn the distribution of the training event data. We also observe that the model fails to generate meaningful event images for unseen text labels, producing only noise instead.

Scaling of Synthesized Training Data An advantage of *ControlEvents* its ability to synthesize realistic event data with pseudo ground-truth labels at low cost. We conducted an ablation study to examine the impact of scaling the amount of synthesized event data from *ControlEvents*$^{\text{skeleton}}$ for training a 2D pose estimator by scaling the dataset size from 0 (0.0x) to 20,000 (2.0x) samples, as illustrated in Fig. 8. The results indicate that up-scaling the synthesized data can enhance the model performance.

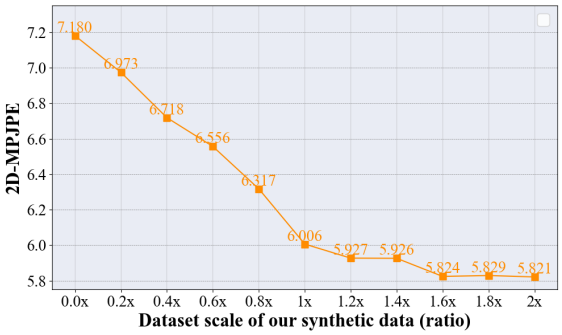


Figure 8. **Ablation on scaling synthetic data of skeleton estimation**. Our synthesized events with pseudo ground-truth can enhance the skeleton estimation along with original training dataset.

5. Conclusion

We present ***ControlEvents***, a controllable event synthesis approach that leverages the foundational priors of pre-trained image diffusion models. We demonstrate that: 1) generating events with pseudo ground-truth labels can effectively enhance the performance of deep learning models, and 2) the diffusion prior from foundational models helps in generalizing to unseen scenarios for events generation

tasks. We believe that our approach paves the way for new directions in data synthesis for event-based vision tasks. We will make our model and synthetic event dataset for all tasks available for future research.

Limitations and future work To best leverage the priors from image diffusion models, we accumulate event streams into images, which may result in the loss of rich temporal information inherent in asynchronous streams. A promising direction for future exploration is to utilize priors from video diffusion models [6, 27–31], which can incorporate temporal information and potentially handle the time steps of event data more effectively.

Acknowledgements We thank G.Tiwari, I.Sarandi, V.Guzov, Y.He, X.Xie, C.Li whose feedback helped improve this paper. This work is made possible by funding from the Carl Zeiss Foundation. This work is also funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - 409792180 (EmmyNoether Programme, project: Real Virtual Humans) and the German Federal Ministry of Education and Research (BMBF): Tübingen AI Center, FKZ: 01IS18039A. The authors thank the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for supporting Y.Xue. G. Pons-Moll is a member of the Machine Learning Cluster of Excellence, EXC number 2064/1 – Project number 390727645.

Y.Hu and Y.Xue contributed equally as the joint first authors. Y.Xue is the corresponding author. Authors with equal contributions are listed in alphabetical order and allowed to change their orders freely on their resume and website. Y.Xue initialized the core idea, organized the project, co-developed the current method, co-supervised the experiments, and wrote the draft. Y.Hu co-developed the current method, implemented most of the prototypes, and conducted experiments.

A. Dataset

A.1. Dataset Description

Classification In this paper, we utilize N-ImageNet [21] and N-Caltech101 [33] as the event data classification datasets. Both datasets are event-based conversions of ImageNet [21] and Caltech101 [12], created by capturing original RGB images displayed on a monitor using an event camera.

These datasets generate events from static RGB images by introducing relative camera-to-image motion, achieved either by moving the event cameras [21] or by moving the monitor [33]. Details of the event acquisition system used in N-ImageNet [21] are illustrated in Fig. 9.

A.2. Processing Datasets

As described in Section 3.1 of the main paper, we convert asynchronous events into RGB images to leverage the priors from pre-trained foundation models. We visualize the processed datasets in Fig. 10, including N-ImageNet [21], N-Caltech101 [33], DHP-19 [9], CDEHP [44], and Event-HPE [62]. The processing procedure is described as:

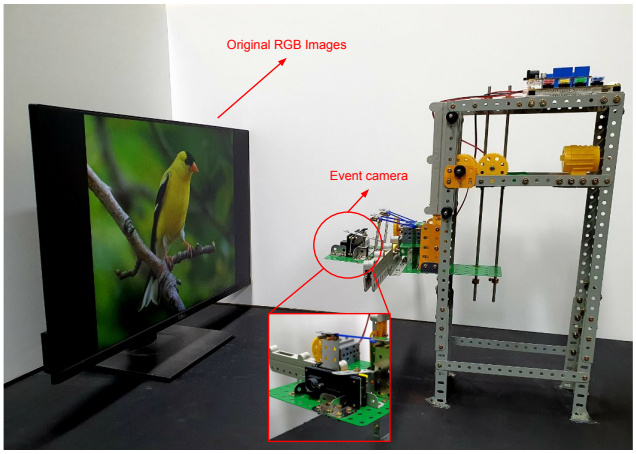


Figure 9. **Conversion hardware system** in N-ImageNet [21]. The event camera is vibrated vertically and horizontally on a plane parallel to the LCD monitor screen to capture events of displayed images.

- N-ImageNet: All events are accumulated into a single image, capturing the original ImageNet image. Positive events are represented in red, and negative events in blue.
- N-Caltech101: Similar to N-ImageNet, all events are accumulated into a single image. Positive events are represented in red, and negative events in blue.
- DHP-19: Following the original paper [9], 7,500 events are accumulated to construct each event image. We follow the original paper and discard polarities. DHP-19 includes 2D skeleton annotations as ground truth but does not include intensity frames.
- CDEHP: Following the original paper [44], events are accumulated over 8.333 ms to construct each event image. We follow the CDEHP includes both 2D skeleton annotations and intensity frames.
- Event-HPE: Following the original paper [62], events are accumulated over 15 ms to construct each event image. We follow the original paper and discard polarities of events. The dataset provides both 3D SMPL annotations and intensity frames.

