EVENT GENERATIVE COMPRESSION

Anonymous authors
Paper under double-blind review

ABSTRACT

Event cameras offer high temporal resolution and dynamic range. However, in dynamic scenes, they can generate event streams with data rates more than two orders of magnitude higher than conventional video, making transmission a major bottleneck for remote perception. While existing approaches primarily aim to preserve the event stream itself, we argue that when the downstream consumer only needs visual content, compression should instead preserve information useful for visual reconstruction and perception, rather than events themselves. We propose *Event Generative Compression* (EGC), which encodes each temporal window into a compact set of discrete symbols and uses a pretrained video diffusion model to recover the missing visual information. EGC learns an event-to-generation interface through latent alignment, discrete quantization, entropy coding, and generative adaptation, while reusing generated history for continuous reconstruction without transmitting reference frames. On DSEC and BS-ERGB, EGC achieves compression ratios of approximately 5,000–8,000 relative to losslessly coded event streams while retaining information for both temporally coherent video reconstruction and downstream tasks. The same low-rate messages support semantic segmentation, optical flow, and depth estimation, demonstrating the effectiveness of EGC in preserving useful visual information under ultra-low bitrate compression.

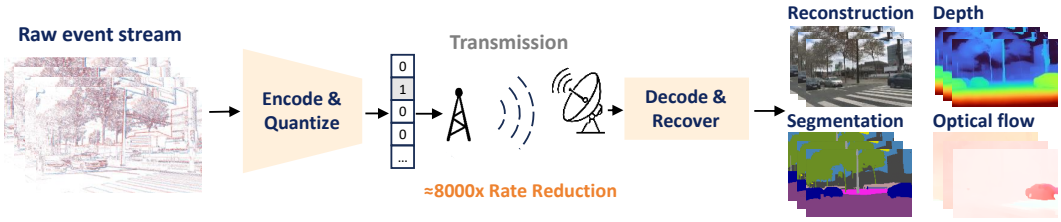

Figure 1: We propose event generative compression (EGC), which encodes events into a compact message for RGB reconstruction and downstream perception including segmentation, optical flow, and depth estimation. On DSEC, EGC achieves 0.004283 bpp, about 1/8,000 of the 35.30 bpp full-event rate, while largely preserving reconstruction and task performance.

1 INTRODUCTION

Event cameras record per-pixel brightness changes asynchronously and offer high temporal resolution, high dynamic range, and low latency (Gallego et al., 2022). The sparse sensing mechanism records only brightness changes and thus avoids the redundant measurements of frame-based cameras. This makes event cameras particularly attractive for capturing fast and dynamic scenes. However, sparse triggering does not necessarily imply a low data rate. Rapid motion and rich textures can repeatedly trigger the same pixels, while every event carries its location, timestamp, and polarity. On DSEC (Gehrig et al., 2021a), the raw event stream averages about 826 Mbps at 64 bits per event, and still 210 Mbps after general-purpose lossless coding, whereas H.264 encodes the co-recorded RGB video at only 2–4 Mbps. For remote perception, fleet-scale data collection, and onboard storage, transmitting and storing events can therefore become a major bottleneck.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Existing approaches reduce this cost through event downsampling (Ghosh et al., 2025) or lossy event coding (Rezaee et al., 2026). However, they still aim to preserve the event stream itself. For visual reconstruction and downstream perception, this objective is unnecessarily restrictive. A human operator needs visual content rather than individual events, while perception models need task-relevant visual features. This opens a different possibility: *events can be compressed much more aggressively if the receiver can recover the missing information without compromising the downstream task*.

The key question is therefore *how to recover visual information from an ultra-low-rate event message*. We propose to utilize rich priors provided by large pretrained video generative models. Specifically, we transmit a compact event message to guide reconstruction and leverage the video prior at the receiver to recover missing visual content. We call this framework *Event Generative Compression* (EGC), as illustrated in Figure 1.

However, this introduces a new challenge: *the event signal and the video generator live in fundamentally different modalities*. The event message must be transformed into a representation that the generator can use effectively, while remaining discrete and compact enough for ultra-low-rate transmission. To solve the problem, we propose a three-stage training framework. First, we align event representations with the latent space of a pretrained video model. Second, we learn a discrete message representation using finite scalar quantization (Mentzer et al., 2024) and entropy coding with a 3D checkerboard context model (He et al., 2021). Third, we adapt the video generator to the compressed event condition through residual flow matching. Generation starts from the decoded event condition rather than pure noise, allowing the generator to use the transmitted information while synthesizing the missing visual content. This training procedure adapts both the event representation and the generative prior without modifying the generator architecture.

Ultra-low-rate transmission also makes temporal consistency challenging. Since no reference frames are transmitted, independently reconstructing each window can lead to discontinuities. EGC therefore reuses the tail of the previously generated output as temporal context for the next window. At inference time, the sender observes only events and never accesses RGB frames. The receiver reconstructs the video entirely from the transmitted messages and its pretrained visual prior.

We evaluate EGC on two real-world event datasets, DSEC (Gehrig et al., 2021a) and BS-ERGB (Tulyakov et al., 2022). EGC achieves ultra-low bitrate while maintaining competitive reconstruction quality. On DSEC, it operates at 0.0043 bpp, achieving a compression ratio of approximately 8,000 relative to the losslessly coded event stream while maintaining comparable visual reconstructions. The same messages further support semantic segmentation, depth, and optical flow estimation, validating its effectiveness on both visual and machine perceptions.

Our contributions are as follows:

- We introduce a new formulation of event compression for visual reconstruction and downstream perception. Instead of preserving the event stream itself, EGC transmits a compact event message and recovers missing visual information from a pretrained video prior.
- We develop a three-stage event-to-generation framework that aligns events with the video latent space, learns an ultra-low-rate discrete message, and adapts a pretrained video generator through residual flow matching. Generated history further enables continuous video reconstruction without transmitting reference frames.
- We systematically evaluate EGC on two real-world event datasets. Experiments show that EGC maintains effective visual reconstruction at ultra-low rates while supporting downstream tasks including semantic segmentation, optical flow, and depth estimation.

