EVENT-GUIDED NEURAL VIDEO COMPRESSION

Anonymous authors
Paper under double-blind review

ABSTRACT

Neural video codecs derive motion and temporal contexts mainly from RGB frames, leaving room for cross-modal guidance from complementary temporal observations. Event streams can provide such observations by recording brightness changes between frames. In this work, we propose an **Event-guided Neural Video Codec (ENVC)** that uses events shared by the encoder and decoder to improve RGB compression efficiency. For motion coding, ENVC forms an event-guided motion prior and codes the remaining motion residual. For frame coding, an event-conditioned predictor supplies multi-scale features for gated temporal context refinement. To support training and evaluation on standard video datasets, we synthesize paired RGB–event data and assess its predictive utility through comparisons with real events. Across six benchmarks, ENVC achieves average BD-rate savings of **39.13%** using PSNR-RGB and **67.63%** using LPIPS relative to DCMVC. Further analyses show that our gains persist on large-motion sequences and that ENVC effectively learns to integrate event information. These results demonstrate the potential of events as a complementary modality for reducing the RGB coding rate. Our model and code will be released upon acceptance.

1 INTRODUCTION

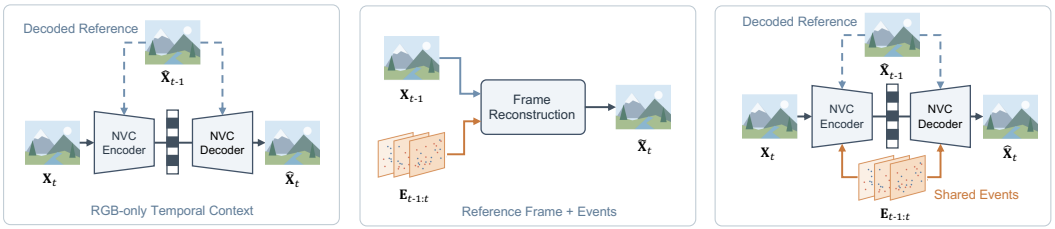
Video accounts for a large share of network traffic, making efficient video compression a central problem in visual communication. Neural video codecs (NVCs) optimize their coding modules jointly under an end-to-end rate–distortion (RD) objective (Lu et al., 2019). Conditional coding methods, which code the current frame using temporal contexts built from previously decoded information, have demonstrated RD gains over conventional codecs (Li et al., 2021; 2023). The quality of these contexts therefore strongly affects the bitrate needed for a given reconstruction quality. Accordingly, recent studies have refined context modeling via multi-scale propagation (Sheng et al., 2022), reference-guided modulation (Tang et al., 2025), and non-local multi-reference mining (Jiang et al., 2025).

Despite these advances, temporal contexts in the RGB-only framework (Fig. 1a) depend only on previously decoded frames and propagated features. Diverse motion patterns can make prediction from these references less reliable, while reconstruction errors can accumulate over long prediction chains. In both cases, more RGB information may be needed to maintain reconstruction quality. We therefore ask whether complementary observations of inter-frame changes can improve temporal contexts and reduce the required RGB bitrate.

To investigate this possibility, we focus on event streams, which asynchronously record brightness changes between video frames. An event is triggered whenever the magnitude of the log-intensity change at a pixel exceeds a contrast threshold, with its polarity indicating the sign of the change. This mechanism yields microsecond-level temporal resolution and a high dynamic range (Gallego et al., 2022). These properties have motivated a range of event-based methods for motion analysis (Gehrig et al., 2021; Wang et al., 2024b) and video reconstruction and interpolation (Rebecq et al., 2019; Weng et al., 2021; Sun et al., 2023).

A reference frame and inter-frame events can be used to reconstruct the current frame directly, as illustrated in Fig. 1b. However, without additional current-frame RGB information, prediction errors can limit reconstruction fidelity (Zhu et al., 2024a; Kong et al., 2025). We instead consider RGB coding with events shared by the encoder and decoder (Fig. 1c). Both sides use these events to construct temporal contexts, while a bitstream carries additional current-frame RGB information to

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.



(a) RGB-only Video Coding (b) Event-based Reconstruction (c) Event-guided Video Coding

Figure 1: **Comparison of Video Coding and Reconstruction Settings.** (a) RGB-only video coding reconstructs the current frame from a decoded reference and an RGB coding bitstream. (b) Event-based reconstruction uses a reference frame and inter-frame events without an additional current-frame RGB bitstream. (c) Event-guided video coding uses events shared by the encoder and decoder to construct temporal contexts and code the current frame under a rate–distortion objective.

improve reconstruction quality. Similar shared-side-information settings have been explored in prior work on cross-modal compression (Gnutti et al., 2024; Jiang et al., 2024; Song et al., 2026), and we discuss practical scenarios in App. A.

In this work, we investigate how shared event observations can reduce the RGB rate required for reconstruction. We propose an **Event-guided Neural Video Codec (ENVC)** that incorporates these observations into motion coding and temporal context construction. ENVC forms an event-guided motion prior at both the encoder and decoder and codes the remaining motion residual. For frame coding, an event-conditioned predictor generates a frame prediction and multi-scale features. Learned gates inject these features into temporal contexts, which are further refined using the frame prediction.

To train and evaluate ENVC on standard NVC datasets, we generate paired events from their RGB videos through frame interpolation and event simulation (Sec. 3). We assess the utility of the synthetic events by comparing event statistics and next-frame prediction quality with those obtained using real events. Across six standard video coding benchmarks paired with synthetic events, ENVC achieves average BD-rate savings of 39.13% using PSNR-RGB and 67.63% using LPIPS (Zhang et al., 2018) relative to DCMVC (Tang et al., 2025) under the shared-event setting. Our evaluation further analyzes event utilization, motion-dependent performance, sensitivity to event perturbations, and RD differences between real and synthetic event inputs.

Our contributions can be summarized as follows:

- We investigate shared events as a complementary modality for RGB video compression and propose ENVC, which combines event-guided residual motion coding with gated temporal context refinement.
- We construct paired RGB–event datasets for standard NVC training and evaluation, and assess the synthetic events through comparisons with real events in terms of event statistics and next-frame prediction quality.
- We demonstrate substantial RGB coding gains across six standard benchmarks using PSNR-RGB and LPIPS under the shared-event setting. These gains persist on large-motion sequences, supporting the use of events as temporal guidance for RGB video coding.

2 Related Work

Neural Video Compression. Beyond conventional hybrid codecs (Wiegand et al., 2003; Sullivan et al., 2012; Bross et al., 2021), NVCs learn motion and residual coding (Lu et al., 2019; Agustsson et al., 2020; Hu et al., 2021; Rippel et al., 2021) or code frames conditioned on temporal contexts (Ladune et al., 2021; Ho et al., 2022; Li et al., 2021). Context modeling has advanced through multi-scale mining (Sheng et al., 2022), spatio-temporal entropy modeling (Li et al., 2022; 2023), reference modulation (Tang et al., 2025), and multi-reference aggregation (Jiang et al., 2025). Other works address quality control and temporal stability (Li et al., 2024), speed (Jia et al., 2025; Li et al., 2026), buffer requirements (Chen et al., 2025), and model capacity (Tang et al., 2026). ENVC uses shared event observations to complement the RGB-derived temporal information used by these methods.