B. Method

B.1. Training details

ControlEvents$^{\text{class}}$ We train our text-based event generative model by adopting the Stable Diffusion v1.4 text-to-image checkpoint [1] and fine-tuning it on our collected event data. The model is trained on 4 NVIDIA A40 GPUs for 80 epochs, with a batch size of 4 and gradient accumulation over 32 steps, resulting in an effective batch size of 512. The training process takes approximately 5 days.

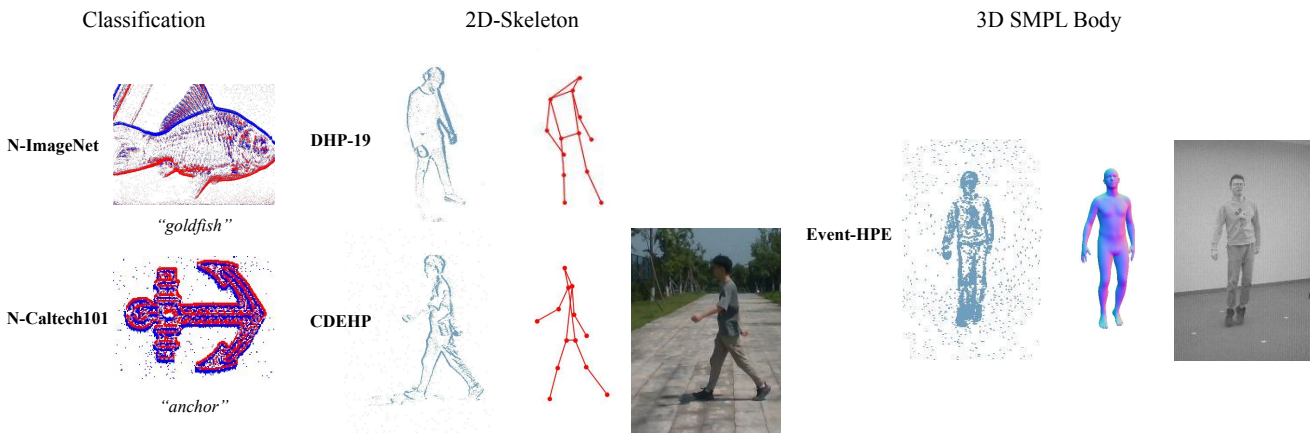


Figure 10. **Visualization of processed datasets.** N-ImageNet [21] and N-Caltech101 [33] provide polarities of events. CDEHP [44] and Event-HPE [62] provide intensity frames.

***ControlEvents*$^{\text{skeleton}}$** We train our skeleton-based event generative model by adopting the ControlNet human skeleton v1.0 checkpoint [3] and fine-tuning it on our collected event data. The model is trained on 4 NVIDIA A40 GPUs for 600 epochs, with a batch size of 2 and gradient accumulation over 64 steps, resulting in an effective batch size of 512. The training process takes approximately 3 days.

***ControlEvents*$^{\text{SMPL}}$** We train our SMPL-based event generative model by adopting the ControlNet normal v1.0 checkpoint [2] and fine-tuning it on our collected event data, using normal maps rendered from SMPL annotations as input. The model is trained on 4 NVIDIA A40 GPUs for 600 epochs, with a batch size of 2 and gradient accumulation over 64 steps, resulting in an effective batch size of 512. The entire training process takes approximately 3 days.

B.2. Inference details

At inference time, we use DDPM [18] as scheduler and set reverse sampling steps to 1000. The whole inference time on NVIDIA A40 (48GB) for *ControlEvents*$^{\text{class}}$ takes 12 seconds, for *ControlEvents*$^{\text{skeleton}}$ takes 17s, and for *ControlEvents*$^{\text{SMPL}}$ takes 17s.

C. Experiments

C.1. Generation from Text

In the main paper, we only present the generated event images from text labels in fig.1. Thus, we visualize more text-conditioned generation results in Fig. 11. Please refer to our supplementary video for more visualization results.

C.2. Zero-Shot Text-based Generation

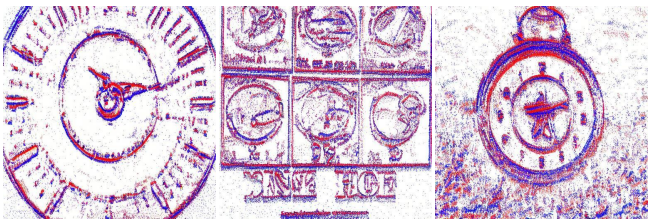
As introduced in section 4.1, our *ControlEvents*$^{\text{class}}$ can generate event images from text labels that were unseen during training on the N-ImageNet [21] dataset. Here, we present additional qualitative zero-shot binary classification results (see Fig. 12) for 10 unseen classes from the N-Caltech101 [33] dataset. In Fig. 12 we show the closest seen text label from N-ImageNet [21] and their cosine similarity in the CLIP [35] space. We use *ControlEvents*$^{\text{class}}$ to generate 500 event images for each class. For comparison, we train the baseline classification model using 50 event images per class from the N-Caltech101 dataset. Quantitative results demonstrate that our zero-shot generated large-scale dataset outperforms the few-shot classification baseline trained on limited data.

