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

## Abstract

*In this supplementary document, we provide further details about the training details, inference details, and promised results in the main paper. Please also refer to our supplementary video for more results animation.*

## 1. Dataset

### 1.1. Dataset Description

**Classification** In this paper, we utilize N-ImageNet [7] and N-Caltech101 [8] as the event data classification datasets. Both datasets are event-based conversions of ImageNet [7] and Caltech101 [5], 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 [7] or by moving the monitor [8]. Details of the event acquisition system used in N-ImageNet [7] are illustrated in Fig. 1.

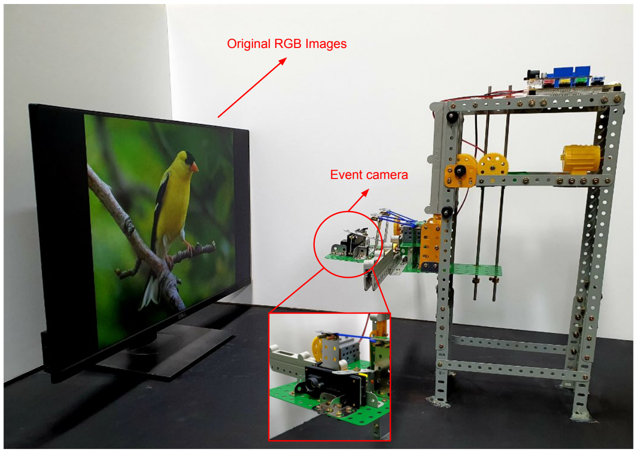


Figure 1. **Conversion hardware system** in N-ImageNet [7]. The event camera is vibrated vertically and horizontally on a plane parallel to the LCD monitor screen to capture events of displayed images.

### 1.2. 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. 2, including N-ImageNet [7], N-Caltech101 [8], DHP-19 [4], CDEHP [10], and Event-HPE [11]. The processing procedure is described as:

- 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 [4], 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 [10], 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 [11], 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.

## 2. Method

### 2.1. Training details

***ControlEvents***$^{\textbf{class}}$ 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.

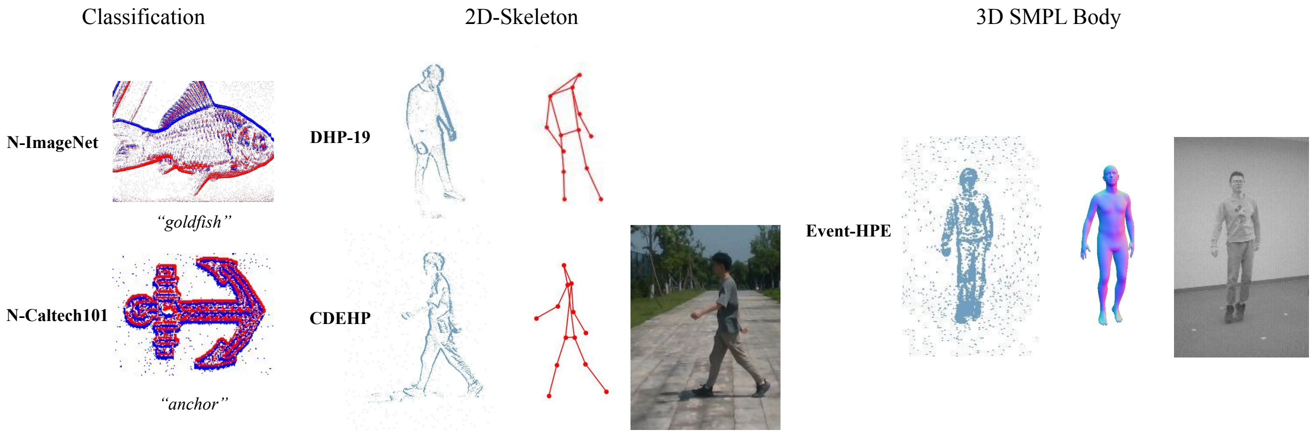


Figure 2. **Visualization of processed datasets.** N-ImageNet [7] and N-Caltech101 [8] provide polarities of events. CDEHP [10] and Event-HPE [11] provide intensity frames.

***ControlEvents*$^{\textbf{skeleton}}$** 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*$^{\textbf{SMPL}}$** 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.

## 2.2. Inference details

At inference time, we use DDPM [6] 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.