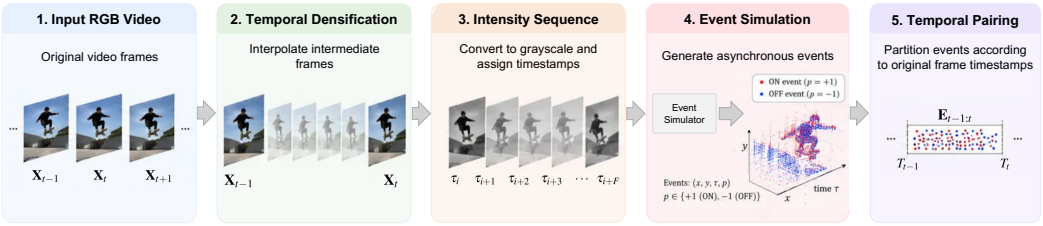


Figure 2: **Synthetic RGB–Event Dataset Construction.** Temporally densified frames are used for event synthesis, while the original RGB frames are retained for coding. The resulting event stream is partitioned at the original frame timestamps into groups $\mathbf{E}_{t-1:t}$ aligned with each coding interval.

Event-based Frame Reconstruction. Event-only methods reconstruct intensity frames from event streams (Rebecq et al., 2019; Weng et al., 2021; Zhu et al., 2024b), although events do not directly measure absolute intensity (Gallego et al., 2022). Frame-guided methods combine RGB images and events for interpolation (Tulyakov et al., 2021; 2022) or prediction from a single reference (Zhu et al., 2024a; Kong et al., 2025; Wang et al., 2025). Single-reference prediction is compatible with sequential decoding, as a previously reconstructed frame can serve as the reference without requiring future RGB frames. ENVC uses event-based prediction to guide compression and codes additional current-frame RGB information under an RD objective.

Cross-modal Guidance for Compression. Joint visible–infrared and RGB–depth codecs exploit inter-modality redundancy (Lu et al., 2022; Zheng & Gao, 2024). Auxiliary modalities also guide RGB compression as geometric (Gnutti et al., 2024; Jiang et al., 2024) or semantic side information (Jiang et al., 2023; Lee et al., 2024; Hu et al., 2026), including semantic residual coding (Ke et al., 2025) and text-conditioned latent refinement for machine vision (Lee et al., 2025). Depth-assisted video compression uses geometric priors for temporal context and entropy modeling (Sheng et al., 2026). Our work first investigates shared events as temporal side information, using their observations of inter-frame brightness changes to guide motion coding and temporal context construction.

3 Synthetic RGB–Event Dataset Construction

Training and evaluating RGB compression with shared events requires paired RGB–event data. Real-event datasets provide such pairs (Tulyakov et al., 2021; 2022; Kim et al., 2023), while paired events are unavailable for Vimeo-90K (Xue et al., 2019), a widely used dataset for NVC training, and standard video coding benchmarks (Flynn et al., 2013; Wang et al., 2016; Mercat et al., 2020; Cheon & Lee, 2018). We therefore synthesize events from these RGB datasets using the pipeline in Fig. 2, enabling RD comparisons with prior work on the same video content.

3.1 Event Synthesis Pipeline

Temporal Densification. We use Super-SloMo (Jiang et al., 2018) to insert $F-1$ intermediate frames between each adjacent pair in the original sequence $\{\mathbf{X}_t\}_{t=1}^{N}$. Given the source frame rate f_{src}, we set $F = \lceil f_{\text{target}}/f_{\text{src}} \rceil$ with $f_{\text{target}} = 200$ fps, yielding a nominal frame rate of at least f_{target}. This gives a densified sequence $\{\mathbf{Y}_i\}_{i=1}^{F(N-1)+1}$, which we convert to 8-bit grayscale and assign timestamps $\tau_i = i \cdot \lfloor 10^6/(F f_{\text{src}}) \rceil$ in microseconds, yielding an intensity sequence $I(\tau)$ at each pixel. The densified grayscale frames are used only for event synthesis.

Event Simulation. We generate events from $I(\tau)$ with DVS-Voltmeter (Lin et al., 2022), a stochastic simulator that models per-pixel voltage changes and event triggering. We use the DVS346[1] preset of the simulator without additional calibration.

Temporal Pairing. Each event is stored as a tuple (x, y, τ, p) with pixel location (x, y), timestamp τ, and polarity $p \in \{+1, -1\}$. Let $T_t = \tau_{1+F(t-1)}$ denote the timestamp assigned to the original frame $\mathbf{X}_t$. We partition the event stream into

$$\mathbf{E}_{t-1:t} = \{(x, y, \tau, p) \mid T_{t-1} < \tau \leq T_t\}, \qquad t = 2, \ldots, N. \tag{1}$$

Each coding interval is thus associated with $(\mathbf{X}_{t-1}, \mathbf{E}_{t-1:t}, \mathbf{X}_t)$.

[1] https://github.com/Lynn0306/DVS-Voltmeter.

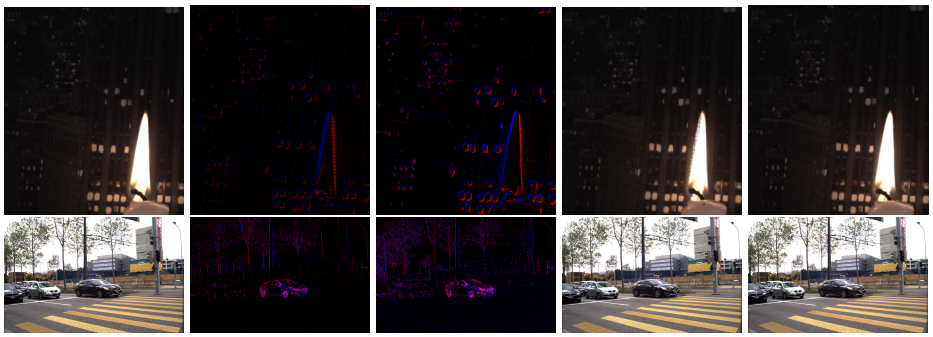


Figure 3: **Qualitative Comparison.** The top and bottom rows show HS-ERGB and BS-ERGB examples, respectively. VFPSIE predicts the target frame $\mathbf{X}_t$ from the same reference frame $\mathbf{X}_{t-1}$ using either real or synthetic events from the corresponding interval.

3.2 Synthetic Event Quality Assessment

We assess whether synthetic events provide useful temporal information for frame prediction. Using the pipeline in Sec. 3.1, we synthesize events from the RGB frames of 15 HS-ERGB test sequences (Tulyakov et al., 2021) and 17 BS-ERGB validation sequences (Tulyakov et al., 2022), and compare them with real events over the same intervals. For each sequence, we evaluate the 31 adjacent intervals within its first 32 frames.

We compare event counts and their spatial distributions (see Tab. 1). For each interval, spatial count maps record event counts per 4×4 cell for HS-ERGB and per pixel for BS-ERGB. We report the pooled synthetic-to-real event count ratio and median per-interval correlation between the corresponding spatial count maps.

Table 1: **Real and Synthetic Events.** PSNR-RGB is measured for the reference frame (Ref.) and VFPSIE predictions using real or synthetic events (Real/Syn.).

Dataset	Event statistics		PSNR-RGB (dB)		
	Count Ratio	Spatial Corr.	Ref.	Real	Syn.
HS-ERGB	1.543	0.507	29.27	31.78	34.13
BS-ERGB	1.587	0.606	22.05	23.91	24.91

Using $\mathbf{X}_{t-1}$ directly as a prediction of $\mathbf{X}_t$ gives a reference-only baseline (Ref.). We compare this baseline with predictions from the same pretrained VFPSIE (Zhu et al., 2024a), conditioned on $\mathbf{X}_{t-1}$ and either real or synthetic events, reporting PSNR-RGB against $\mathbf{X}_t$ averaged across intervals. The PSNR-RGB gain over Ref. reflects the predictive benefit of inter-frame events beyond the information available in the previous RGB frame.

