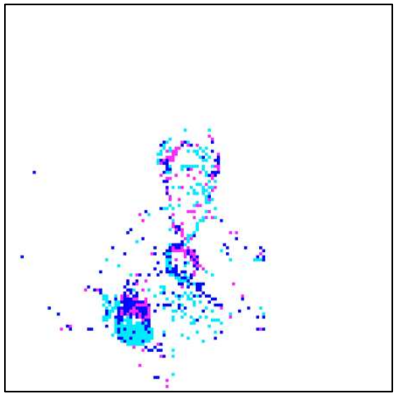
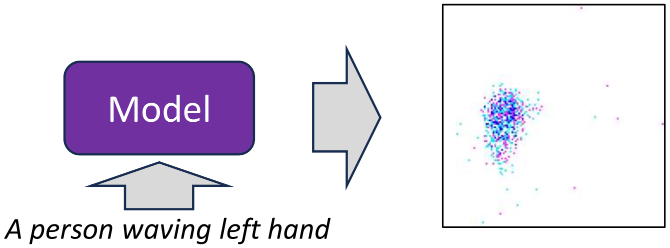
Text-to-Events: Synthetic Event Camera Streams from Conditional Text Input

Joachim Ott, Zuowen Wang & Shih-Chii Liu
Institute of Neuroinformatics,
University of Zurich & ETH Zurich



Text2Events

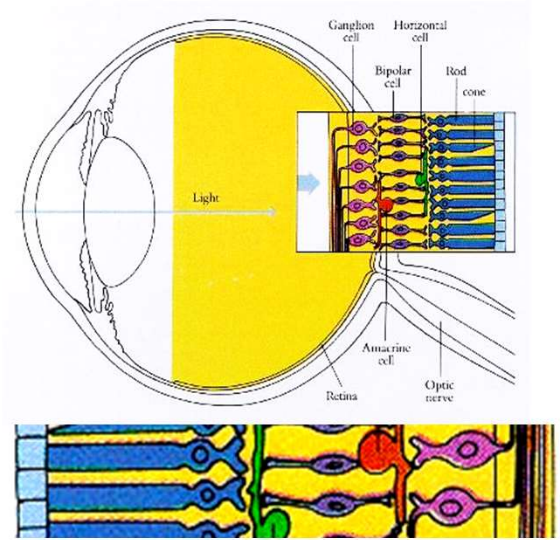


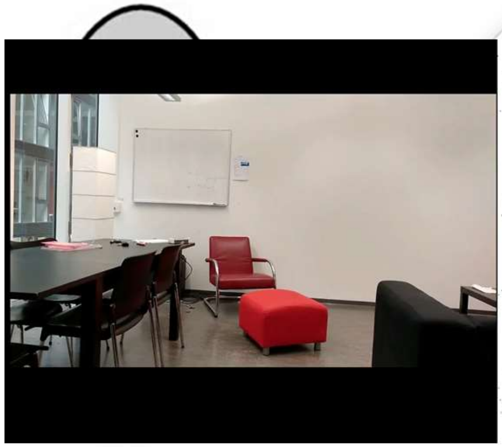
Model

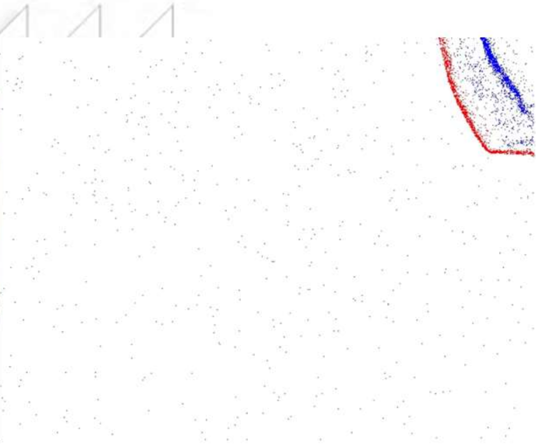
A person waving left hand

Ott et al, NICE Conf, SD, 23.04.2024

Event Camera or Dynamic Vision Sensor







P. Lichtsteiner et al., JSSC, Feb. 2007

Image sources: S. Guo et al. “Low Cost and Latency Event Camera Background Activity Denoising”,2022
E. Mueggler et al. “Event-based, 6-DOF Pose Tracking for High-Speed Maneuvers",2014, Rodieck et al, The first Steps in Seeing, 1998.

‘A ball bouncing on the grass’
[4]

‘A squirrel eating a burger.’
[3]

‘A dog driving a car on a suburban street wearing funny sunglasses’
[2]

‘Drone view of waves crashing against the rugged cliffs along Big Sur’s garay point beach. The crashing blue waters create white-tipped waves, while the golden light of the setting sun illuminates the rocky shore. A small island with a lighthouse sits in the distance, and green shrubbery covers the cliff’s edge. The steep drop from the road down to the beach is a dramatic feat, with the cliff’s edges jutting out over the sea. This is a view that captures the raw beauty of the coast and the rugged landscape of the Pacific Coast Highway.’