2 Related Work

Event-based video reconstruction. Event-to-video reconstruction recovers intensity frames from asynchronous brightness changes using recurrent or transformer-based temporal models (Rebecq et al., 2019; Stoffregen et al., 2020; Weng et al., 2021; Cadena et al., 2021), while EVREAL (Ercan et al., 2023) provides a unified benchmark for these methods. More recently, generative priors have been introduced to improve reconstruction quality. EvDiff (Li et al., 2026) uses a one-step diffusion model to reconstruct color video, and LongE2V (Fan et al., 2026) uses a video diffusion prior for

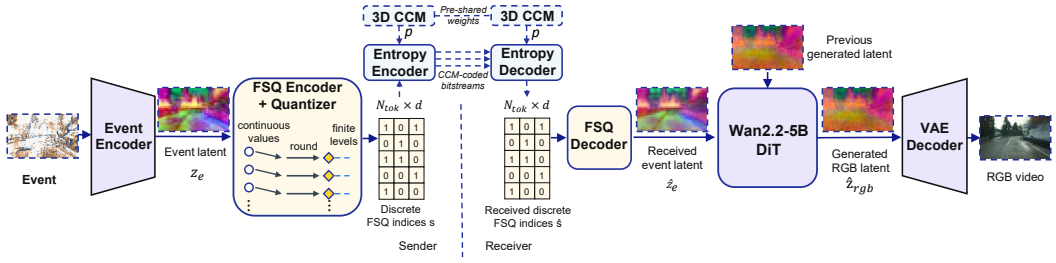

Figure 2: EGC sender and receiver. FSQ quantizes event latents z_e into indices s, and a shared 3D checkerboard context model (CCM) supplies entropy-coding probabilities. The receiver recovers the indices exactly and decodes them into condition $\hat{z}_e$. The adapted Wan2.2-5B (Wan et al., 2025) DiT combines this condition with generated history to produce $\hat{z}_{\mathrm{rgb}}$, after which the VAE decodes into RGB video.

long-horizon reconstruction. These methods assume that the complete event stream is available at the reconstruction site. EGC instead treats reconstruction as decoding from a rate-constrained message. Both the event representation and the generative decoder are adapted, allowing the generator to recover visual content from far fewer transmitted observations.

Event data reduction and compression. Prior work reduces event data through spatial or temporal downsampling (Ghosh et al., 2025) or lossy coding. Examples include compressing event histograms with JPEG 2000 and event point clouds with G-PCC (Rezaee et al., 2026). These approaches treat the event stream as the signal to preserve and introduce distortion in the event domain. EGC follows a different objective. It does not reconstruct or preserve the event stream. Instead, EGC learns a compact event message and adapts a generative receiver to reconstruct visual content from it.

Learned and generative compression. Learned image codecs model latent representations with hyperpriors and context models for entropy coding (Ballé et al., 2018; Minnen et al., 2018; He et al., 2021), and extend to video through temporal context modeling (Li et al., 2021). At low bitrates, generative decoders can improve perceptual quality through adversarial training (Agustsson et al., 2019; Mentzer et al., 2020) or diffusion models (Yang & Mandt, 2023). This benefit is closely related to the rate–distortion–perception trade-off (Blau & Michaeli, 2019). Generative Latent Coding (Qi et al., 2025) further explores compression in the latent space of a generative VQ-VAE. These codecs, however, encode the same signal that the decoder reconstructs. Their generative prior mainly restores detail lost during compression. ENCORE (Ye et al., 2026) uses events as side information for motion estimation, but still compresses RGB video. EGC is fundamentally cross-modal: the sender observes only events, while the receiver reconstructs RGB video that is never available to the sender. The generative prior therefore provides appearance that is absent from the source signal, rather than merely recovering details lost during compression.

3 Method

3.1 Problem Formulation and System Overview

Let the event stream be $\mathcal{E} = \{(x_i, y_i, t_i, p_i)\}_{i=1}^{N}$, where N is the number of events, and (x_i, y_i), t_i, and p_i denote pixel coordinates, timestamp, and polarity, respectively. Let $\mathcal{E}_k$ denote the events in window k. Encoding and reconstruction are defined as

$$b_k = C_\phi(\mathcal{E}_k), \qquad \widehat{X}_k = G_\theta(b_k, h_k; \xi_k), \tag{1}$$

where C_ϕ and G_θ are the sender and receiver with model parameters ϕ and θ, respectively. Here, b_k is the transmitted bitstream, $\widehat{X}_k$ the reconstructed RGB video, h_k the previously generated latent history at the receiver, and ξ_k the sampling noise. During inference, the sender takes only events as input. The receiver initializes the first window's history to zero. RGB video is used only for training supervision and offline evaluation.

3.2 Event Representation and Compression

Event representation. For window k, we convert the events into a ternary Event Status Representation (ESR) $S_k \in \{-1, 0, 1\}^{T \times H \times W}$, defined over the T output-frame intervals of the window. Each entry records the polarity of the most recent event at that pixel within the corresponding frame interval, or zero if no event occurs.

Sender. An event encoder E_ϕ maps S_k to an event latent $z_{e,k} \in \mathbb{R}^{48 \times T' \times H' \times W'}$, which has the same shape as the RGB latent of the corresponding clip. A trainable transform A_ψ then reduces $z_{e,k}$ to d channels and subsamples it temporally by a stride τ, yielding $N_{\text{tok}} = (T'/\tau)\, H'W'$ coding positions with d components each. Finite scalar quantization (FSQ) (Mentzer et al., 2024) rounds the j-th component to one of L_j levels, producing the discrete indices s_k:

$$z_{e,k} = E_\phi(S_k), \qquad s_k = \mathrm{FSQ}(A_\psi(z_{e,k})) \in \textstyle\prod_{j=1}^{d} \{0, \ldots, L_j - 1\}^{N_{\text{tok}}}. \tag{2}$$

Each position costs $\sum_j \log_2 L_j$ bits. For example, levels $[2, 2, 2]$ give eight combinations, or 3 bits per position. A stride of $\tau = 4$ reduces the number of positions, and hence the rate, fourfold relative to $\tau = 1$.