Synthetic events are 1.54 and 1.59 times as numerous as real events on HS-ERGB and BS-ERGB, respectively, while exhibiting spatial agreement with real events. They enable VFPSIE to achieve prediction quality comparable to that obtained with real events (Tab. 1). The examples in Fig. 3 further show that synthetic events closely resemble real events in their dominant spatial structures, capturing similar object contours and regions of concentrated event activity. Together, these results support the utility of synthetic events as temporal guidance for frame prediction. Additional qualitative examples are provided in App. I.

4 Proposed Method

In this section, we introduce **Event-guided Neural Video Codec (ENVC)**, and describe how ENVC leverages shared events to predict motion and refine temporal contexts for conditional frame coding.

4.1 Framework Overview

Fig. 4a illustrates the overview pipeline of ENVC, inspired by the conditional coding framework of DCVC-DC (Li et al., 2023) and DCMVC (Tang et al., 2025). Both encoder and decoder use the reconstructed reference $\hat{\mathbf{X}}_{t-1}$ and shared events $\mathbf{E}_{t-1:t}$. The **E**vent-conditioned **F**rame **P**redictor (**EFP**) produces a frame prediction $\mathbf{P}_t$ and multi-scale event-conditioned contexts. We estimate

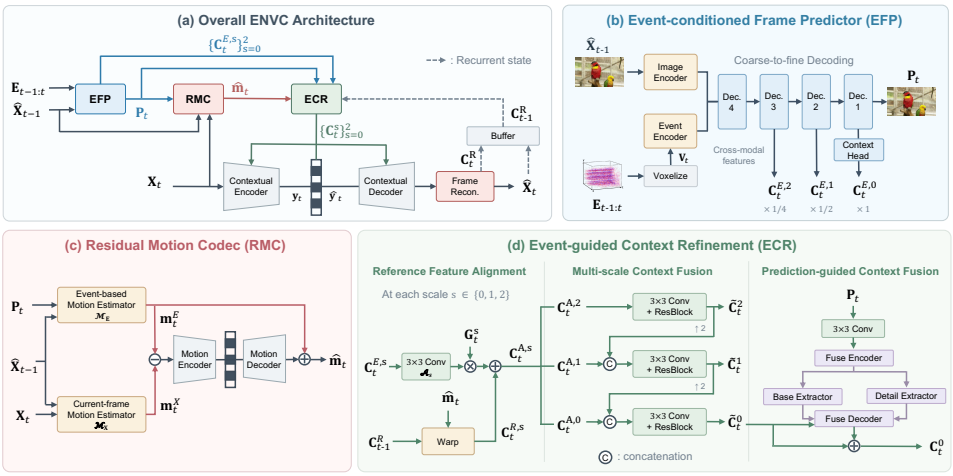


Figure 4: **Overview Pipeline of ENVC.** The overall framework (a) combines event-conditioned frame prediction (b), residual motion coding (c), and event-guided context refinement (d).

motion from $\mathbf{P}_t$ relative to $\hat{\mathbf{X}}_{t-1}$, while the **R**esidual **M**otion **C**odec (**RMC**) codes the difference between the current-frame and event-based motion estimates. Adding the decoded residual to the event-based estimate yields the reconstructed motion (Sec. 4.2).

Event-guided **C**ontext **R**efinement (**ECR**) aligns reference features using the reconstructed motion, then refines them through gated event injection and multi-scale fusion. **P**rediction-guided **C**ontext **F**usion (**PCF**) further incorporates features extracted from $\mathbf{P}_t$ into the $\times 1$ context (Sec. 4.3). The resulting temporal contexts condition frame encoding, entropy modeling, and reconstruction (Sec. 4.4).

4.2 Event-Predictive Motion Coding

Event-conditioned Frame Prediction (EFP). EFP is built on the VFPSIE backbone (Zhu et al., 2024a), with outputs for both frame prediction and multi-scale context construction. We represent $\mathbf{E}_{t-1:t}$ as a signed event voxel grid $\mathbf{V}_t \in \mathbb{R}^{10\times H\times W}$, distributing each event between neighboring temporal bins with linear weights. Given $\hat{\mathbf{X}}_{t-1}$ and $\mathbf{V}_t$, EFP produces

$$\left(\mathbf{P}_t, \{\mathbf{C}_t^{E,s}\}_{s=0}^{2}\right) = \mathcal{P}_\theta\left(\hat{\mathbf{X}}_{t-1}, \mathbf{V}_t\right), \tag{2}$$

where $\mathcal{P}_\theta$ denotes EFP and $\mathbf{C}_t^{E,s}$ denotes its event-conditioned context at scale s.

As shown in Fig. 4b, separate encoders extract image and event features. A coarse-to-fine decoder combines these features through cross-modal attention to estimate flow and synthesize an RGB frame. The prediction $\mathbf{P}_t$ is obtained by blending the synthesized frame with the forward-warped reference.

We extract $\mathbf{C}_t^{E,1}$ and $\mathbf{C}_t^{E,2}$ from the decoder outputs at $\times 1/2$ and $\times 1/4$ resolution, respectively. A separate context head generates $\mathbf{C}_t^{E,0}$ at $\times 1$ resolution from intermediate features of the final decoding stage.

Residual Motion Coding (RMC). For reference alignment, we estimate backward motion rather than reuse the internal forward flow from EFP. The **E**vent-based **M**otion **E**stimator (**EME**) $\mathcal{M}_E$ and **C**urrent-frame **M**otion **E**stimator (**CME**) $\mathcal{M}_X$ use separate SPyNet-based networks (Ranjan & Black, 2017):

$$\mathbf{m}_t^E = \mathcal{M}_E\left(\mathbf{P}_t, \hat{\mathbf{X}}_{t-1}\right), \qquad \mathbf{m}_t^X = \mathcal{M}_X\left(\mathbf{X}_t, \hat{\mathbf{X}}_{t-1}\right). \tag{3}$$

The event-based estimate is computed at both the encoder and decoder, whereas the current-frame estimate is computed only at the encoder.

RMC codes their difference:

$$\Delta\mathbf{m}_t = \mathbf{m}_t^X - \mathbf{m}_t^E, \qquad \hat{\mathbf{m}}_t = \mathbf{m}_t^E + \Delta\hat{\mathbf{m}}_t, \tag{4}$$

where $\Delta\hat{\mathbf{m}}_t$ is reconstructed from the quantized motion latent (Fig. 4c). The motion entropy model uses a hyperprior and the previously reconstructed motion latent when available (Li et al., 2023).

Table 2: **BD-Rate (%) against DCMVC.** Each comparison uses PSNR-RGB (left) and LPIPS (right). "–" indicates no overlapping quality range. Averages are computed over datasets with valid BD-rate values.

Method	UVG	MCL-JCV	HEVC-B	HEVC-C	HEVC-D	HEVC-E	Average
DCMVC	0.00 / 0.00	0.00 / 0.00	0.00 / 0.00	0.00 / 0.00	0.00 / 0.00	0.00 / 0.00	0.00 / 0.00
DCVC	+176.33 / +188.87	+133.70 / +139.12	+141.41 / +176.24	+222.23 / +168.15	+213.26 / +166.59	– / +228.29	+177.39 / +177.88
DCVC-TCM	+59.17 / +29.03	+59.66 / +38.33	+49.37 / +38.54	+91.85 / +52.10	+84.92 / +55.88	+101.45 / +49.84	+74.40 / +43.95
DCVC-DC	+0.01 / −11.50	+1.05 / −9.55	−1.85 / −8.89	+3.03 / −5.70	−8.83 / −4.50	−3.64 / −14.50	−1.71 / −9.10
DCVC-FM	+6.16 / −11.46	+10.75 / −4.57	+8.14 / −8.73	+11.67 / −3.76	−10.62 / +2.28	−1.70 / −14.34	+4.07 / −6.76
DCVC-RT	+20.74 / −7.72	+28.64 / +4.58	+31.32 / +2.92	+66.42 / +13.64	+44.64 / +15.91	+20.16 / −5.49	+35.32 / +3.97
ENVC (Ours)	**−29.73 / −82.21**	**−52.28 / −84.64**	**−51.96 / −77.12**	**−49.74 / −60.65**	**−43.30 / −49.54**	**−7.76 / −51.60**	**−39.13 / −67.63**

4.3 Event-guided Context Refinement

Fig. 4d shows how motion-aligned reference information is refined using EFP contexts and predictions.