[1]

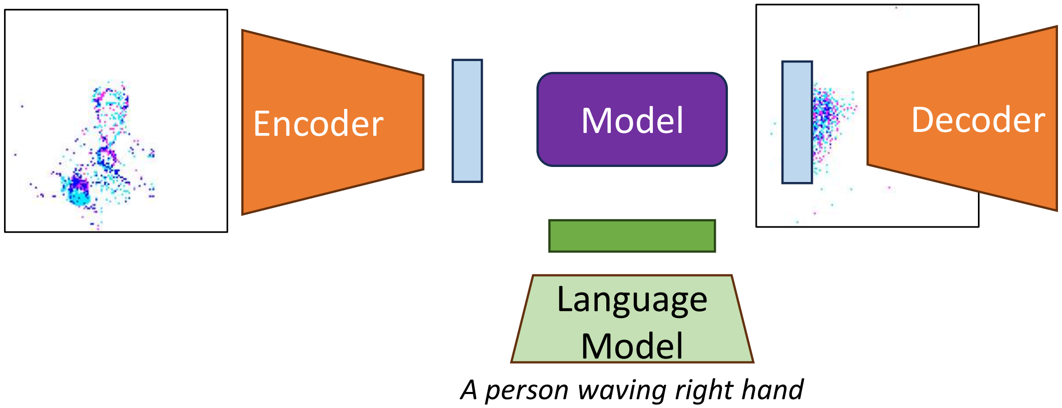
time

2023

2024

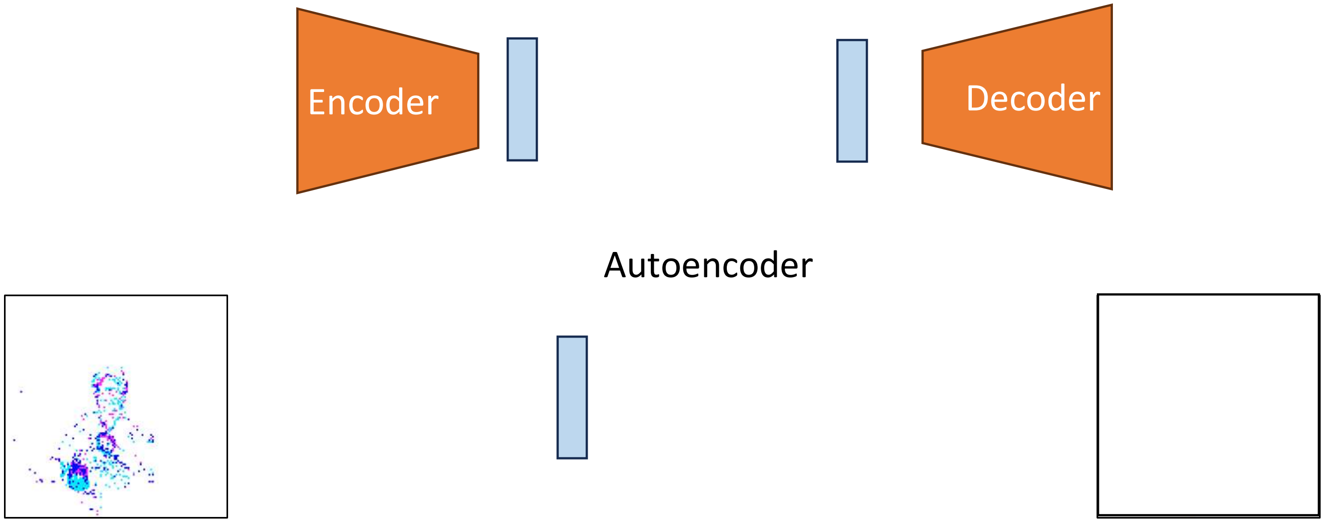
Prompt texts and videos copied from https://research.nvidia.com/labs/toronto-ai/VideoLDM/, https://lumiere-video.github.io/, and https://openai.com/sora.
Publications: [1]Brooks, Tim et al. “Video generation models as world simulators”, (2024), and [2] Bar-Tal, Omer, et al. "Lumiere: A space-time diffusion model for video generation." (2024), and [3] Blattmann, Andreas, et al. "Align your latents: High-resolution video synthesis with latent diffusion models." 2023, and [4] Luo, Zhengxiong, et al. "VideoFusion: Decomposed Diffusion Models for High-Quality Video Generation." 2023.

Text2Events – Latent Diffusion Model (LDM)