Entropy coding. The indices s_k are further entropy-coded into the bitstream b_k. A 3D checkerboard context model (CCM) (He et al., 2021), shared by the sender and receiver, estimates the probability $p_\eta(s_k)$ used for arithmetic coding:

$$b_k = \mathrm{EntEnc}(s_k;\, p_\eta), \qquad \hat{s}_k = \mathrm{EntDec}(b_k;\, p_\eta) = s_k. \tag{3}$$

The CCM partitions the N_{tok} positions into anchors and non-anchors in a spatiotemporal checkerboard pattern, predicting anchor probabilities first and non-anchor probabilities from the decoded anchors via masked 3D convolutions. The d components at each position are modeled independently. It is trained on frozen indices by minimizing $\mathcal{L}_{\text{CCM}} = -\mathbb{E}\big[\log_2 p_\eta(s_k)\big]$, which estimates the bits needed to encode the quantized symbols. The CCM reduces the bitrate through context modeling while lossless decoding preserves the generator's condition.

Receiver. Given the decoded indices $\hat{s}_k$, $\mathrm{Dequant}$ maps them to their quantized values, and a trainable transform B_ψ restores the original latent shape, producing the event condition for the generator:

$$\hat{z}_{e,k} = B_\psi(\mathrm{Dequant}(\hat{s}_k)) \in \mathbb{R}^{48 \times T' \times H' \times W'}. \tag{4}$$

Operating points. Each operating point is determined by the quantization levels and the temporal stride, and is trained as a separate model. We name configurations by their levels and stride: `222-ts1` denotes three binary components at $\tau = 1$, and `2x12-ts4` denotes twelve binary components at $\tau = 4$. We refer to `222-ts1` as the base configuration.

3.3 Generative Receiver

The decoded condition $\hat{z}_e$ lies in the generator's latent space and already specifies the structure and motion of the scene. What it lacks is appearance, such as color and fine texture, which events do not record or quantization discards. Rather than generating the target RGB latent z_c from pure noise, the receiver therefore refines $\hat{z}_e$ into z_c. Specifically, we apply flow matching (Lipman et al., 2023) to the residual $x_0 = z_c - \hat{z}_e$,

$$x_t = (1 - \sigma_t)\, x_0 + \sigma_t\, \epsilon, \qquad \epsilon \sim \mathcal{N}(0, I), \tag{5}$$

and feed $\hat{z}_e + x_t$ to the pretrained Diffusion Transformer (DiT). Sampling thus starts from $\hat{z}_e + \epsilon$ and ends at $\hat{z}_{\text{rgb}} = \hat{z}_e + \hat{x}_0$, after which the VAE decoder maps to the output video $\widehat{X} = D(\hat{z}_{\text{rgb}})$. Unlike residual diffusion for image restoration (Yue et al., 2023; Liu et al., 2024), the starting point here is a compressed cross-modal condition. Because the condition enters by addition, the DiT requires no architecture changes.

To keep the generated frames continuous without transmitting reference frames, each window is conditioned on the last T_h latent frames of the previous output, $h_k = \mathrm{Tail}_{T_h}(\hat{z}_{\text{rgb},k-1})$, with zero history for the first window. The history is concatenated with the current input along time as $[h_k;\, \hat{z}_e + x_t]$, remains noise-free, and is not supervised.

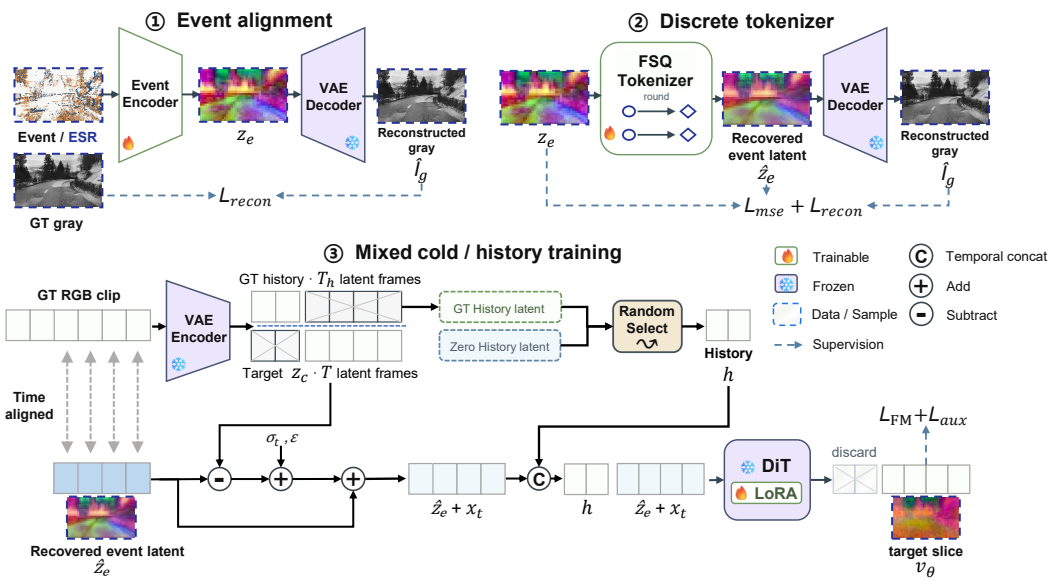

Figure 3: Three-stage training of EGC. (1) The event encoder is aligned with the video latent space through grayscale reconstruction. (2) The FSQ transforms learn a discrete message that preserves the aligned latent. (3) LoRA adapters teach the DiT to refine the recovered event condition into the RGB latent, with zero or perturbed ground-truth history. Modules trained in previous stages are frozen.