Reference Feature Alignment. From the propagated decoder feature $\mathbf{C}_{t-1}^{R}$, we construct reference features at $\times 1$, $\times 1/2$, and $\times 1/4$ resolution. When this feature is unavailable, we instead extract reference features from $\hat{\mathbf{X}}_{t-1}$. For numerical stability, we also regenerate the reference features from $\hat{\mathbf{X}}_{t-1}$ when the maximum absolute value of the initial $\times 1$ context exceeds 5. We align the reference features by backward warping with $\hat{\mathbf{m}}_t$, adjusting the motion field and its displacement to each scale. At $\times 1$, offset diversity (Li et al., 2023) further refines the alignment with learned sampling offsets and masks. We denote the aligned reference contexts by $\{\mathbf{C}_t^{R,s}\}_{s=0}^{2}$.

Gated Multi-scale Event Injection. We inject EFP contexts into the aligned reference contexts through gated residual connections:

$$\mathbf{C}_t^{A,s} = \mathbf{C}_t^{R,s} + \mathbf{G}_t^{s} \odot \mathcal{A}_s\left(\mathbf{C}_t^{E,s}\right), \tag{5}$$

where $\mathcal{A}_s$ is a zero-initialized feature adapter and $\odot$ denotes element-wise multiplication. Each gate $\mathbf{G}_t^s$ is a single-channel sigmoid map generated from $\mathbf{V}_t$ and $\hat{\mathbf{X}}_{t-1}$, broadcast across feature channels. Multi-scale context fusion combines the augmented contexts along a top-down path from $\times 1/4$ through $\times 1/2$ to $\times 1$, producing $\{\widetilde{\mathbf{C}}_t^{s}\}_{s=0}^{2}$.

Prediction-guided Context Fusion (PCF). PCF further refines the $\times 1$ context ($\widetilde{\mathbf{C}}_t^{0}$) using the frame prediction $\mathbf{P}_t$. Features extracted from $\mathbf{P}_t$ are fused with $\widetilde{\mathbf{C}}_t^{0}$ through the base/detail fusion module of DCMVC (Tang et al., 2025), and the fused representation is added back to $\widetilde{\mathbf{C}}_t^{0}$. The three contexts are finally clipped to $[-10, 10]$, yielding $\{\mathbf{C}_t^{s}\}_{s=0}^{2}$ for conditional frame coding.

4.4 Frame Coding and Optimization

We use the conditional frame codec of DCVC-DC (Li et al., 2023), conditioned on $\{\mathbf{C}_t^{s}\}_{s=0}^{2}$. The contextual encoder transforms $\mathbf{X}_t$ into a latent using contexts at the corresponding scales, and the contextual decoder reconstructs the frame from the quantized latent and the same contexts. The $\times 1/4$ context also provides a temporal prior for entropy modeling.

The reconstructed frame and decoder feature form the reference state for the next frame. Excluding shared events, the estimated RGB coding rate for each P-frame is

$$R_t = R(\hat{\mathbf{y}}_t^m) + R(\hat{\mathbf{z}}_t^m) + R(\hat{\mathbf{y}}_t) + R(\hat{\mathbf{z}}_t), \tag{6}$$

where $R(\cdot)$ is the entropy-model rate estimate, and $(\hat{\mathbf{y}}_t^m, \hat{\mathbf{z}}_t^m)$ and $(\hat{\mathbf{y}}_t, \hat{\mathbf{z}}_t)$ are the motion and frame latents with their hyperlatents, respectively. The main rate–distortion term is

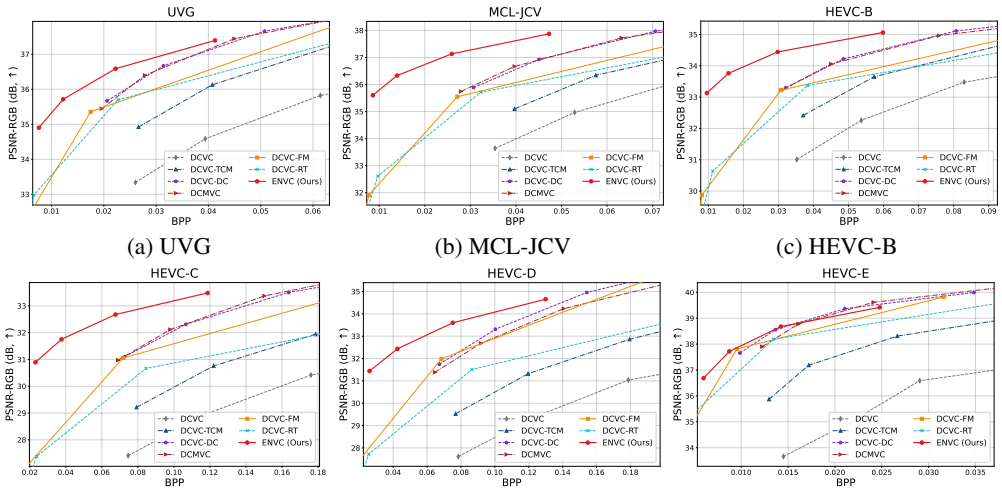
$$\mathcal{L}_{\mathrm{RD},t} = R_t + \lambda D(\mathbf{X}_t, \hat{\mathbf{X}}_t), \tag{7}$$

where D measures reconstruction distortion and λ controls the rate–distortion trade-off. Stage-wise objectives, detailed training stages and optimization settings are provided in App. B.

5 Experiments

5.1 Experimental Details

Training Details. We train ENVC on Vimeo-90K (Xue et al., 2019) septuplet sequences paired with synthetic events. We adopt staged training followed by cascaded fine-tuning with up to 32



(a) UVG (b) MCL-JCV (c) HEVC-B

(d) HEVC-C (e) HEVC-D (f) HEVC-E

Figure 5: **Rate–Distortion Performance.** PSNR-RGB (↑) versus bitrate on six benchmarks.

frames (Li et al., 2024), using longer clips extracted from the original Vimeo videos[2]. The stage-wise objectives and training schedule are detailed in App. B.

Evaluation Details. We evaluate on UVG (Mercat et al., 2020), MCL-JCV (Wang et al., 2016), and HEVC Classes B, C, D, and E (Flynn et al., 2013), using RGB frames paired with synthetic events. Unless otherwise specified, all models are evaluated on the first 96 frames of each sequence with an intra period of 32. ENVC uses the pretrained intra-frame codec of DCVC-DC (Li et al., 2023) for I-frame coding.

We report PSNR-RGB and LPIPS (Zhang et al., 2018), summarizing RD performance using BD-rate (Bjontegaard, 2001) relative to DCMVC (Tang et al., 2025). For all models, coding rates are estimated from entropy models and reported in bits per pixel (bpp), including both I- and P-frame coding. Results using additional quality metrics and datasets are provided in App. H.

We compare ENVC with DCVC (Li et al., 2021), DCVC-TCM (Sheng et al., 2022), DCVC-DC (Li et al., 2023), DCVC-FM (Li et al., 2024), DCMVC (Tang et al., 2025), and DCVC-RT (Jia et al., 2025).