C.3. Generation from Skeleton & SMPL

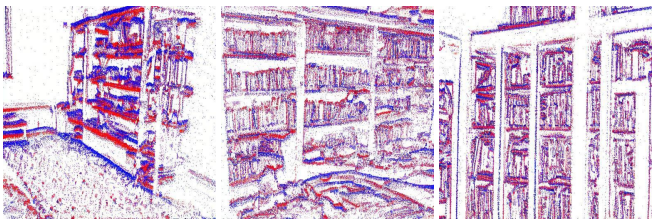
We visualize more skeleton-conditioned event images generation results in Fig. 13 and SMPL-conditioned generation results in Fig. 14. Please refer to our supplementary video for animation results. We also present the animation results of text-to-events in motion, which we describe in Sec.4.3 in the main paper. Please refer to our supplementary video for the details.

N-Caltech101 zero-shot set [33]		
Class	Baseline Acc. ↑	Our Acc. ↑
"bonsai"	55.00%	**76.25%**
"ceiling fan"	**85.06%**	67.82%
"ferry"	60.22%	**68.82%**
"joshua tree"	61.54%	**70.51%**
"minaret"	54.44%	**60.00%**
"pagoda"	65.15%	**80.30%**
"sunflower"	60.00%	**74.67%**
"water lilly"	**72.41%**	55.17%
"windsor chair"	**62.16%**	58.11%
"chandelier"	58.02%	**70.37%**

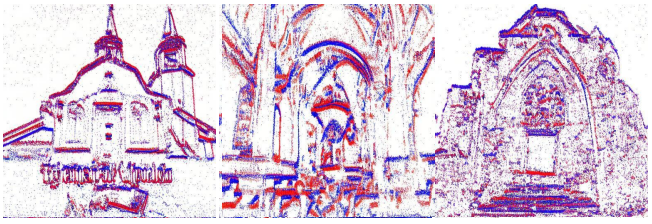
Table 5. **Zero-shot binary classification** on unseen classes from N-Caltech101 [33].



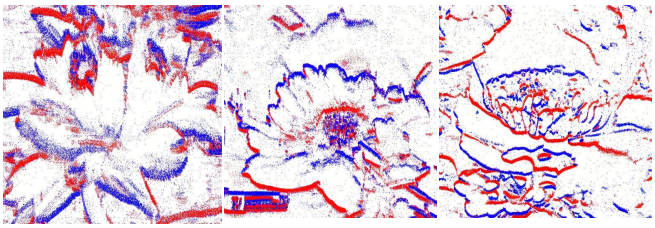
“analog clock”



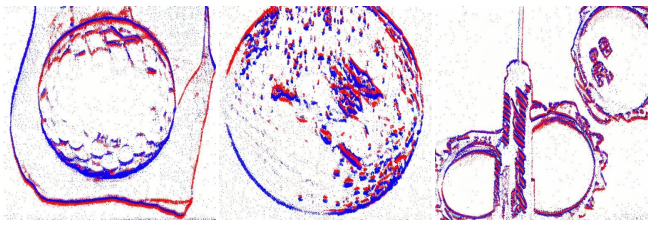
“bookcase”



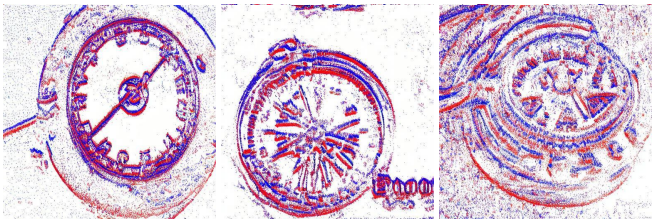
“church”



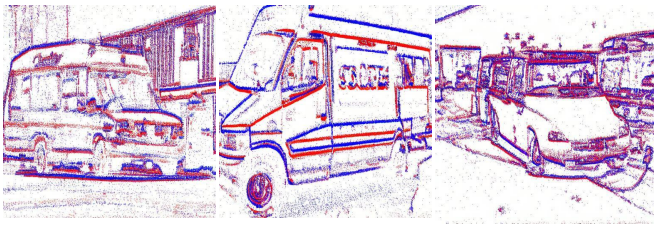
“daisy”



“golf ball”



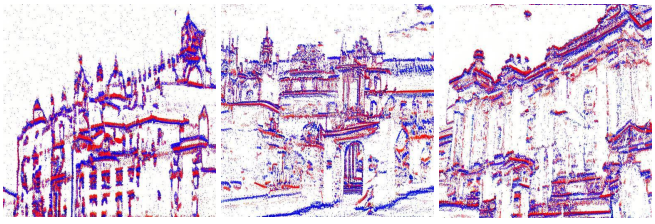
“magnetic compass”



“minibus”

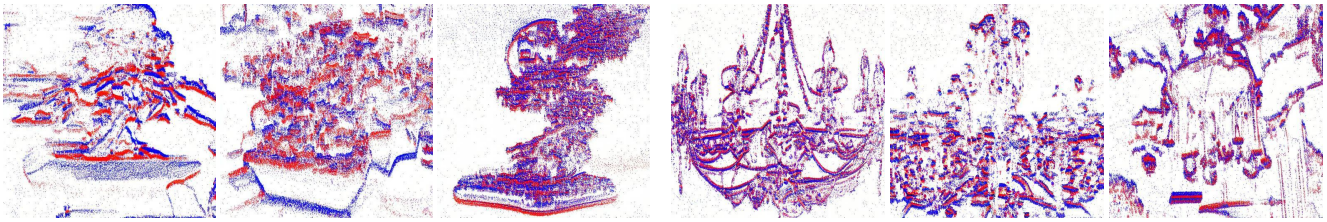
“teapot”

“running shoes”



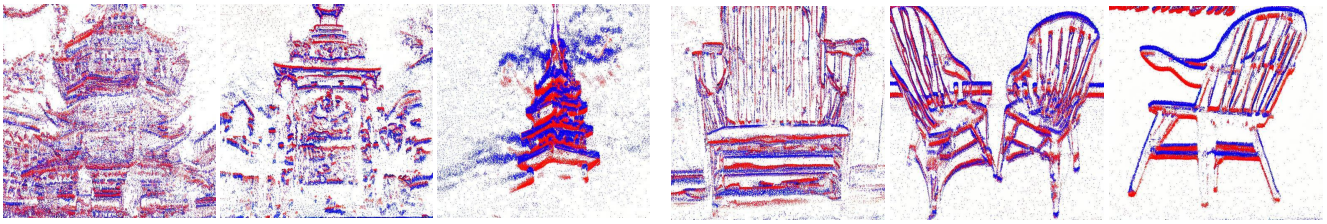
“palace”

Figure 11. **Generation of event images from text class labels** in N-ImageNet [21].