3.4 Three-Stage Training

We train the system following the data flow in Figure 3. Each stage freezes the components learned in the previous stages. This ensures that the message is fixed to better fit the entropy model and the generator. RGB frames are used only at training phase.

Stage 1: Event latent alignment. We train the event encoder E_ϕ to produce a latent representation that can be decoded into grayscale frames by the frozen VAE decoder: $D(z_e) \approx I_g$. The encoder is optimized with MSE and LPIPS (Zhang et al., 2018) losses. We use grayscale targets because events contain no color information. Color is recovered later by the generative prior.

Stage 2: Discrete message learning. We freeze E_ϕ and D, and train the transforms A_ψ and B_ψ to learn a compact discrete representation. The recovered latent $\hat{z}_e$ is encouraged to remain close to z_e and to preserve the grayscale reconstruction. We additionally apply temporal and quantization regularization. After obtaining the frozen indices, we train the CCM to model their distribution for entropy coding.

Stage 3: Generative adaptation. We freeze the compressor and the DiT backbone, and train LoRA adapters (Hu et al., 2022) to adapt the generator to the compressed event condition. Training objective includes the flow matching loss $\mathcal{L}_{\text{FM}}$ with an auxiliary perceptual and color loss $\mathcal{L}_{\text{aux}}$. During training, each clip uses randomly selected zero history representing the first window, or perturbed ground-truth history, simulating the imperfect generated history encountered at deployment. Full objectives and hyperparameters are provided in Supplementary Section A.2.

4 Experiments

4.1 Experimental Setup

Data and training. We evaluate EGC for RGB reconstruction and downstream tasks. For downstream evaluation, we use frozen public task models on reconstructed RGB video and lightweight task heads trained on multiscale features from the frozen VAE decoder. The latter operates directly on generated latents without RGB decoding. RGB reconstruction uses the DSEC test sequences (Gehrig

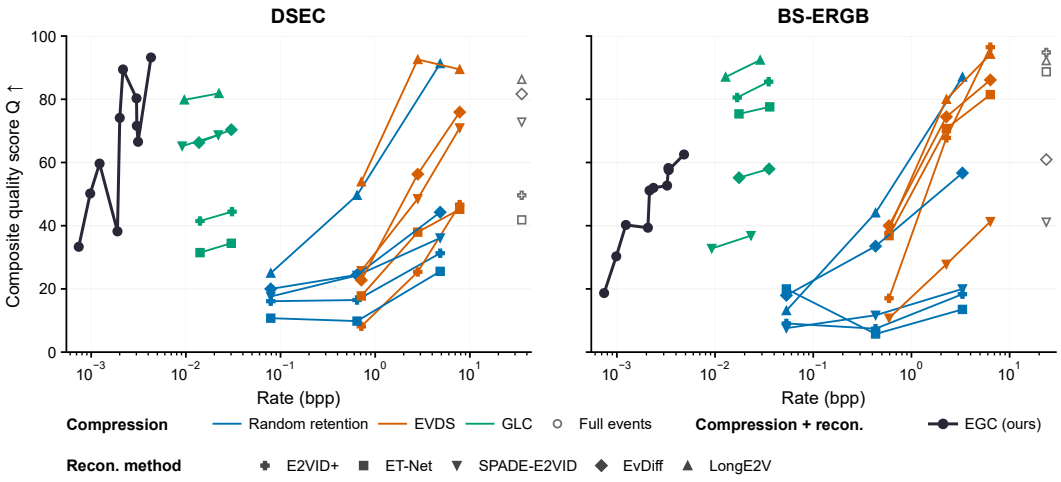

Figure 4: Grayscale quality versus bitrate on DSEC and BS-ERGB. Points denote event reconstruction methods, and colors denote compression methods. Higher Q indicates better relative quality within each dataset.

et al., 2021a) and BS-ERGB validation sequences (Tulyakov et al., 2022) from EVREAL (Ercan et al., 2023). On DSEC, we evaluate segmentation, optical flow, and depth estimation. Training combines simulated events from BDD100K (Yu et al., 2020) with real event-image pairs from BS-ERGB and DSEC, using disjoint training and evaluation sequences. Generative adaptation uses 121-frame clips at 384×640 with rank-16 LoRA. We use the pretrained Wan2.2-5B (Wan et al., 2025) as the generator, and the compressor and DiT weights remain frozen. All models use 60k-step checkpoints and 25-step flow-matching inference. Unless otherwise stated, we use the base configuration `222-ts1` for controlled comparisons.

Baselines. We compare EGC with E2VID+ (Stoffregen et al., 2020), ET-Net (Weng et al., 2021), SPADE-E2VID (Cadena et al., 2021), EvDiff (Li et al., 2026), and LongE2V (Fan et al., 2026), combined with different compression approaches including random event retention, EVDS temporal downsampling (Ghosh et al., 2025), or subsequent GLC compression (Qi et al., 2025). For downstream tasks, we evaluate frozen task models on reconstructed RGB. We also compare against task-specific event-based models: TESPEC (Mohammadi et al., 2025) for semantic segmentation, E-RAFT (Gehrig et al., 2021b) for optical flow, and RecDAv2 (Bartolomei et al., 2025) for depth estimation.

Metrics and bitrate. Reconstruction methods are evaluated at the same ground-truth timestamps and over the same valid image regions. We report framewise metrics averaged equally across sequences. We summarize reconstruction quality with Q, a weighted rank score of LPIPS (Zhang et al., 2018), MS-SSIM (Wang et al., 2003), tOF, and tLP, with weights 0.50/0.20/0.20/0.10. Q captures both framewise fidelity and temporal consistency relative to the ground truth. Bitrates are reported in bpp, using measured CCM bitstream bytes for EGC, measured transmitted bytes for event baselines, and likelihood estimates for GLC.