5.2 Rate–Distortion Performance

Figs. 5 and 6, together with Tab. 2, summarize the RD performance. Across six benchmarks, ENVC achieves average BD-rate savings of 39.13% using PSNR-RGB and 67.63% using LPIPS relative to DCMVC. These results show that shared events can reduce the RGB rate required for reconstruction under both pixel-wise and perceptual quality measures. An analysis of P-frame bit allocation in App. G further shows reduced motion coding rates relative to DCMVC.

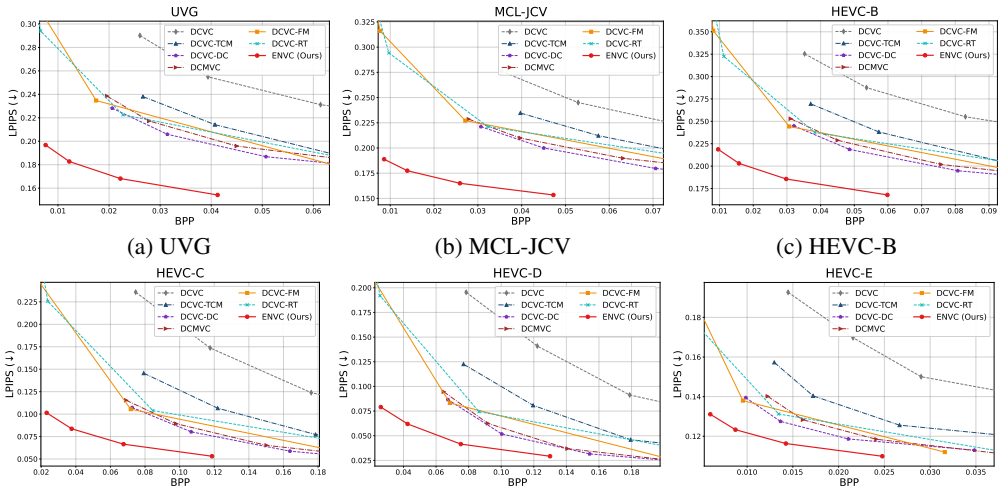
On HEVC-E, PSNR-RGB savings are smaller (7.76%), while LPIPS-based savings remain substantial (51.60%). The smaller PSNR-RGB gains are consistent with limited inter-frame motion and the high I-frame share of the total RGB rate (56.2% at Q_0), which limits the impact of event-guided P-frame improvements. Sequence-level gains and failure cases are further analyzed in App. C.

Since ENVC is trained on synthetic events, we further examine how its RD performance changes when real events are used instead. Comparisons using real and synthetic event inputs on paired RGB–event datasets are provided in App. D.

5.3 How Does ENVC Use Events?

To understand how ENVC incorporates events into its temporal context, we examine its multi-scale event gates ($\mathbf{G}_t^s$). We use quality factor Q_1 and 12 sequences.

[2] https://github.com/anchen1011/toflow/blob/master/data/original_vimeo_links.txt.

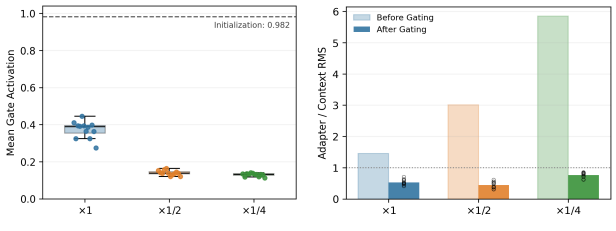


(a) UVG (b) MCL-JCV (c) HEVC-B

(d) HEVC-C (e) HEVC-D (f) HEVC-E

Figure 6: **Rate–Distortion Performance.** LPIPS (↓) versus bitrate on six benchmarks.

The mean gate activations at $\times 1$, $\times 1/2$, and $\times 1/4$ are 0.373, 0.140, and 0.130, respectively (Fig. 7a). All three means are substantially below the initialization of 0.982, showing that ENVC learns to attenuate the adapter outputs. Fig. 7b compares adapter-feature root-mean-square (RMS) magnitudes before and after gating, normalized by the RMS of the aligned reference contexts before injection. After gating, these ratios are 0.533, 0.444, and 0.765, respectively.



(a) Learned Gate Activation (b) Relative Feature Magnitude

Figure 7: **Event Gating at Q_1.** (a) Mean gate activations. (b) Adapter-to-reference-context RMS ratios before and after gating.

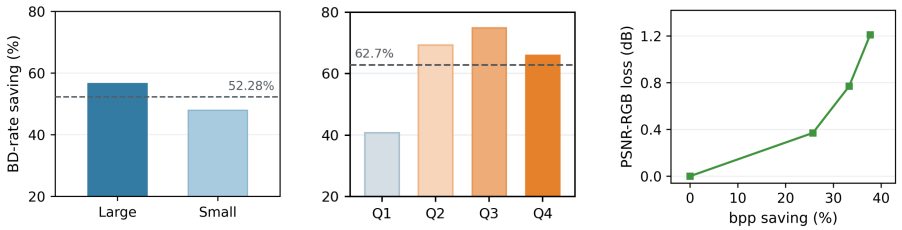
Although the ungated adapter features exceed the reference contexts in RMS magnitude at all three scales, gating brings their magnitudes below those of the reference contexts while retaining substantial event-conditioned signals. The $\times 1/4$ branch has the largest adapter-to-context RMS ratio both before and after gating, despite its smallest mean gate activation. Mean gate activation alone therefore does not characterize the magnitude of event injection, which also depends on the adapter features being modulated. We also replace all injection gates with constants $G = 0$ or $G = 1$ at inference, without retraining. Across both interventions, PSNR decreases by 2.30–3.89 dB on four tested sequences, indicating that the trained model relies on the learned modulation for reconstruction. Additional analyses and detailed intervention results are provided in Apps. E and F.

5.4 Effects of Motion Magnitude and Intra Period

Motion Magnitude. We examine how our coding gains vary with the motion magnitude. Following DCVC-RT (Jia et al., 2025), we measure motion using SEA-RAFT (Wang et al., 2024a) optical flow between consecutive original frames. On 30 MCL-JCV sequences (96 frames each, intra period 32), ENVC achieves an overall BD-rate saving of 52.28% over DCMVC.

At the sequence level, the large-motion group achieves a saving of 56.60%, compared with 47.96% for the small-motion group (Fig. 8a). We further examine this relationship at the frame level by grouping only P-frames into four motion quartiles. From the lowest to the highest motion quartile, BD-rate savings are 40.8%, 69.3%, 74.8%, and 65.9%, respectively (Fig. 8b). Although the largest saving occurs in the third quartile, the highest-motion quartile still shows substantial gains. Together, these results indicate that event guidance remains useful across different motion magnitudes, including sequences and frames with large motion.

Longer Intra Periods. On seven UVG sequences, we examine how longer intra periods affect RGB coding rate and reconstruction quality. Relative to $K = 32$, increasing K to 64, 96, and 192



(a) Sequence-level Motion (b) Frame-level Motion (c) Intra Period

Figure 8: **Motion-dependent BD-rate Savings and Longer Intra Periods.** (a–b) BD-rate savings over DCMVC. Dashed lines indicate the sequence mean and the mean over four quartiles, respectively. The frame-level analysis in (b) includes only P-frames. (c) Relative bpp saving and PSNR-RGB loss for single-GOP coding, using a 32-frame GOP as the reference. Points from left to right correspond to GOP lengths of 32, 64, 96, and 192 frames, each containing one I-frame.

reduces mean bpp by 25.7%, 33.3%, and 37.7%, with PSNR-RGB losses of 0.37, 0.77, and 1.21 dB, respectively (Fig. 8c). Increasing K from 32 to 64 provides the largest incremental rate saving. Further increases yield smaller additional savings while reconstruction quality continues to decline.