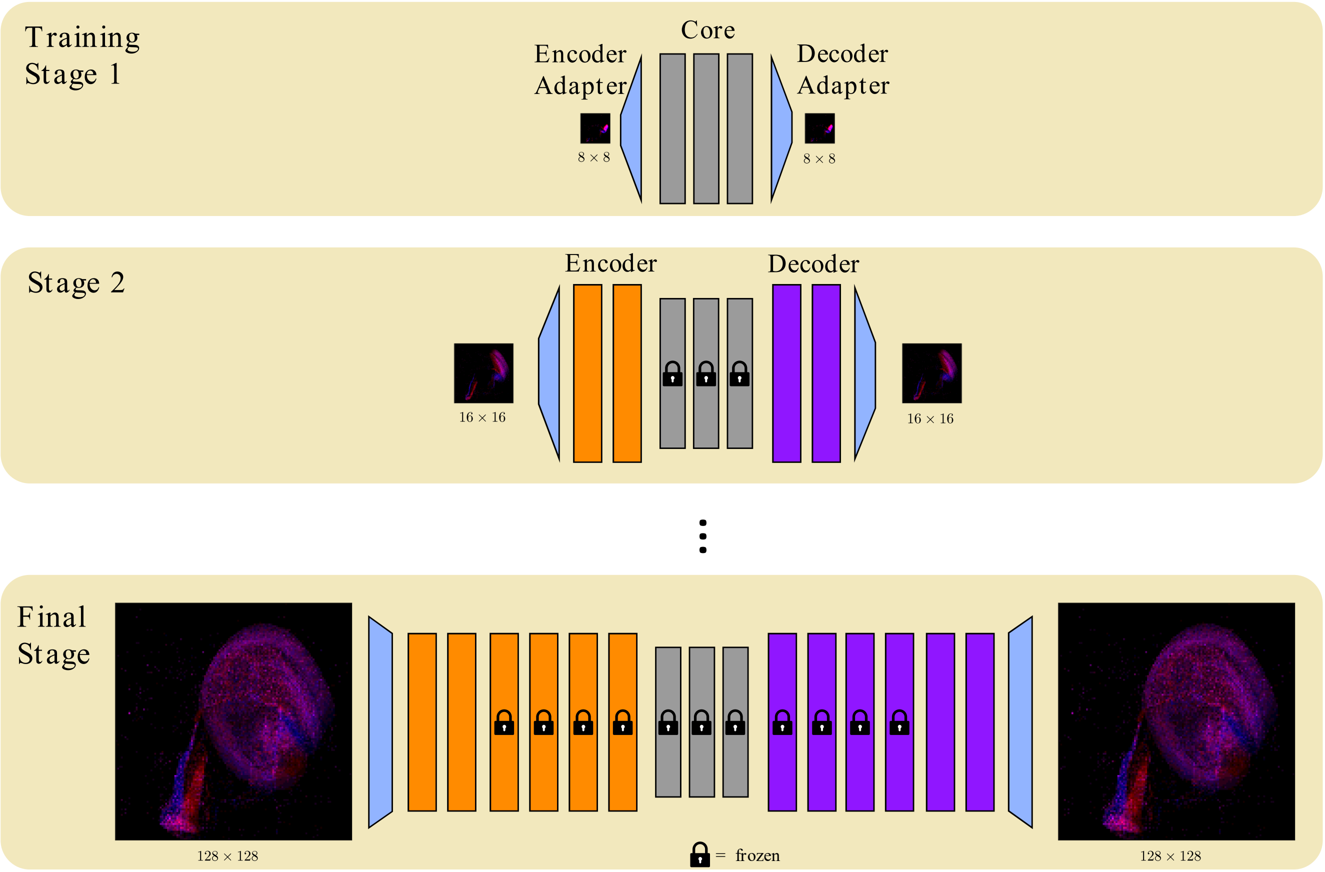
Ott et al, NICE Conf, SD, 23.04.2024

Text2Events – Autoencoder



autoencoder training collapse

Text2Events - Autoencoder

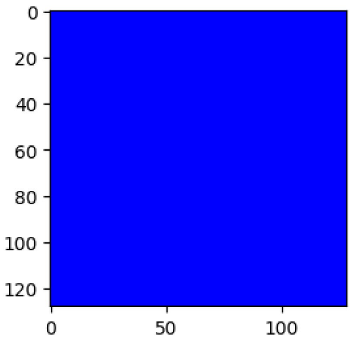


Ott et al, NICE Conf, SD, 23.04.2024

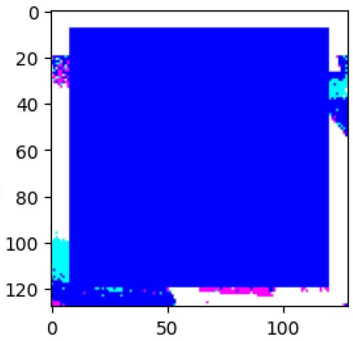
Text2Events – Warm-Up

Epoch

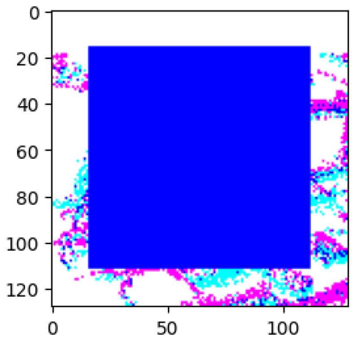
1



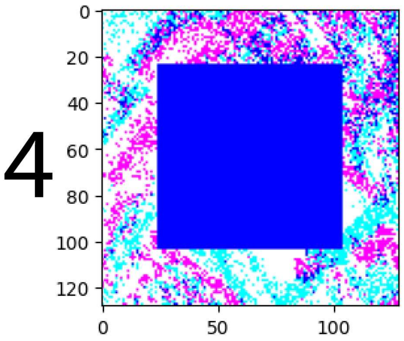
2



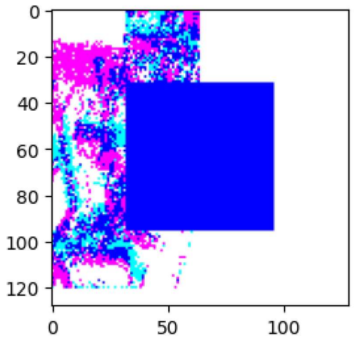
3



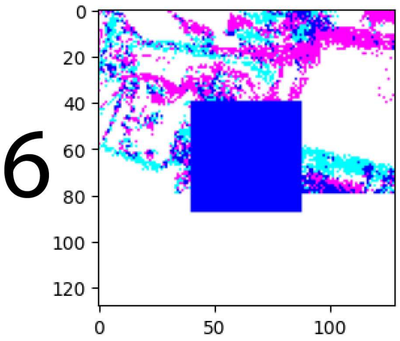
4



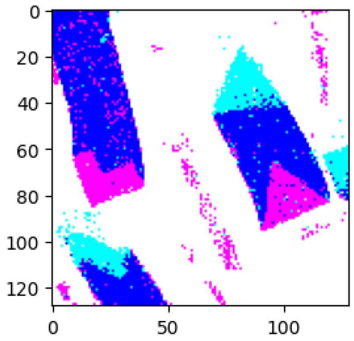
5



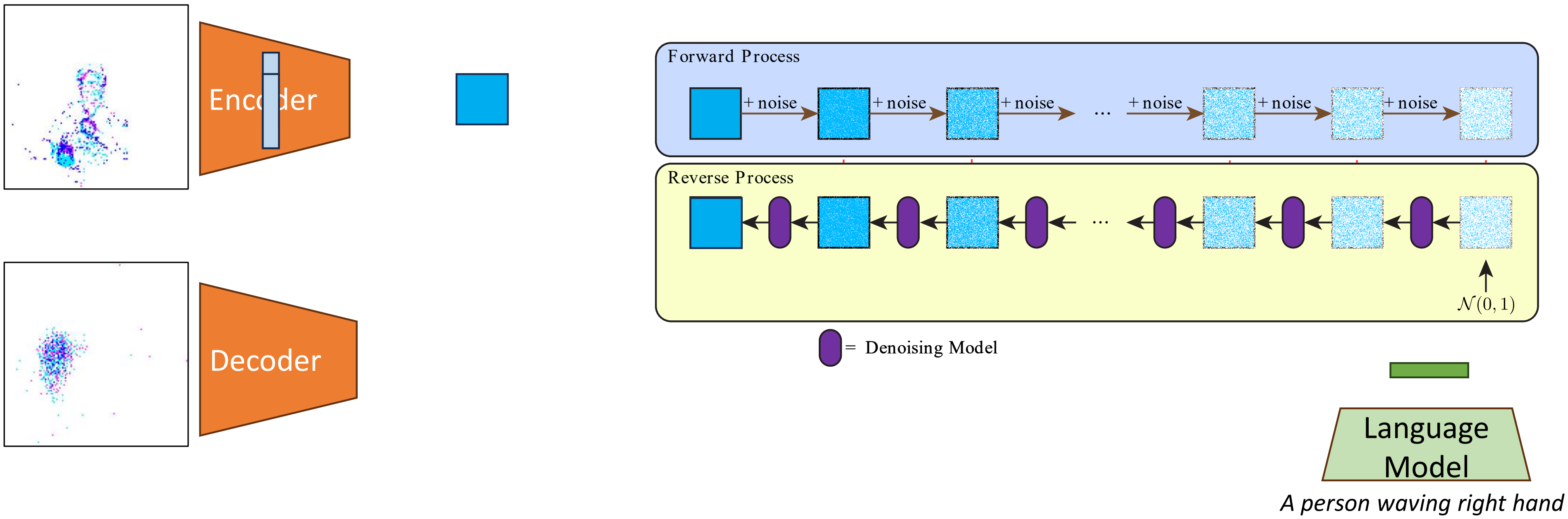
6



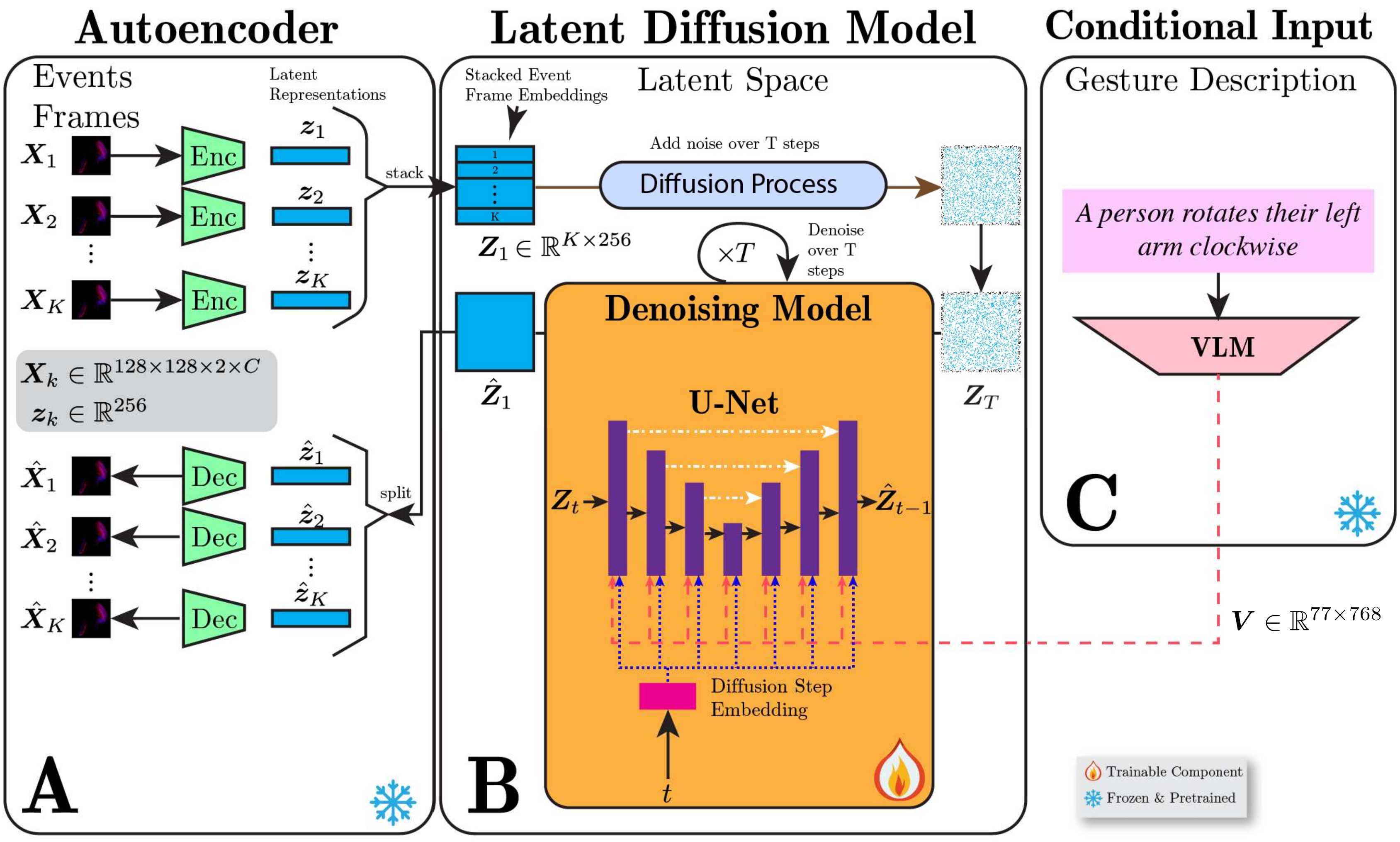