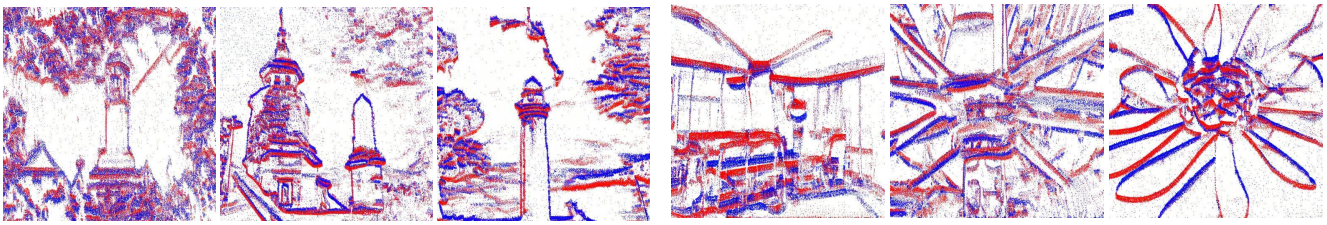
“bonsai”
(“broccoli”, 0.65)

“chandelier”
(“candle”, 0.70)



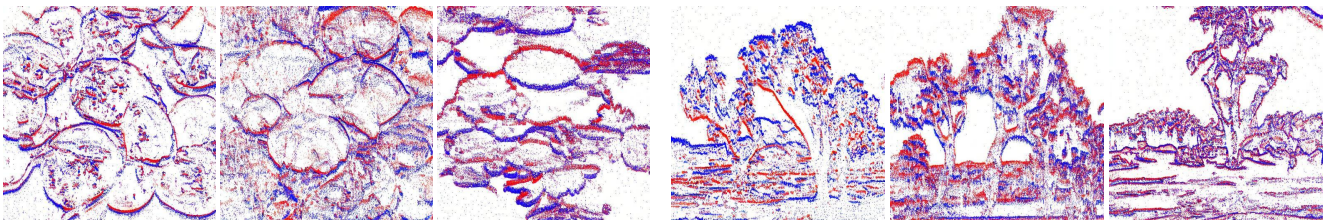
“pagoda”
(“palace”, 0.70)

“windsor chair”
(“windsor tie”, 0.68)



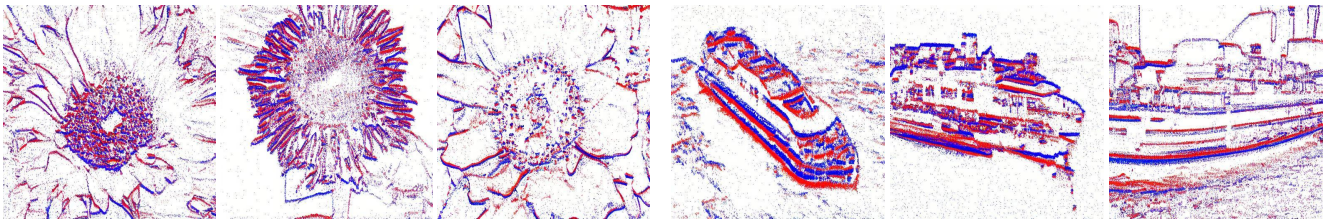
“minaret”
(“mini”, 0.67)

“ceiling fan”
(“electric fan”, 0.60)



“water lily”
(“water snake”, 0.75)

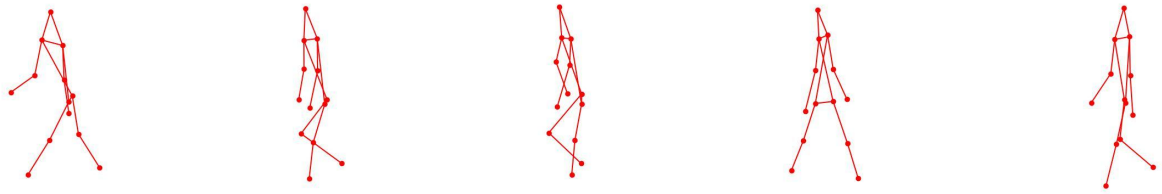
“joshua tree”
(“joystick”, 0.55)