5.5 Complexity Analysis

Tab. 3 reports model size and encoding/decoding throughput measured on the 1920×1080 UVG Beauty sequence using a single RTX 3090. ENVC processes events in addition to video frames and has 33.52M parameters, with encoding and decoding speeds of 0.323 and 0.386 FPS, respectively. Both encoder and decoder run EFP and EME modules to construct the shared motion prior, adding computation to both coding stages. The additional event-processing modules increase model size and computational cost, making efficiency a limitation of the current implementation.

Table 3: **Complexity Analysis.** Model size and encoding/decoding speed on UVG Beauty using a single RTX 3090.

Model	Params.	Enc. FPS	Dec. FPS
DCVC	7.94M	0.153	0.023
DCVC-TCM	10.71M	1.883	2.749
DCVC-DC	19.78M	1.691	2.082
DCVC-FM	18.34M	2.719	3.298
DCMVC	21.02M	1.061	1.208
DCVC-RT	20.69M	57.221	51.198
ENVC	33.52M	0.323	0.386

6 Discussion

Our results suggest that events are a promising complementary modality to guide RGB video compression. This motivates further research on efficient event coding and joint compression of RGB frames and event streams. Jointly modeling the dependencies between these modalities could further exploit cross-modal redundancy and improve overall coding efficiency, extending the present framework toward joint RGB–event compression.

As an initial exploration of RGB coding with shared events, ENVC focuses on integrating event information into motion coding and temporal context construction. The current implementation requires event-conditioned prediction and motion estimation at both the encoder and decoder, introducing additional computational overhead. Developing lightweight event-guidance modules and efficient coding pipelines to reduce complexity and end-to-end latency remains an important direction for future work.

7 Conclusion

In this work, we presented ENVC, which incorporated events as a complementary modality for RGB video compression through event-guided residual motion coding and gated temporal context refinement. Across six benchmarks with synthetic events, ENVC achieved average BD-rate savings of 39.13% using PSNR-RGB and 67.63% using LPIPS relative to DCMVC. Further analysis showed scale-dependent event injection, while longer intra periods revealed a growing rate–quality trade-off. These results demonstrate the potential of shared event observations as a complementary modality for reducing the RGB coding rate.

AI USE STATEMENT

We use generative AI tools for minor English editing and assistance with baseline implementation and coding. We do not use them to develop the research idea, conduct comparisons, or analyze and interpret the results. All AI-generated code and text are reviewed and verified by the authors, who take full responsibility for the final content.

REPRODUCIBILITY STATEMENT

We address all the data preparation, model design, training procedure, and evaluation protocols in the main paper and appendices. Sec. 3 describes the synthetic RGB–event data construction pipeline, including temporal densification, the event simulator and its preset, and temporal pairing with RGB frames. Sec. 4 presents the model components, their interactions, and the rate–distortion objective. App. B details the training stages, module initialization and freezing, cascaded sequence lengths, training data, event representation, crop sizes, multi-rate loss settings, optimizer, and training hardware.

Sec. 5.1 specifies the evaluation datasets, compared codecs, frame counts, intra period, quality metrics, and entropy-estimated rate reporting. We also describe the protocols for comparing real and synthetic events in App. D, analyzing learned event gating in App. E, and applying spatiotemporal event interventions in App. F. App. H provides additional evaluations on 4K video and supplementary quality metrics. Our source code and pretrained model checkpoints will be publicly released upon acceptance.

REFERENCES

E. Agustsson, D. Minnen, N. Johnston, J. Ballé, S. J. Hwang, and G. Toderici. Scale-space flow for end-to-end optimized video compression. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8500–8509. IEEE, 2020.

S. H. Ahmed, J. Finkbeiner, and E. Neftci. Efficient event-based object detection: A hybrid neural network with spatial and temporal attention. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13970–13979. IEEE, 2025.

L. Bartolomei, E. Mannocci, F. Tosi, M. Poggi, and S. Mattoccia. Depth AnyEvent: A cross-modal distillation paradigm for event-based monocular depth estimation. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 19669–19678. IEEE, 2025.

G. Bjontegaard. Calculation of average psnr differences between rd-curves. *ITU SG16 Doc. VCEG-M33*, 2001.

C. Brites and J. Ascenso. Neuromorphic vision data coding: Classifying and reviewing the literature. *IEEE Access*, 13:14626–14657, 2025.

B. Bross, Y.-K. Wang, Y. Ye, S. Liu, J. Chen, G. J. Sullivan, and J.-R. Ohm. Overview of the versatile video coding (VVC) standard and its applications. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(10):3736–3764, 2021.

T. Chen, B. Xu, C. Zhang, and Carlos Guestrin. Training deep nets with sublinear memory cost. *arXiv preprint arXiv:1604.06174*, 2016.

Y.-H. Chen, Y.-C. Yao, K.-W. Ho, C.-H. Wu, H.-T. Phung, M. Benjak, J. Ostermann, and W.-H. Peng. HyTIP: Hybrid temporal information propagation for masked conditional residual video coding. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 17889–17898. IEEE, 2025.

M. Cheon and J.-S. Lee. Subjective and objective quality assessment of compressed 4K UHD videos for immersive experience. *IEEE Transactions on Circuits and Systems for Video Technology*, 28 (7):1467–1480, 2018.

D. Flynn, K. Sharman, and C. Rosewarne. Common test conditions and software reference configurations for HEVC range extensions. Document JCTVC-N1006, Joint Collaborative Team Video Coding ITU-T SG 16, 2013.

G. Gallego, T. Delbrück, G. Orchard, C. Bartolozzi, B. Taba, A. Censi, S. Leutenegger, A. J. Davison, J. Conradt, K. Daniilidis, and D. Scaramuzza. Event-based vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(1):154–180, 2022.

M. Gehrig, M. Millhäusler, D. Gehrig, and D. Scaramuzza. E-RAFT: Dense optical flow from event cameras. In *International Conference on 3D Vision (3DV)*, pp. 197–206. IEEE, 2021.

A. Gnutti, S. Della Fiore, M. Savardi, Y.-H. Chen, R. Leonardi, and W.-H. Peng. Lidar depth map guided image compression model. In *IEEE International Conference on Image Processing (ICIP)*, pp. 1890–1896, 2024.

Y.-H. Ho, C.-P. Chang, P.-Y. Chen, A. Gnutti, and W.-H. Peng. CANF-VC: Conditional augmented normalizing flows for video compression. In *European Conference on Computer Vision (ECCV)*, pp. 207–223. Springer, 2022.

J. Hu, H. Hu, H. Sun, and L. Mu. MoVie: Multimodal video compression with text guidance. In *International Conference on Machine Learning*, 2026.

Z. Hu, G. Lu, and D. Xu. FVC: A new framework towards deep video compression in feature space. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1502–1511, 2021.

B. Huang and T. Ebrahimi. Event data stream compression based on point cloud representation. In *IEEE International Conference on Image Processing (ICIP)*, pp. 3120–3124, 2023.

B. Huang, D. Lazzarotto, and T. Ebrahimi. Evaluation of the impact of lossy compression on event camera-based computer vision tasks. In *Applications of Digital Image Processing XLVI*, volume 12674. International Society for Optics and Photonics, SPIE, 2023.

Z. Jia, B. Li, J. Li, W. Xie, L. Qi, H. Li, and Y. Lu. Towards practical real-time neural video compression. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12543–12552. IEEE, 2025.