# 3. Experiments

## 3.1. 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. 3. Please refer to our supplementary video for more visualization results.

## 3.2. Zero-Shot Text-based Generation

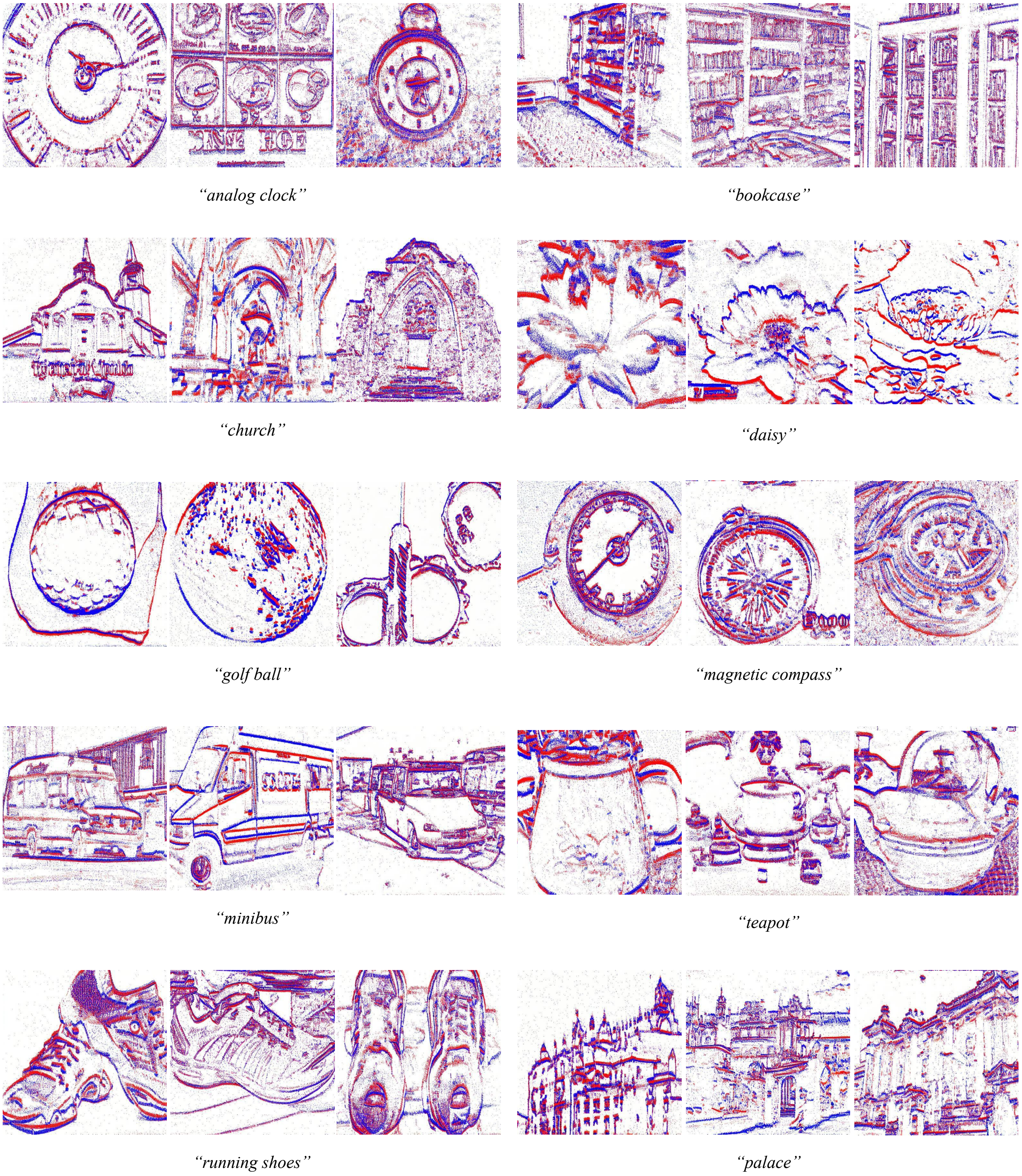
As introduced in section 4.1, our *ControlEvents*$^{\text{class}}$ can generate event images from text labels that were unseen during training on the N-ImageNet [7] dataset. Here, we present additional qualitative zero-shot binary classification results (see Fig. 4) for 10 unseen classes from the N-Caltech101 [8] dataset. In Fig. 4 we show the closest seen text label from N-ImageNet [7] and their cosine similarity in the CLIP [9] 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.