4.2 Comparison with Existing Methods

Video reconstruction. We compare video reconstruction performances in Figure 4. EGC retains visual content after substantial rate reduction. On DSEC, `2x12-ts1` achieves $Q = 93.2$ at 0.004283 bpp, approximately 8,000-fold compression relative to the 35.30 bpp full-event channel. Its LPIPS of 0.4699 is close to full-event EvDiff's 0.4591. On BS-ERGB, the same configuration achieves approximately 5,000-fold compression at 0.004830 bpp, with $Q = 62.5$. Although EGC does not surpass LongE2V and E2VID+ compressed by GLC, its performance remains comparable to SPADE-E2VID and EvDiff while operating at ultra-low bitrates.

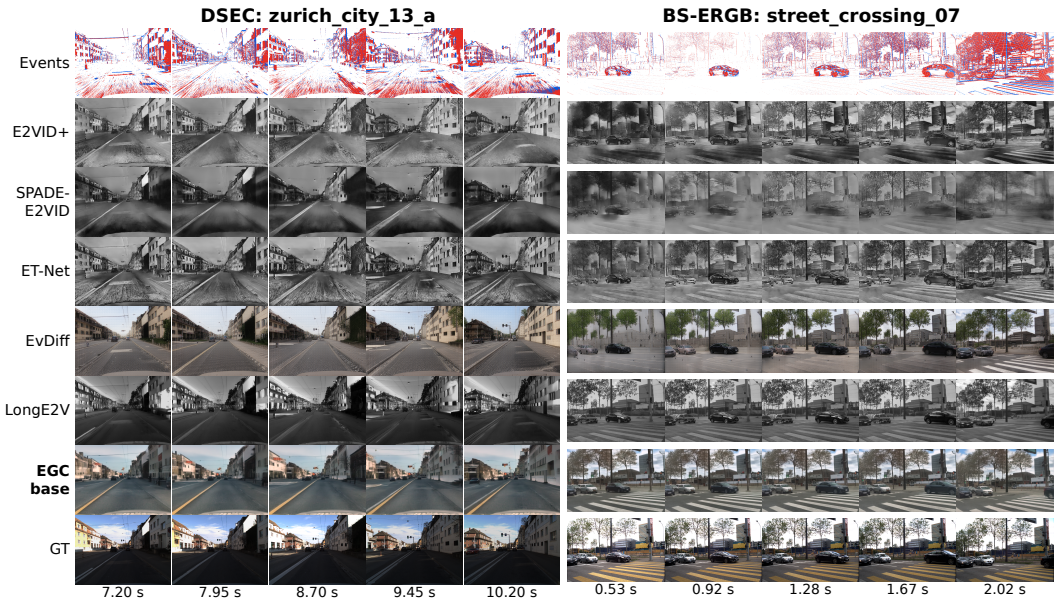

Figure 5: Temporally matched reconstructions on DSEC (left) and BS-ERGB (right). Other reconstruction methods use the complete event stream without event reduction, while EGC base uses compressed event messages. Columns follow time, with timestamps below GT.

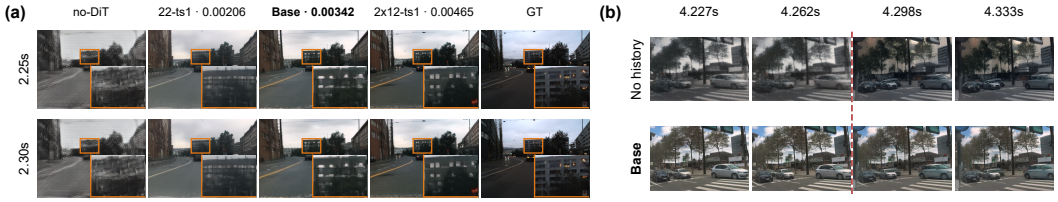

Figure 6: Receiver, quantization, and history comparisons. (a) Two matched DSEC times comparing quantization settings and direct VAE decoding without DiT (No DiT). Rates are in bpp. (b) BS-ERGB with and without history; the dashed line marks the window boundary.

Figure 5 qualitatively compares the reconstruction on two datasets. In the early BS-ERGB frames, EvDiff exhibits a washed-out appearance that gradually clears during initialization, along with noticeable flicker across consecutive frames. LongE2V incorrectly reconstructs the structure of the car on the left. In contrast, EGC recovers recognizable scenes and their temporal evolution from highly compressed messages. These examples also show that access to the full event stream does not guarantee correct recovery of every scene structure.

Downstream tasks. In Figure 7, we evaluate EGC on semantic segmentation, optical flow, and depth estimation. EGC shows a clear advantage in optical flow and achieves comparable performance on segmentation and depth estimation at extremely low bitrates. This demonstrates that EGC can preserve task-relevant visual information across different perception tasks under extreme compression.

Computational cost. At 640×464 output resolution running on H200, EGC's sender requires 39.0 ms/frame, 1.14 TFLOPs/frame, and 150.3M parameters, with latency close to E2VID+GLC's 38.0 ms/frame (Table 1a). Since EvDiff and LongE2V rely on computationally intensive generative priors to reconstruct color video, combining them with GLC requires both the reconstruction and compression to be performed at the sender, resulting in sender-side processing times of 165.9 and 1272.1 ms/frame, respectively. EGC instead encodes events directly and leaves expensive generation to the receiver, requiring 142.2 ms/frame and 29.20 TFLOPs/frame at the receiver. This design