H. Jiang, D. Sun, V. Jampani, M.-H. Yang, E. Learned-Miller, and J. Kautz. Super SloMo: High quality estimation of multiple intermediate frames for video interpolation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9000–9008, 2018.

W. Jiang, J. Li, K. Zhang, and L. Zhang. ECVC: Exploiting non-local correlations in multiple frames for contextual video compression. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7331–7341. IEEE, 2025.

X. Jiang, W. Tan, T. Tan, B. Yan, and L. Shen. Multi-modality deep network for extreme learned image compression. In *AAAI Conference on Artificial Intelligence*, volume 37, pp. 1033–1041, 2023.

Y. Jiang, H. Zhang, L. Li, D. Liu, and Z. Li. Sparse point clouds assisted learned image compression. *IEEE Transactions on Circuits and Systems for Video Technology*, 35(5):4000–4010, 2024.

J.-Y. Kang, H. Cho, and K.-J. Yoon. Unleashing the temporal potential of stereo event cameras for continuous-time 3d object detection. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6869–6881. IEEE, 2025.

A. Ke, X. Zhang, T. Chen, M. Lu, C. Zhou, J. Gu, and Z. Ma. Ultra lowrate image compression with semantic residual coding and compression-aware diffusion. In *International Conference on Machine Learning*, 2025.

T. Kim, Y. Chae, H.-K. Jang, and K.-J. Yoon. Event-based video frame interpolation with cross-modal asymmetric bidirectional motion fields. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18032–18042. IEEE, 2023.

J. Kong, J.-H. Kim, and J.-S. Lee. DESSERT: Diffusion-based event-driven single-frame synthesis via residual training. *arXiv preprint arXiv:2512.17323*, 2025.

T. Ladune, P. Philippe, W. Hamidouche, L. Zhang, and O. Déforges. Conditional coding for flexible learned video compression. *arXiv preprint arXiv:2104.07930*, 2021.

H. Lee, M. Kim, J.-H. Kim, S. Kim, D. Oh, and J. Lee. Neural image compression with text-guided encoding for both pixel-level and perceptual fidelity. In *International Conference on Machine Learning*, 2024.

H. Lee, J.-H. Kim, and J.-S. Lee. SLIM: Semantic-based low-bitrate image compression for machines by leveraging diffusion. *arXiv preprint arXiv:2512.18200*, 2025.

J. Li, B. Li, and Y. Lu. Deep contextual video compression. In *Advances in Neural Information Processing Systems*, volume 34, pp. 18114–18125, 2021.

J. Li, B. Li, and Y. Lu. Hybrid spatial-temporal entropy modelling for neural video compression. In *Proceedings of the 30th ACM International Conference on Multimedia*, pp. 1503–1511, 2022.

J. Li, B. Li, and Y. Lu. Neural video compression with diverse contexts. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22616–22626. IEEE, 2023.

J. Li, B. Li, and Y. Lu. Neural video compression with feature modulation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26099–26108. IEEE, 2024.

J. Li, W. Xie, Z. Jia, B. Li, Z. Guo, X. Zhang, and Y. Lu. Ultra-fast neural video compression. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 41311–41321. IEEE, 2026.

S. Lin, Y. Ma, Z. Guo, and B. Wen. DVS-Voltmeter: Stochastic process-based event simulator for dynamic vision sensors. In *European Conference on Computer Vision (ECCV)*, pp. 578–593. Springer, 2022.

I. Loshchilov and F. Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

D. Lu, R. Li, A. Liang, L. Kong, W. Yin, L. X. Ng, B. R Cottereau, C. S. Chane, and W. T. Ooi. EventDrive: Event cameras for vision-language driving intelligence. *arXiv preprint arXiv:2606.18242*, 2026.

G. Lu, W. Ouyang, D. Xu, X. Zhang, C. Cai, and Z. Gao. DVC: An end-to-end deep video compression framework. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10998–11007. IEEE, 2019.

G. Lu, T. Zhong, J. Geng, Q. Hu, and D. Xu. Learning based multi-modality image and video compression. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6073–6082. IEEE, 2022.

A. Mercat, M. Viitanen, and J. Vanne. UVG dataset: 50/120fps 4K sequences for video codec analysis and development. In *ACM Multimedia Systems Conference*, pp. 297–302, 2020.

M. Mohammadi, Z. Wu, and I. Gilitschenski. TESPEC: Temporally-enhanced self-supervised pretraining for event cameras. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7782–7793. IEEE, 2025.

A. Ranjan and M. J Black. Optical flow estimation using a spatial pyramid network. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2720–2729. IEEE, 2017.

H. Rebecq, R. Ranftl, V. Koltun, and D. Scaramuzza. Events-To-Video: Bringing modern computer vision to event cameras. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3852–3861. IEEE, 2019.

O. Rippel, A. G Anderson, K. Tatwawadi, S. Nair, C. Lytle, and L. Bourdev. ELF-VC: Efficient learned flexible-rate video coding. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 14459–14468. IEEE, 2021.

A. Seleem, A. F. R. Guarda, N. M. M. Rodrigues, and F. Pereira. Deep learning-based event data coding: A joint spatiotemporal and polarity solution. *IEEE Open Journal of Signal Processing*, 7:222–237, 2026.

Abdelrahman Seleem, André FR Guarda, Nuno MM Rodrigues, and Fernando Pereira. A double deep learning-based solution for efficient event data coding and classification. *IEEE Access*, 13: 48703–48719, 2025.

A. Sezavar, C. Brites, and J. Ascenso. Learning-based lossless event data compression. In *IEEE International Conference on Visual Communications and Image Processing (VCIP)*, pp. 1–5, 2024.

X. Sheng, J. Li, B. Li, L. Li, D. Liu, and Y. Lu. Temporal context mining for learned video compression. *IEEE Transactions on Multimedia*, 25:7311–7322, 2022.

X. Sheng, P. Chen, L. Xu, S. Wang, and D. O. Wu. Depth map-assisted learned video compression. *IEEE Transactions on Multimedia*, pp. 1–11, 2026.

R. Song, Y. Wang, T. Xu, Z. Liu, Z. Lin, and J. Zhang. Low-latency neural lidar compression with 2d context models. In *International Conference on Learning Representations*, 2026.

G. J. Sullivan, J.-R. Ohm, W.-J. Han, and T. Wiegand. Overview of the high efficiency video coding (HEVC) standard. *IEEE Transactions on Circuits and Systems for Video Technology*, 22(12): 1649–1668, 2012.

L. Sun, C. Sakaridis, J. Liang, P. Sun, J. Cao, K. Zhang, Q. Jiang, K. Wang, and L. Van Gool. Event-based frame interpolation with ad-hoc deblurring. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18043–18052, 2023.

C. Tang, Z. Li, Y. Bian, L. Li, and D. Liu. Neural video compression with context modulation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12553–12563. IEEE, 2025.

C. Tang, X. Sheng, L. Li, D. Liu, and F. Wu. NVC-1B: Scaling up neural video coding models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48(9):10535–10550, 2026.

S. Tulyakov, D. Gehrig, S. Georgoulis, J. Erbach, M. Gehrig, Y. Li, and D. Scaramuzza. Time Lens: Event-based video frame interpolation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16150–16159. IEEE, 2021.

S. Tulyakov, A. Bochicchio, D. Gehrig, S. Georgoulis, Y. Li, and D. Scaramuzza. Time Lens++: Event-based frame interpolation with parametric nonlinear flow and multi-scale fusion. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17734–17743. IEEE, 2022.