7



Text2Events – Latent Diffusion Model

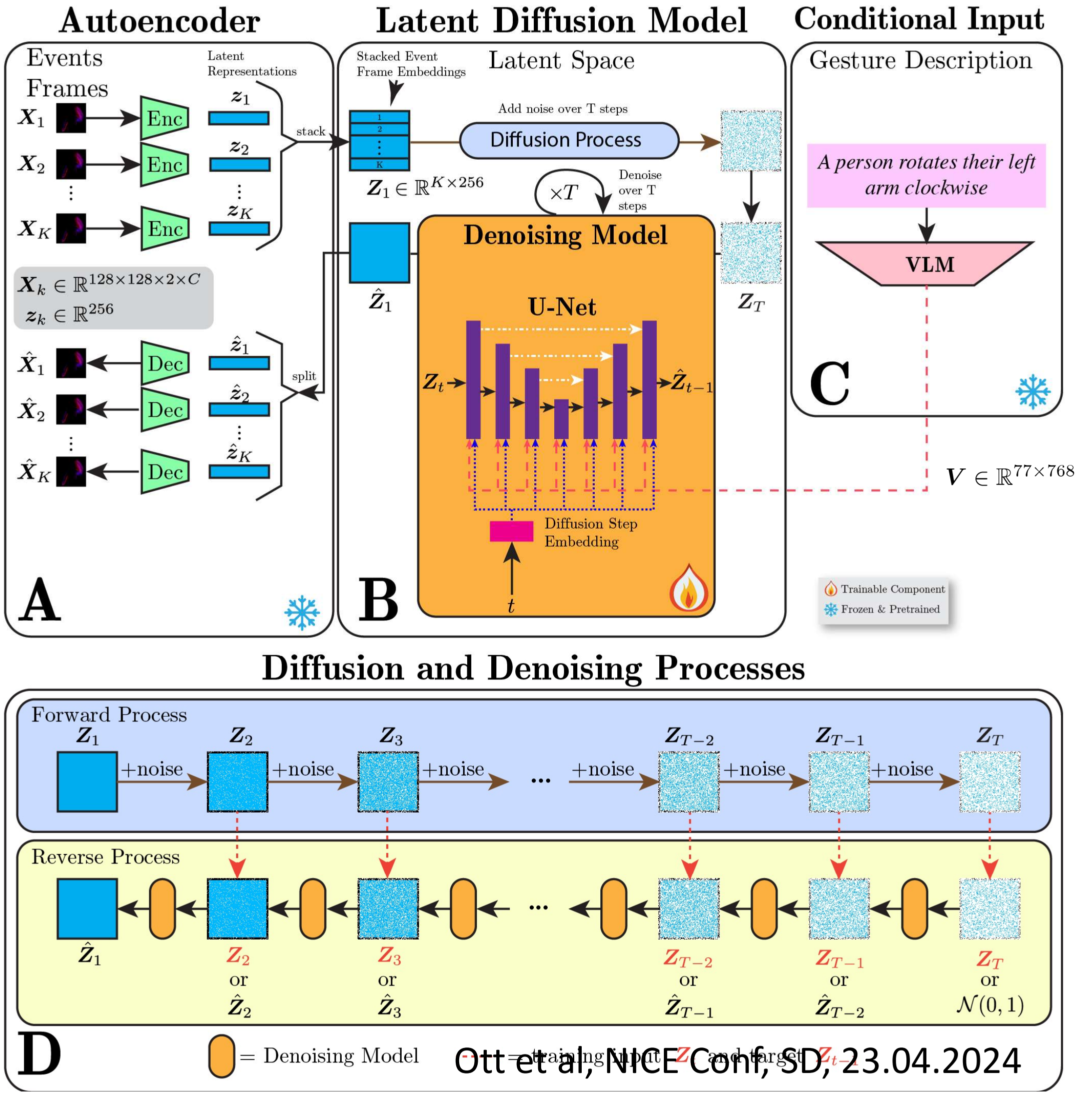


Text2Events – Full Model



Ott et al, NICE Conf, SD, 23.04.2024

Text2Events – Full Model


Autoencoder
Latent Diffusion Model
Conditional Input
Events
Frames
Latent Representations
X_1 Enc z_1
X_2 Enc z_2
X_K Enc z_K
stack
$X_k \in \mathbb{R}^{128\times128\times2\times C}$
$z_k \in \mathbb{R}^{256}$
$\hat{X}_1$ Dec $\hat{z}_1$
$\hat{X}_2$ Dec $\hat{z}_2$
$\hat{X}_K$ Dec $\hat{z}_K$
split
A
Stacked Event Frame Embeddings
Latent Space
Add noise over T steps
Diffusion Process
$Z_1 \in \mathbb{R}^{K\times256}$
$\times T$