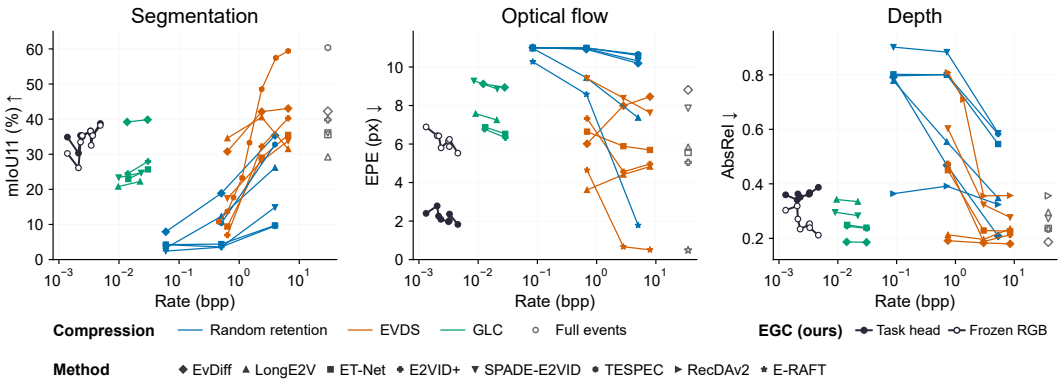

Figure 7: Downstream rate–performance on DSEC: segmentation (mIoU, %↑), optical flow (EPE, pixels↓), and depth estimation (AbsRel↓).

Table 1: Computation and ablation results. (a) Sender and receiver computational costs. (b) Effect of generated history on temporal continuity. (c) Downstream performance with different input representations.

(a) Computational cost

	Sender			Receiver	
Method	Time↓ (ms per frame)	TFLOPs↓ (/frame)	Params (M)	Time↓ (ms per frame)	TFLOPs↓ (/frame)
E2VID+	–	–	–	4.1	0.27
E2VID+ → GLC	38.0	1.07	151.5	25.8	1.36
ET-Net	–	–	–	45.1	0.84
ET-Net → GLC	76.9	1.65	162.9	25.8	1.36
EvDiff	–	–	–	141.0	6.15
EvDiff → GLC	165.9	6.96	2364.3	25.8	1.36
LongE2V	–	–	–	1021.9	122.48
LongE2V → GLC	1272.1	123.28	10689.3	25.8	1.36
EGC + CCM	39.0	1.14	150.3	142.2	29.20

(b) Generated history

Metric	None	Gen.
DSEC seam↓	6.35	2.59
BS-ERGB seam↓	3.33	1.69
BS-ERGB b slope → 0	−76.8	−8.6

(c) Task readout

Input	mIoU11↑ (%)	MAE↓ (px)	EPE↓ (px)
Real video	51.83	3.069	1.349
Cont. + DiT	42.62	3.443	1.556
base	37.26	3.853	2.370
No DiT	34.33	4.412	2.885

substantially reduces the sender-side computational and transmission burden by shifting expensive generation to the receiver.

Table 2: Reconstruction ablations. Each cell reports DSEC / BS-ERGB. (a) Receiver and message conditions. (b) Transmitted representations with a fixed receiver under the full-stream protocol. LPIPS, MS-SSIM, and tOF are evaluated under grayscale image, while FID uses RGB features.

(a) Receiver and message

Condition	LPIPS↓	MS-SSIM↑	tOF↓ (10^{-2})	FID↓
Wrong	0.62/0.74	0.42/0.21	1.42/3.92	89/198
No DiT	0.50/0.53	0.50/0.37	1.22/3.71	–
base	0.51/0.62	0.56/0.41	1.03/3.05	83/165

(b) Transmitted representation

Input	bpp↓ (10^{-3})	LPIPS↓	MS-SSIM↑
base	3.01/3.27	0.51/0.62	0.55/0.40
CRF35	856/875	0.50/0.62	0.56/0.40
CRF51	182/145	0.54/0.65	0.55/0.39

4.3 Ablation Studies

In this ablation, we examine the transmitted representation, generative receiver, message correspondence, history, task readout, and vary quantization capacity across different configurations.

Representation and symbol coding. From Figure 4, it can be observed that increasing the FSQ condition capacity generally improves reconstruction quality on both DSEC and BS-ERGB. In Table 2b, we compare the learned message representation with directly coding the ESR representation using the HEVC-based x265 encoder (CRF35 and CRF51). With the condition extractor and receiver fixed, the learned representation achieves substantially better rate-quality efficiency. For example, the full-stream base achieves LPIPS of 0.5085/0.6202 at 0.003014/0.003269 bpp on DSEC/BS-ERGB, whereas ESR+x265 (CRF35) requires substantially higher bitrates for comparable quality.

Table 3: Measured container rates for the base configuration under the same reconstruction protocol. Both coders preserve the same quantized symbols.

Dataset	zstd (bpp)	CCM (bpp)	Reduction
DSEC	0.003309	0.003033	8.3%
BS-ERGB	0.003499	0.003247	7.2%

We compare CCM with Zstandard (zstd) (Collet & Kucherawy, 2021), a general-purpose lossless compressor, using the same quantized symbols. CCM reduces base container rates by 8.3% on DSEC and 7.2% on BS-ERGB (Table 3). CCM losslessly codes the quantized symbols, reducing transmitted bytes without changing the generative condition.

Generative receiver. The DiT improves MS-SSIM and tOF on both datasets, although grayscale LPIPS increases from 0.4975/0.5278 without DiT to 0.5078/0.6178 with the base model (Table 2a). Figure 6a shows the corresponding visual differences. Without DiT, the outputs are grayscale and exhibit pronounced grid artifacts and blurred facades and windows. DiT restores color, sharper structures, and more temporally consistent results, highlighting the benefit of the generative prior beyond framewise perceptual similarity.

The importance of message. Keeping generated history and per-window noise fixed, we replace the current message with one from a non-overlapping window of the same sequence. Wrong-message LPIPS reaches 0.6235 on DSEC and 0.7407 on BS-ERGB, with MS-SSIM and tOF also degrading (Wrong row in Table 2a). These results show that the generator relies on event cues corresponding to the current window: its pure visual prior cannot substitute for the correct transmitted content.