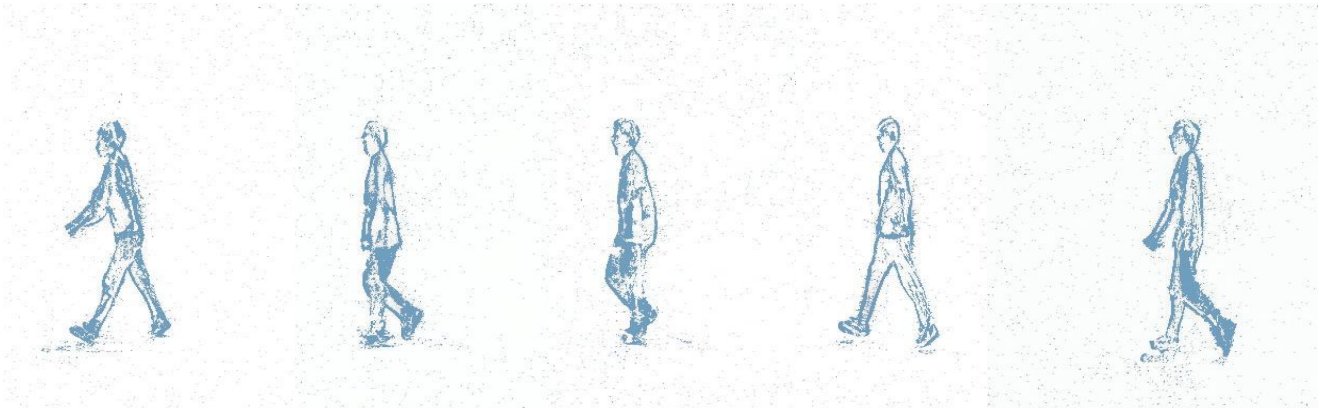
“sunflower”
(“daisy”, 0.71)

“ferry”
(“jetty”, 0.70)

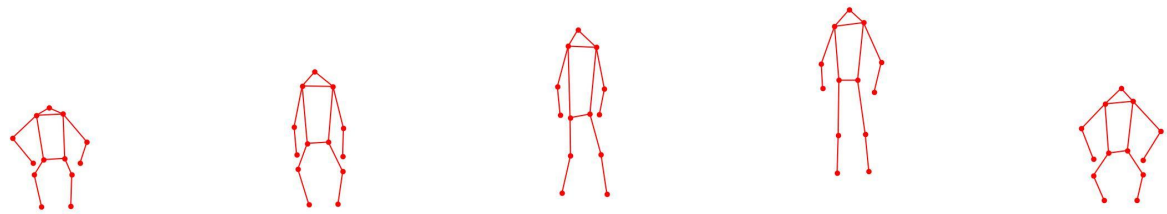
Figure 12. **Zero-shot generation of unseen text label** from N-Caltech101 [33] dataset. We determine the closest seen text label during training based on the CLIP cosine similarity.



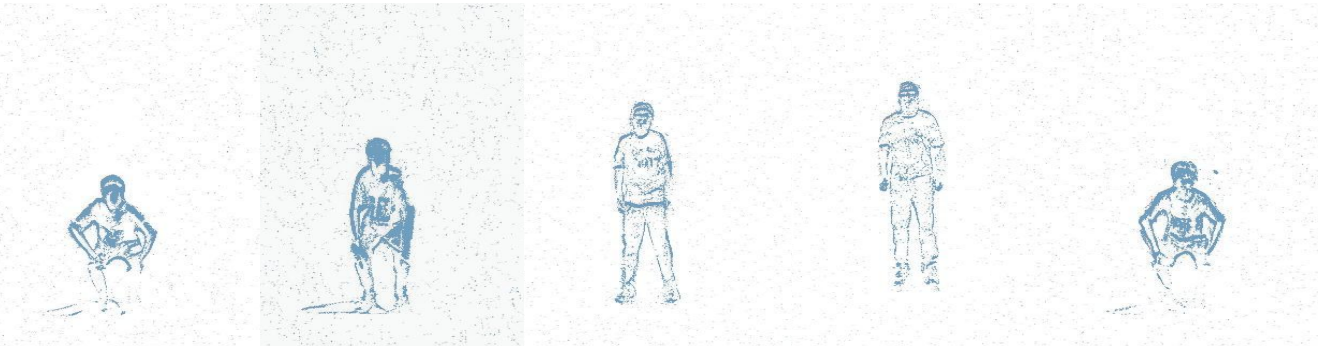
Condition: 2D Skeleton



Generation of events



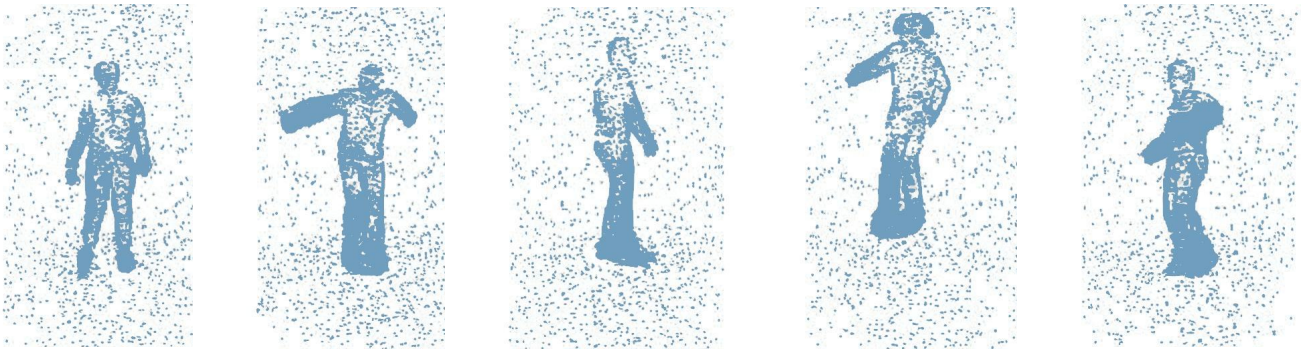
Condition: 2D Skeleton



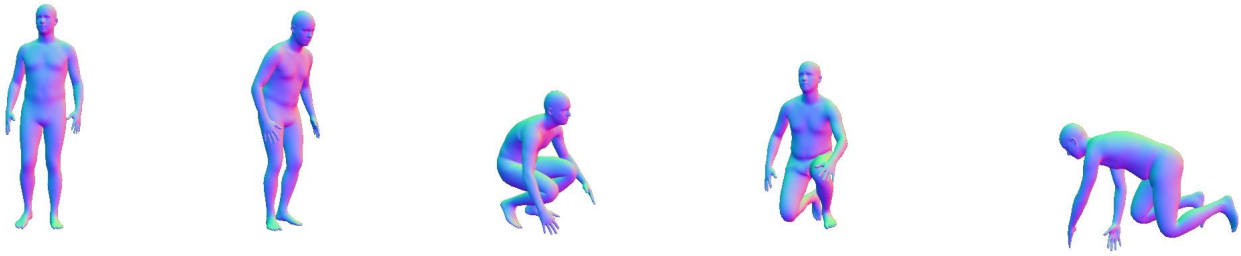
Generation of events

Figure 13. **Generation of event images from 2D skeleton.** Please refer to supplementary video for animation results.