Denoise over T steps
Denoising Model
$\hat{Z}_1$
U-Net
Z_t
$\hat{Z}_{t-1}$
Z_T
Diffusion Step Embedding
t
B
Gesture Description
A person rotates their left arm clockwise
VLM
C
$V \in \mathbb{R}^{77\times768}$
Trainable Component
Frozen & Pretrained
Diffusion and Denoising Processes
Forward Process
Z_1 +noise Z_2 +noise Z_3 +noise ... +noise Z_{T-2} +noise Z_{T-1} +noise Z_T
Reverse Process
$\hat{Z}_1$
Z_2 or $\hat{Z}_2$
Z_3 or $\hat{Z}_3$
...
Z_{T-2} or $\hat{Z}_{T-1}$
Z_{T-1} or $\hat{Z}_{T-2}$
Z_T or $\mathcal{N}(0,1)$
D
= Denoising Model
= training input Z_t and target $\hat{Z}_{t-1}$

Ott et al, NICE Conf, SD, 23.04.2024

Parameter Count

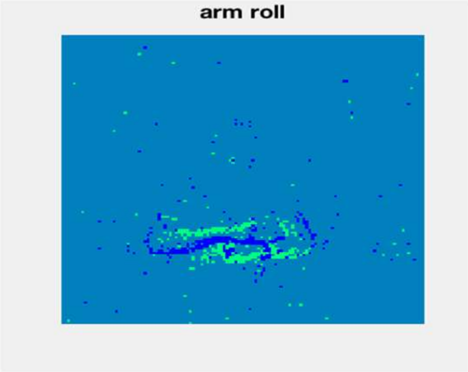
Autoencoder:
3.3 million

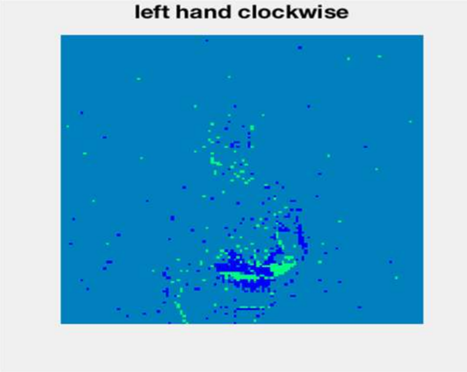
VLM:
12.3 million

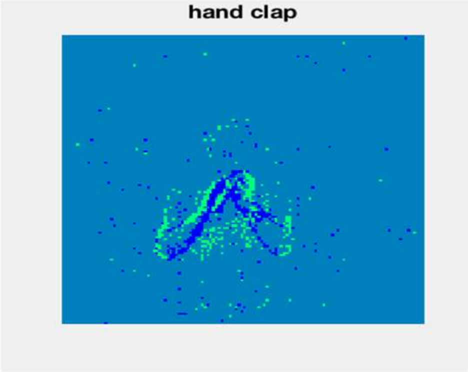
LDM:
120.8 million

Text2Events – Experiments

DVS 128 Dataset







a person is rotating the left arm clockwise

someone is rotating the left arm clockwise

this person is rotating their left arm clockwise

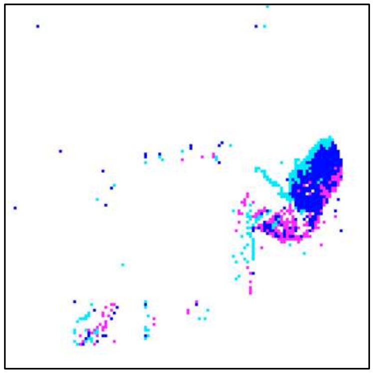
→ **36 new labels per class**

Images: Amir, Arnon, et al. "A low power, fully event-based gesture recognition system." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,* 2017.