Generated history. Table 1b shows that reusing generated history substantially improves temporal continuity across windows. Here, the seam ratio measures boundary-frame differences relative to ordinary adjacent-frame differences, and values closer to zero for the b-channel error slope indicate less color drift. It can be seen that with the history, the seam ratio decreases from 6.35/3.33 to 2.59/1.69 on DSEC/BS-ERGB at nearly unchanged bitrates, while the Lab b-channel error slope also indicates improved color stability. Figure 6b further shows a pronounced color jump at the window boundary without history, which is substantially reduced when preceding outputs are reused.

Task-adapted readout. Table 1c compares four representations for downstream tasks: RGB input, full-event input with DiT, the quantized base model, and the quantized model without DiT. All settings use capacity-matched heads, identical task splits, and the same model-selection rules. RGB provides the most complete visual information, while full-event+DiT serves as a continuous-condition reference. Relative to the quantized model without DiT, the base model gains 2.93 segmentation mIoU points, reduces depth-task MAE (measured in disparity space) by 12.7%, and reduces flow EPE by 17.9%. These results show that the generative receiver improves the usefulness of quantized conditions for downstream tasks, not only their visual appearance.

5 Conclusion

We present *Event Generative Compression* (EGC), a new framework for event compression that focuses on preserving visual information rather than the event stream itself. EGC combines a compact discrete event message with a pretrained video generative prior. Its three-stage training framework aligns event representations with the video latent space, learns an ultra-low-rate message, and adapts the generator to use the compressed condition. Reusing generated history further enables continuous reconstruction without transmitting reference frames. Experiments on DSEC and BS-ERGB show that EGC can operate at extremely low bitrates while maintaining competitive visual reconstruction quality. The transmitted messages also support semantic segmentation, depth estimation, and optical flow, demonstrating that the learned representation retains information beyond frame reconstruction.

AI USE STATEMENT

We used generative AI assistants to search for relevant literature; to draft, translate, and edit parts of the manuscript; to write the program used to simulate event data from existing videos; and to create evaluation and figure-generation scripts. The core EGC method, experimental parameter choices, interpretation of results, and the content and overall layout of the figures were determined by the authors. The lead author cross-checked and manually reviewed AI-assisted code, outputs, and text before inclusion. The authors take responsibility for the final manuscript, code, figures, results, and claims.

ETHICS STATEMENT

This study uses the existing BDD100K, BS-ERGB, and DSEC datasets; it involves no new recordings, recruited participants, or human-subject experiments. We acknowledge Prof. Davide Scaramuzza and the University of Zurich as the source of BS-ERGB. We do not redistribute the original datasets. The supplementary presentation video includes generated clips alongside other methods for qualitative comparison, rather than distributing them as standalone videos. The qualitative figures were checked for discernible faces and legible license plates. Because events incompletely specify scene appearance, the generative receiver may produce plausible but inaccurate content. Reconstructed frames should therefore not be treated as documentary evidence or used without independent validation in safety-critical settings.

REFERENCES

Eirikur Agustsson, Michael Tschannen, Fabian Mentzer, Radu Timofte, and Luc Van Gool. Generative adversarial networks for extreme learned image compression. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 221–231, 2019. doi: 10.1109/ICCV.2019.00031.

Johannes Ballé, David Minnen, Saurabh Singh, Sung Jin Hwang, and Nick Johnston. Variational image compression with a scale hyperprior. In *International Conference on Learning Representations*, 2018. URL https://arxiv.org/abs/1802.01436.

Luca Bartolomei, Enrico Mannocci, Fabio Tosi, Matteo Poggi, and Stefano Mattoccia. Depth anyevent: A cross-modal distillation paradigm for event-based monocular depth estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 19669–19678, October 2025.

Yochai Blau and Tomer Michaeli. Rethinking lossy compression: The rate-distortion-perception tradeoff. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 675–685. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/blau19a.html.

Pablo Rodrigo Gantier Cadena, Yeqiang Qian, Chunxiang Wang, and Ming Yang. Spade-e2vid: Spatially-adaptive denormalization for event-based video reconstruction. *IEEE Transactions on Image Processing*, 30:2488–2500, 2021. doi: 10.1109/TIP.2021.3052070.

Y. Collet and M. Kucherawy. Zstandard Compression and the 'application/zstd' Media Type. RFC 8878, February 2021. URL https://www.rfc-editor.org/info/rfc8878/.

Burak Ercan, Onur Eker, Aykut Erdem, and Erkut Erdem. Evreal: Towards a comprehensive benchmark and analysis suite for event-based video reconstruction. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 3943–3952, 2023. doi: 10.1109/CVPRW59228.2023.00410.

Cheng-De Fan, Chun-Wei Tuan Mu, Chen-Wei Chang, Chin-Yang Lin, Kun-Ru Wu, Yu-Chee Tseng, and Yu-Lun Liu. Longe2v: Long-horizon event-based video reconstruction, prediction, and frame interpolation with video diffusion models, 2026. URL https://arxiv.org/abs/2607.08770.

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 Transactions on Pattern Analysis and Machine Intelligence*, 44(1):154–180, 2022. doi: 10.1109/TPAMI.2020.3008413.

Mathias Gehrig, Willem Aarents, Daniel Gehrig, and Davide Scaramuzza. Dsec: A stereo event camera dataset for driving scenarios. *IEEE Robotics and Automation Letters*, 6(3):4947–4954, 2021a. doi: 10.1109/LRA.2021.3068942.

Mathias Gehrig, Mario Millhäusler, Daniel Gehrig, and Davide Scaramuzza. E-raft: Dense optical flow from event cameras. In *2021 International Conference on 3D Vision (3DV)*, pp. 197–206, 2021b. doi: 10.1109/3DV53792.2021.00030.

Anindya Ghosh, Thomas Nowotny, and James Knight. Evdownsampling: A robust method for downsampling event camera data. In Alessio Del Bue, Cristian Canton, Jordi Pont-Tuset, and Tatiana Tommasi (eds.), *Computer Vision – ECCV 2024 Workshops*, pp. 377–390, Cham, 2025. Springer Nature Switzerland. ISBN 978-3-031-92460-6.