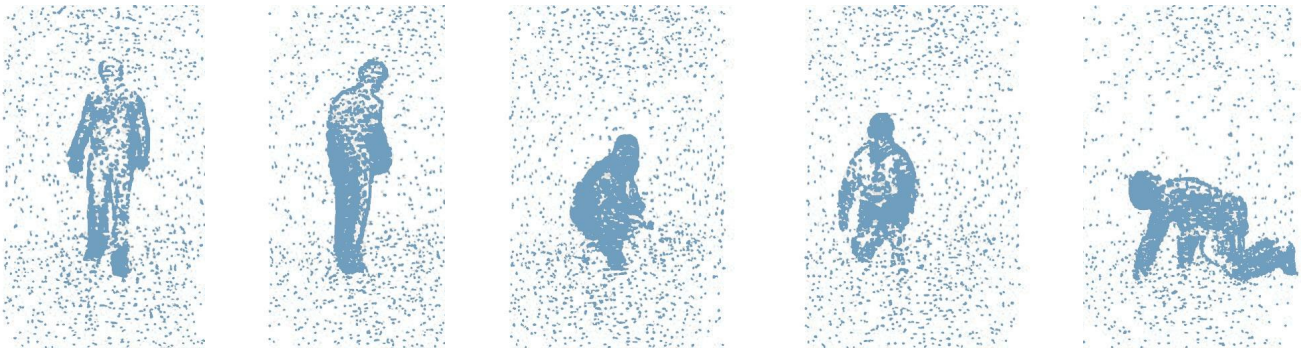
Condition: SMPL normal map



Generation of events



Condition: SMPL normal map



Generation of events

Figure 14. **Generation of event images from 3D SMPL normal maps.** Please refer to supplementary video for animation results.

References

[1] *HuggingFace, CompVis/stable-diffusion-v1-4*, 2022. 9

[2] *HuggingFace, lllyasviel/sd-controlnet-normal*, 2023. 10

[3] *HuggingFace, lllyasviel/sd-controlnet-openpose*, 2023. 10

[4] Patrick Bardow, Andrew J. Davison, and Stefan Leutenegger. Simultaneous optical flow and intensity estimation from an event camera. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 884–892. IEEE Computer Society, 2016. 2

[5] Michael J. Black, Priyanka Patel, Joachim Tesch, and Jinlong Yang. BEDLAM: A synthetic dataset of bodies exhibiting detailed lifelike animated motion. In *Proceedings IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 8726–8737, 2023. 3

[6] Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. Video generation models as world simulators. 2024. 9

[7] Samuel Bryner, Guillermo Gallego, Henri Rebecq, and Davide Scaramuzza. Event-based, direct camera tracking from a photometric 3d map using nonlinear optimization. In *International Conference on Robotics and Automation, ICRA 2019, Montreal, QC, Canada, May 20-24, 2019*, pages 325–331. IEEE, 2019. 2

[8] Zhongang Cai, Mingyuan Zhang, Jiawei Ren, Chen Wei, Daxuan Ren, Zhengyu Lin, Haiyu Zhao, Lei Yang, Chen Change Loy, and Ziwei Liu. Playing for 3d human recovery. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–12, 2024. 3

[9] Enrico Calabrese, Gemma Taverni, Christopher Awai Easthope, Sophie Skriabine, Federico Corradi, Luca Longinotti, Kynan Eng, and Tobi Delbrück. DHP19: dynamic vision sensor 3d human pose dataset. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 1695–1704. Computer Vision Foundation / IEEE, 2019. 2, 3, 4, 5, 6, 7, 9

[10] Blender Online Community. *Blender - a 3D modelling and rendering package*. Blender Foundation, Stichting Blender Foundation, Amsterdam, 2018. 5

[11] Prafulla Dhariwal and Alexander Quinn Nichol. Diffusion models beat GANs on image synthesis. In *Advances in Neural Information Processing Systems*, 2021. 4

[12] Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. *Computer Vision and Pattern Recognition Workshop*, 2004. 9

[13] Guillermo Gallego, Tobi Delbrück, Garrick Orchard, Chiara Bartolozzi, Brian Taba, Andrea Censi, Stefan Leutenegger, Andrew J. Davison, Jörg Conradt, Kostas Daniilidis, and Davide Scaramuzza. Event-based vision: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(1):154–180, 2022. 2

[14] Mathias Gehrig, Mario Millhäusler, Daniel Gehrig, and Davide Scaramuzza. E-raft: Dense optical flow from event cameras. In *International Conference on 3D Vision (3DV)*, 2021. 2

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 770–778. IEEE Computer Society, 2016. 6

[16] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 6626–6637, 2017. 5, 7

[17] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021. 4

[18] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. 3, 10

[19] Bingxin Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. Repurposing diffusion-based image generators for monocular depth estimation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 9492–9502. IEEE, 2024. 4

[20] Hanme Kim, Stefan Leutenegger, and Andrew J. Davison. Real-time 3d reconstruction and 6-dof tracking with an event camera. In *European Conference on Computer Vision (ECCV)*, pages 349–364, 2016. 2