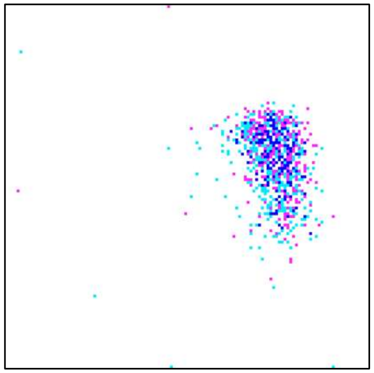
Text2Events - Results

Real

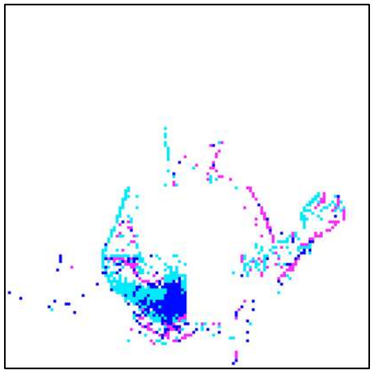
Generated



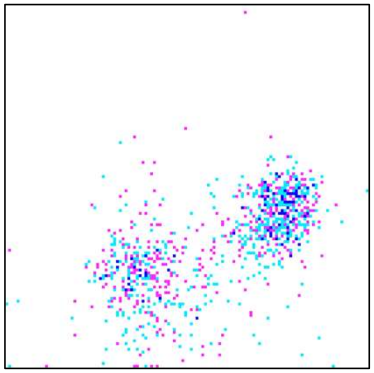
left arm clockwise



'front view of someone rotating their left arm clockwise'

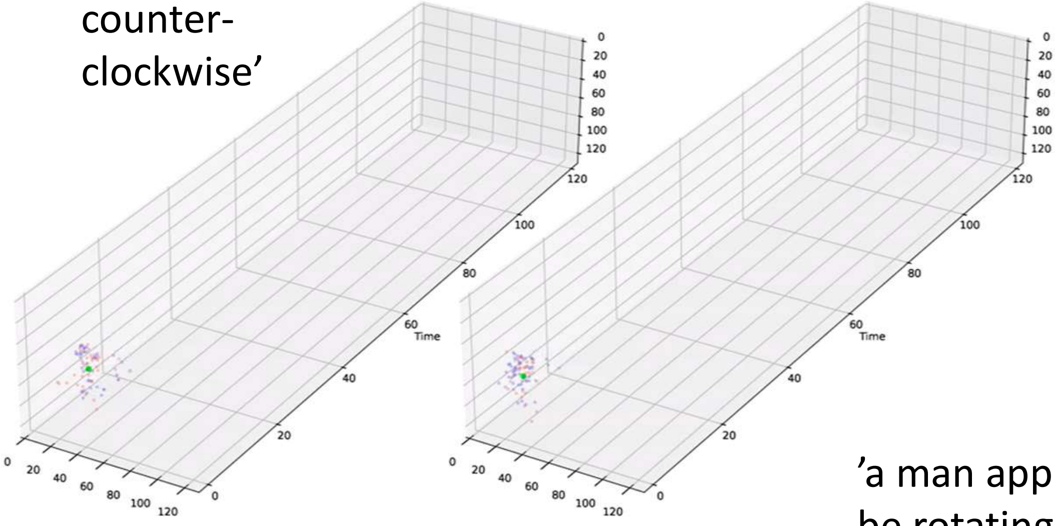


air guitar



'front view of someone playing air guitar'

’right arm counter-clockwise’