Dailan He, Yaoyan Zheng, Baocheng Sun, Yan Wang, and Hongwei Qin. Checkerboard context model for efficient learned image compression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14771–14780, June 2021.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Jiahao Li, Bin Li, and Yan Lu. Deep contextual video compression. In *Advances in Neural Information Processing Systems*, volume 34, pp. 18114–18125, 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/96b250a90d3cf0868c83f8c965142d2a-Paper.pdf.

Weilun Li, Lei Sun, Ruixi Gao, Qi Jiang, Yuqin Ma, Kaiwei Wang, Ming-Hsuan Yang, Luc Van Gool, and Danda Pani Paudel. Evdiff: Event-based video reconstruction using one-step diffusion models, 2026. URL https://arxiv.org/abs/2511.17492.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling, 2023. URL https://arxiv.org/abs/2210.02747.

Jiawei Liu, Qiang Wang, Huijie Fan, Yinong Wang, Yandong Tang, and Liangqiong Qu. Residual denoising diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2773–2783, June 2024.

Fabian Mentzer, George D Toderici, Michael Tschannen, and Eirikur Agustsson. High-fidelity generative image compression. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 11913–11924. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/8a50bae297807da9e97722a0b3fd8f27-Paper.pdf.

Fabian Mentzer, David Minnen, Eirikur Agustsson, and Michael Tschannen. Finite scalar quantization: Vq-vae made simple. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 51772–51783, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/e2dd53601de57c773343a7cdf09fae1c-Paper-Conference.pdf.

David Minnen, Johannes Ballé, and George Toderici. Joint autoregressive and hierarchical priors for learned image compression. In *Advances in Neural Information Processing Systems*, volume 31, pp. 10794–10803, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/53edebc543333dfbf7c5933af792c9c4-Abstract.html.

Mohammad Mohammadi, Ziyi Wu, and Igor Gilitschenski. Tespec: Temporally-enhanced self-supervised pretraining for event cameras. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7782–7793, October 2025.

Linfeng Qi, Zhaoyang Jia, Jiahao Li, Bin Li, Houqiang Li, and Yan Lu. Generative latent coding for ultra-low bitrate image and video compression. *IEEE Transactions on Circuits and Systems for Video Technology*, 35(10):10500–10515, 2025. doi: 10.1109/TCSVT.2025.3571944.

Henri Rebecq, René Ranftl, Vladlen Koltun, and Davide Scaramuzza. Events-to-video: Bringing modern computer vision to event cameras. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3852–3861, 2019. doi: 10.1109/CVPR.2019.00398.

Zahra Rezaee, Catarina Brites, and João Ascenso. Lossy event compression: From event stream distortion to task performance, 2026. URL https://arxiv.org/abs/2608.28429.

Timo Stoffregen, Cedric Scheerlinck, Davide Scaramuzza, Tom Drummond, Nick Barnes, Lindsay Kleeman, and Robert Mahony. Reducing the sim-to-real gap for event cameras. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm (eds.), *Computer Vision – ECCV 2020*, pp. 534–549, Cham, 2020. Springer International Publishing. ISBN 978-3-030-58583-9.

Stepan Tulyakov, Alfredo Bochicchio, Daniel Gehrig, Stamatios Georgoulis, Yuanyou Li, and Davide Scaramuzza. Time lens++: Event-based frame interpolation with parametric nonlinear flow and multi-scale fusion. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17734–17743, 2022. doi: 10.1109/CVPR52688.2022.01723.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, Mengyang Feng, Ningyi Zhang, Pandeng Li, Pingyu Wu, Ruihang Chu, Ruili Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, Tianyi Gui, Tingyu Weng, Tong Shen, Wei Lin, Wei Wang, Wei Wang, Wenmeng Zhou, Wente Wang, Wenting Shen, Wenyuan Yu, Xianzhong Shi, Xiaoming Huang, Xin Xu, Yan Kou, Yangyu Lv, Yifei Li, Yijing Liu, Yiming Wang, Yingya Zhang, Yitong Huang, Yong Li, You Wu, Yu Liu, Yulin Pan, Yun Zheng, Yuntao Hong, Yupeng Shi, Yutong Feng, Zeyinzi Jiang, Zhen Han, Zhi-Fan Wu, and Ziyu Liu. Wan: Open and advanced large-scale video generative models, 2025. URL https://arxiv.org/abs/2503.20314.

Z. Wang, E.P. Simoncelli, and A.C. Bovik. Multiscale structural similarity for image quality assessment. In *The Thirty-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, volume 2, pp. 1398–1402 Vol.2, 2003. doi: 10.1109/ACSSC.2003.1292216.

Wenming Weng, Yueyi Zhang, and Zhiwei Xiong. Event-based video reconstruction using transformer. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2543–2552, 2021. doi: 10.1109/ICCV48922.2021.00256.

Ruihan Yang and Stephan Mandt. Lossy image compression with conditional diffusion models. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://papers.nips.cc/paper_files/paper/2023/hash/ccf6d8b4a1fe9d9c8192f00c713872ea-Abstract-Conference.html.

Shuhan Ye, Hongbin Yu, Chenqi Kong, Pingchuan Ma, Chong Wang, Jun Wan, and Qixin Zhang. Encore: Event-assisted complementary motion refinement for learned video compression, 2026. URL https://arxiv.org/abs/2607.28020.

Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2633–2642, 2020. doi: 10.1109/CVPR42600.2020.00271.

Zongsheng Yue, Jianyi Wang, and Chen Change Loy. Resshift: Efficient diffusion model for image super-resolution by residual shifting. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 13294–13307. Curran Associates, Inc., 2023. doi: 10.52202/075280-0583. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/2ac2eac5098dba08208807b65c5851cc-Paper-Conference.pdf.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.