[21] Junho Kim, Jaehyeok Bae, Gangin Park, Dongsu Zhang, and Young Min Kim. N-imagenet: Towards robust, fine-grained object recognition with event cameras. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 2126–2136. IEEE, 2021. 2, 3, 4, 5, 6, 9, 10, 11

[22] Nikita Kister, István Sárándi, Anna Khoreva, and Gerard Pons-Moll. Are pose estimators ready for the open world? STAGE: synthetic data generation toolkit for auditing 3d human pose estimators. *CoRR*, abs/2408.16536, 2024. 3

[23] Chuqiao Li, Julian Chibane, Yannan He, Naama Pearl, Andreas Geiger, and Gerard Pons-Moll. Unimotion: Unifying 3d human motion synthesis and understanding. *CoRR*, abs/2409.15904, 2024. 7

[24] Patrick Lichtsteiner, Christoph Posch, and Tobi Delbruck. A 128× 128 120 db 15 μs latency asynchronous temporal contrast vision sensor. *IEEE Journal of Solid-State Circuits*, 43 (2):566–576, 2008. 2, 3

[25] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: a skinned multi-person linear model. *ACM Trans. Graph.*, 34(6):248:1–248:16, 2015. 3, 4, 5

[26] Wufei Ma, Qihao Liu, Jiahao Wang, Angtian Wang, Xiaoding Yuan, Yi Zhang, Zihao Xiao, Guofeng Zhang, Beijia Lu, Ruxiao Duan, Yongrui Qi, Adam Kortylewski, Yaoyao Liu,

and Alan L. Yuille. Generating images with 3d annotations using diffusion models. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. 3

[27] Yue Ma, Yingqing He, Xiaodong Cun, Xintao Wang, Siran Chen, Xiu Li, and Qifeng Chen. Follow your pose: Pose-guided text-to-video generation using pose-free videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 4117–4125, 2024. 9

[28] Yue Ma, Hongyu Liu, Hongfa Wang, Heng Pan, Yingqing He, Junkun Yuan, Ailing Zeng, Chengfei Cai, Heung-Yeung Shum, Wei Liu, et al. Follow-your-emoji: Fine-controllable and expressive freestyle portrait animation. In *SIGGRAPH Asia 2024 Conference Papers*, pages 1–12, 2024.

[29] Yue Ma, Kunyu Feng, Xinhua Zhang, Hongyu Liu, David Junhao Zhang, Jinbo Xing, Yinhan Zhang, Ayden Yang, Zeyu Wang, and Qifeng Chen. Follow-your-creation: Empowering 4d creation through video inpainting. *arXiv preprint arXiv:2506.04590*, 2025.

[30] Yue Ma, Yingqing He, Hongfa Wang, Andong Wang, Leqi Shen, Chenyang Qi, Jixuan Ying, Chengfei Cai, Zhifeng Li, Heung-Yeung Shum, et al. Follow-your-click: Open-domain regional image animation via motion prompts. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 6018–6026, 2025.

[31] Yue Ma, Yulong Liu, Qiyuan Zhu, Ayden Yang, Kunyu Feng, Xinhua Zhang, Zhifeng Li, Sirui Han, Chenyang Qi, and Qifeng Chen. Follow-your-motion: Video motion transfer via efficient spatial-temporal decoupled finetuning. *arXiv preprint arXiv:2506.05207*, 2025. 9

[32] Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. AMASS: archive of motion capture as surface shapes. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 5441–5450. IEEE, 2019. 5, 6, 7

[33] Garrick Orchard, Ajinkya Jayawant, Gregory Cohen, and Nitish V. Thakor. Converting static image datasets to spiking neuromorphic datasets using saccades. *CoRR*, abs/1507.07629, 2015. 2, 4, 6, 9, 10, 12

[34] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3d hands, face, and body from a single image. *CoRR*, abs/1904.05866, 2019. 3, 4, 5

[35] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, pages 8748–8763. PMLR, 2021. 6, 10

[36] Alexander Raistrick, Lahav Lipson, Zeyu Ma, Lingjie Mei, Mingzhe Wang, Yiming Zuo, Karhan Kayan, Hongyu Wen, Beining Han, Yihan Wang, Alejandro Newell, Hei Law, Ankit Goyal, Kaiyu Yang, and Jia Deng. Infinite photorealistic worlds using procedural generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12630–12641, 2023. 3

[37] Alexander Raistrick, Lingjie Mei, Karhan Kayan, David Yan, Yiming Zuo, Beining Han, Hongyu Wen, Meenal Parakh, Stamatis Alexandropoulos, Lahav Lipson, Zeyu Ma, and Jia Deng. Infinigen indoors: Photorealistic indoor scenes using procedural generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21783–21794, 2024. 3

[38] Henri Rebecq, Timo Horstschaefer, and Davide Scaramuzza. Real-time visual-inertial odometry for event cameras using keyframe-based nonlinear optimization. In *British Machine Vision Conference (BMVC)*, 2017. 2

[39] Henri Rebecq, Daniel Gehrig, and Davide Scaramuzza. ESIM: an open event camera simulator. *Conf. on Robotics Learning (CoRL)*, 2018. 2, 3

[40] Mike Roberts, Jason Ramapuram, Anurag Ranjan, Atulit Kumar, Miguel Angel Bautista, Nathan Paczan, Russ Webb, and Joshua M. Susskind. Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding. In *ICCV*, 2021. 3

[41] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021. 2, 3, 4, 5, 6, 7, 8

[42] Viktor Rudnev, Vladislav Golyanik, Jiayi Wang, Hans-Peter Seidel, Franziska Mueller, Mohamed Elgharib, and Christian Theobalt. Eventhands: Real-time neural 3d hand pose estimation from an event stream. In *International Conference on Computer Vision (ICCV)*, 2021. 2, 3