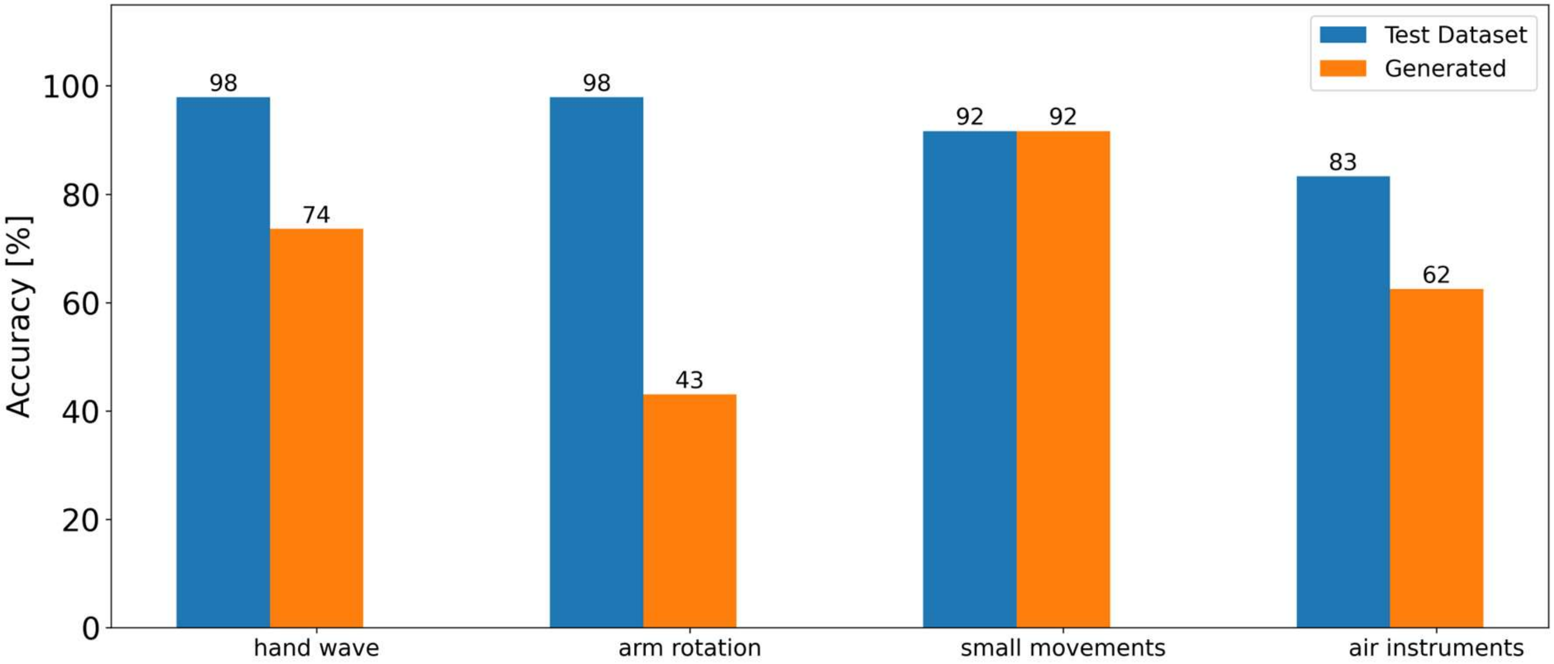
Real

Generated

’a man appears to be rotating his right arm counter-clockwise’

Ott et al, NICE Conf, SD, 23.04.2024

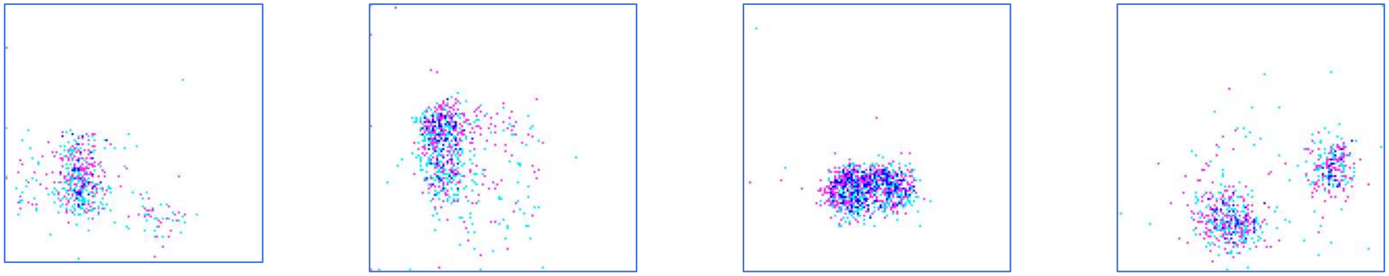
Text2Events - Results



Test dataset: real event sequences

Generated: Event sequences generated by model

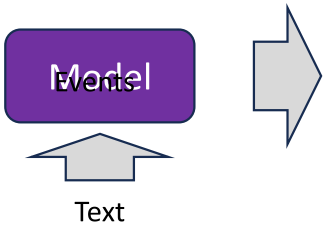
Generated examples

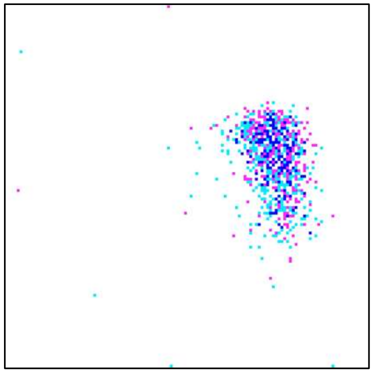


Ott et al, NICE Conf, SD, 23.04.2024

Text2Events – Summary

- First Text-To-Events model: Method of synthesizing vision event datasets using a latent diffusion model
- Method generates gesture event streams from text prompts instead of using a static text-to-events approach.
- Proposed a progressive training method for the autoencoder to encode from and decode to event frames

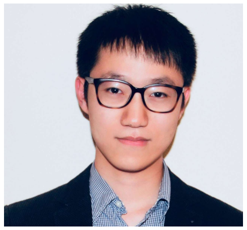




Text2Events - Outlook

- Further improvements can be done, e.g. by combining intensity frames and events during training such as the frames and DVS event outputs of the DAVIS event camera
- Training the model to capture the behavior of each physical camera
- Universal text-to-events pipeline – can be applied to any sensor that produces events

Joachim Ott



Zuowen Wang

FNSNF
Fonds national suisse
Schweizerischer Nationalfonds
Fondo nazionale svizzero
Swiss National Science Foundation

Horizon 2020
European Union Funding
for Research & Innovation

ETH zürich

https://sensors.ini.ch/

sensors
Institute of Neuroinformatics Zurich

University of Zurich UZH

Ott et al, NICE Conf, SD, 23.04.2024