## 3.3. Generation from Skeleton & SMPL

We visualize more skeleton-conditioned event images generation results in Fig. 5 and SMPL-conditioned generation results in Fig. 6. 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 [8] | | |
|---|---|---|
| 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 1. **Zero-shot binary classification** on unseen classes from N-Caltech101 [8].



*"analog clock"*

*"bookcase"*

*"church"*

*"daisy"*

*"golf ball"*

*"magnetic compass"*

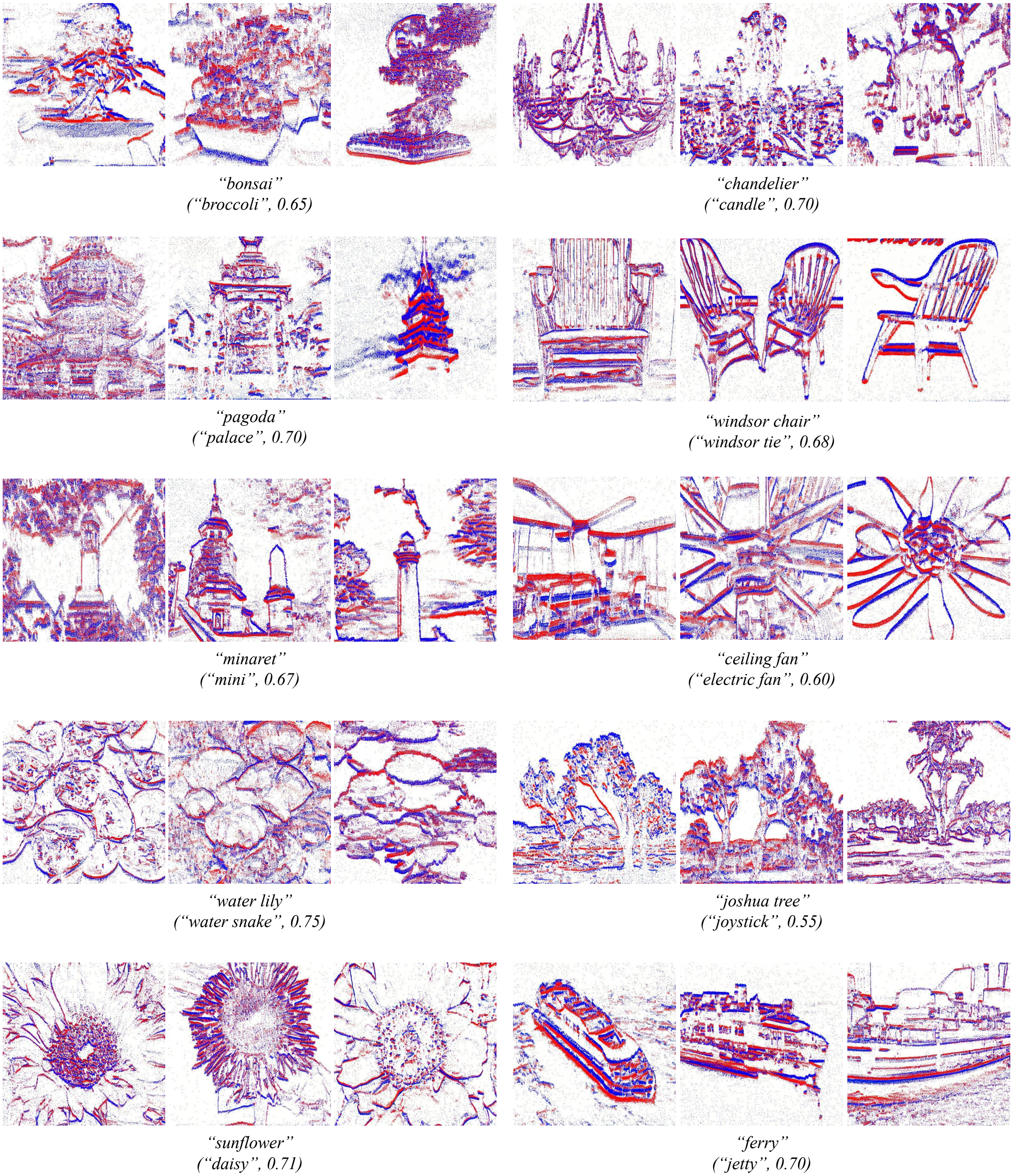
*"minibus"*

*"teapot"*

*"running shoes"*

*"palace"*

Figure 3. **Generation of event images from text class labels** in N-ImageNet [7].