[43] Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, Devi Parikh, and Dhruv Batra. Habitat: A Platform for Embodied AI Research. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019. 3

[44] Zhanpeng Shao, Xueping Wang, Wen Zhou, Wuzhen Wang, Jianyu Yang, and Youfu Li. A temporal densely connected recurrent network for event-based human pose estimation. *Pattern Recognit.*, 147:110048, 2024. 2, 3, 4, 5, 6, 7, 9, 10

[45] Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Daniel Cohen-Or, and Amit H. Bermano. Human motion diffusion model. *CoRR*, abs/2209.14916, 2022. 7

[46] Shuhei Tsuchida, Satoru Fukayama, Masahiro Hamasaki, and Masataka Goto. AIST dance video database: Multi-genre, multi-dancer, and multi-camera database for dance information processing. In *Proceedings of the 20th International Society for Music Information Retrieval Conference, ISMIR 2019, Delft, The Netherlands, November 4-8, 2019*, pages 501–510, 2019. 5

[47] Gül Varol, Javier Romero, Xavier Martin, Naureen Mahmood, Michael J. Black, Ivan Laptev, and Cordelia Schmid. Learning from synthetic humans. In *CVPR*, 2017. 3

[48] Antoni Rosinol Vidal, Henri Rebecq, Timo Horstschaefer, and Davide Scaramuzza. Ultimate slam? combining events, images, and IMU for robust visual SLAM in HDR and high-speed scenarios. *IEEE Robotics Autom. Lett.*, 3(2):994–1001, 2018. 2

[49] Yuxuan Xue, Haolong Li, Stefan Leutenegger, and Joerg Stueckler. Event-based non-rigid reconstruction from contours. In *33rd British Machine Vision Conference 2022, BMVC 2022, London, UK, November 21-24, 2022*, page 78. BMVA Press, 2022. 2, 3, 5, 7

[50] Yuxuan Xue, Haolong Li, Stefan Leutenegger, and Jörg Stückler. Event-based non-rigid reconstruction of low-rank parametrized deformations from contours. *Int. J. Comput. Vis.*, 132(8):2943–2961, 2024. 2, 3, 5, 6, 7, 8

[51] Yuxuan Xue, Xianghui Xie, Riccardo Marin, and Gerard Pons-Moll. Human-3diffusion: Realistic avatar creation via explicit 3d consistent diffusion models. 2024. 4

[52] Yuxuan Xue, Xianghui Xie, Riccardo Marin, and Gerard Pons-Moll. Gen-3diffusion: Realistic image-to-3d generation via 2d & 3d diffusion synergy. *CoRR*, abs/2412.06698, 2024.

[53] Yuxuan Xue, Xianghui Xie, Margaret Kostyrko, and Gerard Pons-Moll. Infinihuman: Infinite 3d human creation with precise control. *SIGGRAPH Asia 2025 Conference Papers*, 2025.

[54] Pradyumna Yalandur Muralidhar, Yuxuan Xue, Xianghui Xie, Margaret Kostyrko, and Gerard Pons-Moll. Physic: Physically plausible 3d human-scene interaction and contact from a single image. *SIGGRAPH Asia 2025 Conference Papers*, 2025. 4

[55] Zhitao Yang, Zhongang Cai, Haiyi Mei, Shuai Liu, Zhaoxi Chen, Weiye Xiao, Yukun Wei, Zhongfei Qing, Chen Wei, Bo Dai, Wayne Wu, Chen Qian, Dahua Lin, Ziwei Liu, and Lei Yang. Synbody: Synthetic dataset with layered human models for 3d human perception and modeling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 20282–20292, 2023. 3

[56] Chongjie Ye, Lingteng Qiu, Xiaodong Gu, Qi Zuo, Yushuang Wu, Zilong Dong, Liefeng Bo, Yuliang Xiu, and Xiaoguang Han. Stablenormal: Reducing diffusion variance for stable and sharp normal. *CoRR*, abs/2406.16864, 2024. 4

[57] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 3813–3824. IEEE, 2023. 2, 3, 4, 5

[58] Yang Zheng, Adam W. Harley, Bokui Shen, Gordon Wetzstein, and Leonidas J. Guibas. Pointodyssey: A large-scale synthetic dataset for long-term point tracking. In *ICCV*, 2023. 3

[59] Alex Zihao Zhu, Liangzhe Yuan, Kenneth Chaney, and Kostas Daniilidis. Unsupervised event-based optical flow using motion compensation. In *Computer Vision - ECCV 2018 Workshops - Munich, Germany, September 8-14, 2018, Proceedings, Part VI*, pages 711–714. Springer, 2018. 2

[60] Alex Zihao Zhu, Ziyun Wang, Kaung Khant, and Kostas Daniilidis. Eventgan: Leveraging large scale image datasets for event cameras. In *IEEE International Conference on Computational Photography, ICCP 2021, Haifa, Israel, May 23-25, 2021*, pages 1–11. IEEE, 2021. 2, 5, 6, 7, 8

[61] Shenhao Zhu, Junming Leo Chen, Zuozhuo Dai, Yinghui Xu, Xun Cao, Yao Yao, Hao Zhu, and Siyu Zhu. Champ: Controllable and consistent human image animation with 3d parametric guidance. *CoRR*, abs/2403.14781, 2024. 5

[62] Shihao Zou, Chuan Guo, Xinxin Zuo, Sen Wang, Pengyu Wang, Xiaoqin Hu, Shoushun Chen, Minglun Gong, and Li Cheng. Eventhpe: Event-based 3d human pose and shape estimation. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 10976–10985. IEEE, 2021. 2, 3, 5, 6, 7, 9, 10