*"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 4. **Zero-shot generation of unseen text label** from N-Caltech101 [8] dataset. We determine the closest seen text label during training based on the CLIP cosine similarity.

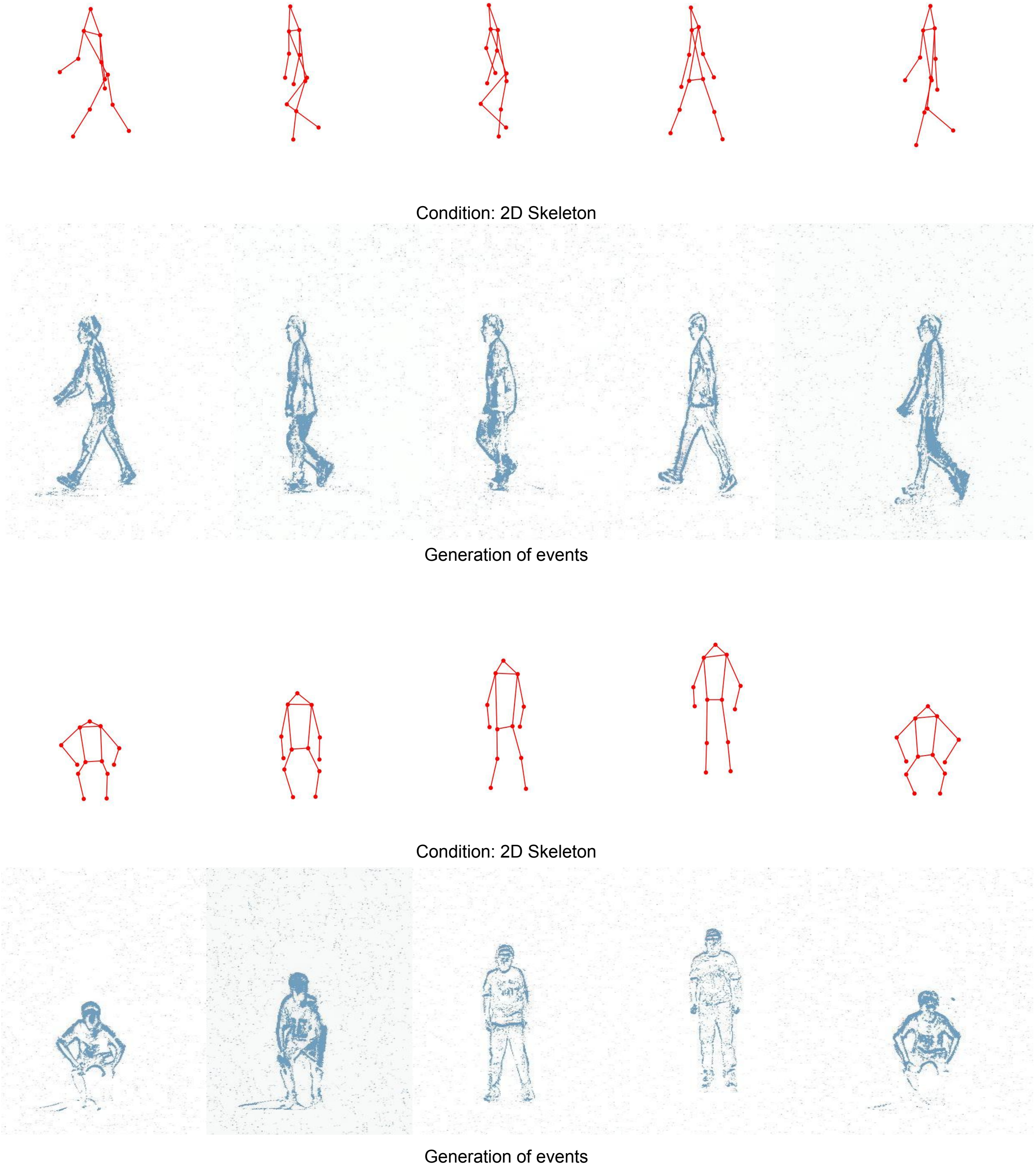


Figure 5. **Generation of event images from 2D skeleton.** Please refer to supplementary video for animation results.

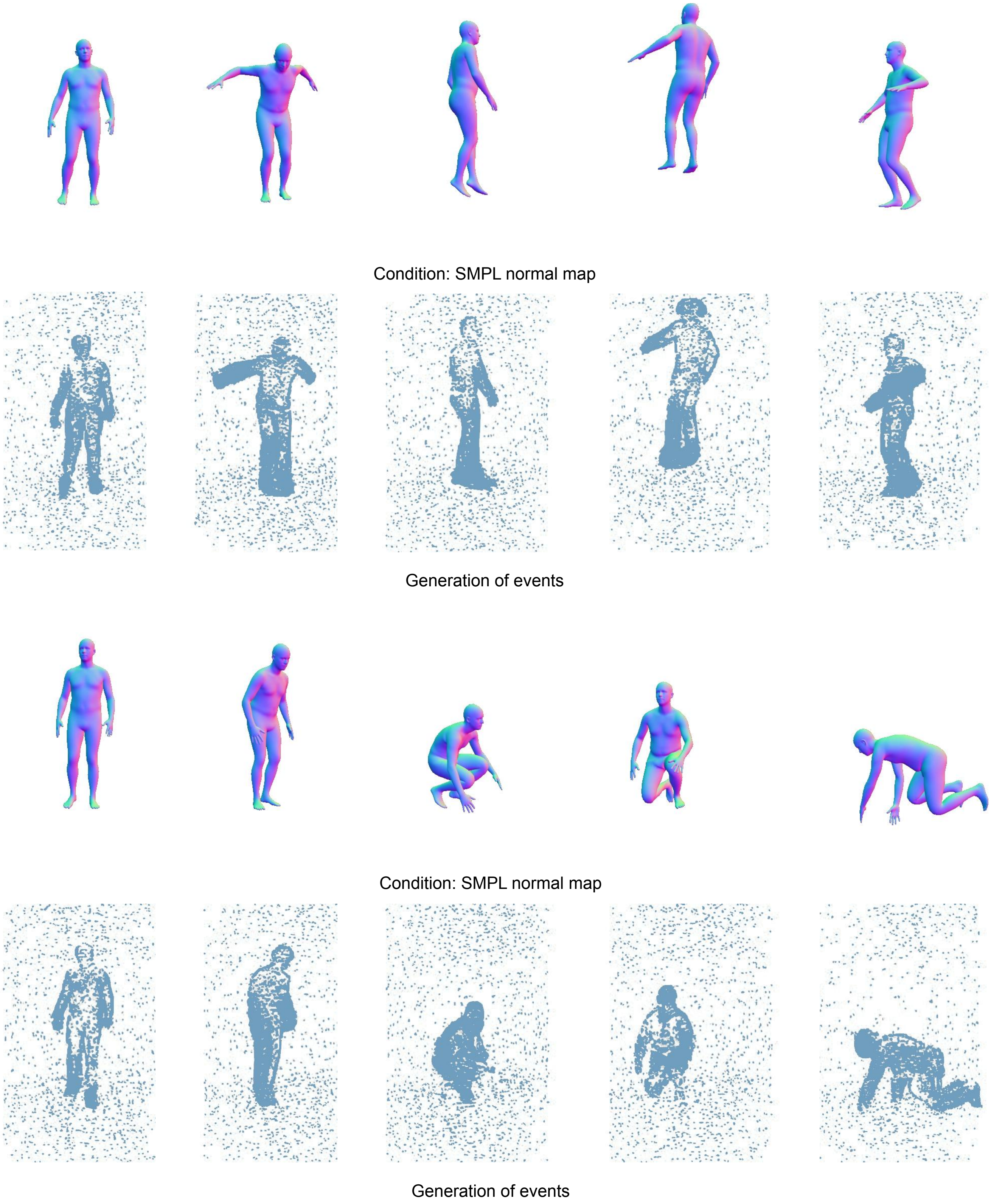


Figure 6. **Generation of event images from 3D SMPL normal maps.** Please refer to supplementary video for animation results.

## References

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

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

[3] *HuggingFace, lllyasviel/sd-controlnet-openpose*, 2023. 2

[4] 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. 1

[5] 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. 1

[6] 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. 2

[7] 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. 1, 2, 3

[8] 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. 1, 2, 4

[9] 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. 2

[10] 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. 1, 2

[11] 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. 1, 2