H. Wang, W. Gan, S. Hu, J. Y. Lin, L. Jin, L. Song, P. Wang, I. Katsavounidis, A. Aaron, and C.-C. J. Kuo. MCL-JCV: A JND-based H.264/AVC video quality assessment dataset. In *IEEE International Conference on Image Processing (ICIP)*, pp. 1509–1513, 2016.

Y. Wang, L. Lipson, and J. Deng. SEA-RAFT: Simple, efficient, accurate raft for optical flow. In *European Conference on Computer Vision*, pp. 36–54, 2024a.

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.

Z. Wang, J. Guo, and K. Daniilidis. Un-EVIMO: Unsupervised event-based independent motion segmentation. In *European Conference on Computer Vision (ECCV)*, pp. 228–245. Springer, 2024b.

Z. Wang, F. Hamann, K. Chaney, W. Jiang, G. Gallego, and K. Daniilidis. Event-based continuous color video decompression from single frames. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 4977–4987. IEEE, 2025.

W. Weng, Y. Zhang, and Z. Xiong. Event-based video reconstruction using transformer. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2543–2552. IEEE, 2021.

T. Wiegand, G.J. Sullivan, G. Bjontegaard, and A. Luthra. Overview of the H.264/AVC video coding standard. *IEEE Transactions on Circuits and Systems for Video Technology*, 13(7):560–576, 2003.

T. Xue, B. Chen, J. Wu, D. Wei, and W. T Freeman. Video enhancement with task-oriented flow. *International Journal of Computer Vision*, 127(8):1106–1125, 2019.

R. Zhang, P. Isola, A. A Efros, E. Shechtman, and O. Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 586–595. IEEE, 2018.

H. Zheng and W. Gao. End-to-end RGB-D image compression via exploiting channel-modality redundancy. In *AAAI Conference on Artificial Intelligence*, volume 38, pp. 7562–7570, 2024.

J. Zhu, Z. Wan, and Y. Dai. Video frame prediction from a single image and events. In *AAAI Conference on Artificial Intelligence*, volume 38, pp. 7748–7756, 2024a.

L. Zhu, Y. Zheng, Y. Zhang, X. Wang, L. Wang, and H. Huang. Temporal residual guided diffusion framework for event-driven video reconstruction. In *European Conference on Computer Vision (ECCV)*, pp. 411–427. Springer, 2024b.

A PRACTICAL SCENARIOS

In this paper, we mainly consider systems that already transmit or retain events for downstream processing and additionally require RGB video. Event cameras offer high temporal resolution, wide dynamic range, and reduced motion blur (Gallego et al., 2022), with commercial implementations including DAVIS346[3] and Sony–Prophesee-based industrial cameras.[4]

Recent studies demonstrate event-based object detection (Ahmed et al., 2025), continuous-time 3D detection (Kang et al., 2025), and monocular depth estimation (Bartolomei et al., 2025), while TESPEC studies representations across detection, segmentation, and depth estimation (Mohammadi et al., 2025). These capabilities give applications a reason to preserve event streams independently of RGB compression.

For example, a robotic or industrial system may send events to a server for perception while providing RGB video for human inspection. An event-based detector can identify a moving object, but an operator may still need its color, appearance, and surrounding context to interpret the prediction. Displaying detections or trajectories over RGB frames connects machine outputs to an interpretable view of the scene. Both streams may also support machine processing: EventDrive combines events, RGB, and language for driving-related perception, prediction, and planning (Lu et al., 2026). When these computations share a receiving endpoint, the RGB decoder can reuse the events already delivered for perception.

Related cross-modal compression methods use shared LiDAR depth maps (Gnutti et al., 2024) or sparse point clouds (Jiang et al., 2024) to assist RGB image coding, while RangeCM uses RGB camera context for LiDAR compression (Song et al., 2026). ENVC applies this principle to shared event information. With the event stream held fixed across RGB codecs, its transmission or storage cost is common to the compared systems. Our reported rate therefore measures the additional RGB coding cost. Both coding ends use the same available event representation, and App. A.1 examines sensitivity to compression-induced event degradation.

A.1 SENSITIVITY TO EVENT COMPRESSION

Motivation. In the transmission and archival scenarios discussed above, the events available at the receiver may themselves have undergone compression. Existing approaches include temporal aggregation followed by video coding (Brites & Ascenso, 2025), point-cloud-based coding (Huang & Ebrahimi, 2023), and learned lossless and lossy codecs (Sezavar et al., 2024; Seleem et al., 2026), and the JPEG XE initiative further reflects the practical need for efficient event representation and interoperability.[5] Lossless coding preserves the coded representation, whereas lossy coding trades event fidelity for rate reduction. Prior work therefore evaluates compressed events through their effects on downstream tasks, including recognition, video reconstruction, optical flow, and depth estimation (Huang et al., 2023; Seleem et al., 2025). In our setting, the decoded event stream can likewise be reused as shared side information at both RGB coding ends. This motivates evaluating how event coding quality affects its usefulness for RGB reconstruction in ENVC.

Settings. To examine this setting, we evaluate a fixed ENVC checkpoint on the first 96 frames of seven UVG sequences, using synthetic events, an intra period of 32, and quality Q_0. For each inter-frame interval, positive and negative events are represented as separate (x, y, t) point clouds (Huang & Ebrahimi, 2023) and compressed using G-PCC (TMC13)[6]. Timestamps are mapped to an integer grid before coding, and the decoded events undergo the original voxelization and normalization and are supplied to both coding ends without fine-tuning. We vary the event quality factor q_E, corresponding to the quantization scale, where smaller values produce coarser coordinate quantization, and $q_E = 1$ is lossless for the coded integer-coordinate representation.

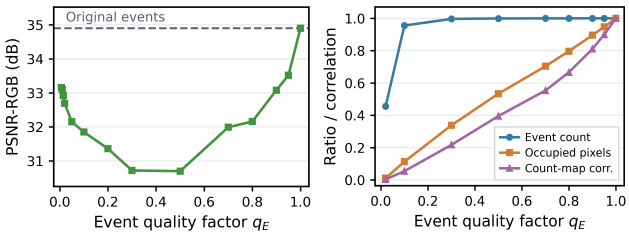
Results. Fig. 9 shows that $q_E = 1$ reproduces the original-event PSNR-RGB of 34.90 dB.

[3] https://docs.inivation.com/hardware/current-products/davis346.html.

[4] https://www.prophesee.ai/event-based-cameras-prophesee-partners/.

[5] https://jpeg.org/jpegxe/index.html.

[6] https://github.com/MPEGGroup/mpeg-pcc-tmc13.



(a) RGB Quality (b) Event Fidelity

Figure 9: **Sensitivity to Event Compression.** (a) PSNR-RGB on UVG, with original-event performance shown by the dashed line. (b) Normalized event and occupied-pixel counts, and event-count map correlation. Smaller q_E indicates coarser quantization.

At $q_E = 0.95$, PSNR-RGB decreases to 33.52 dB, and at $q_E = 0.5$ it reaches 30.70 dB, a 4.20 dB reduction. Event statistics over 21 sampled intervals help characterize this sensitivity. At $q_E = 0.5$, 99.9% of events remain, but the number of occupied pixels falls to 53.4% of its original value and the count-map correlation is only 0.396. Thus, preserving event count does not ensure preservation of the spatial structure used for guidance. The response is also non-monotonic: more aggressive quantization at $q_E = 0.02$ yields 32.69 dB despite poorer event fidelity. One possible explanation is that strongly suppressed guidance can be less disruptive than structured but inaccurate guidance, although this experiment does not isolate that mechanism.

These results reveal sensitivity of the current checkpoint to coordinate-quantized events. They motivate further research into efficient event codecs that preserve task-relevant spatiotemporal structure, together with compression-aware training of event-guided video codecs. Such developments will be important for realizing the value of events in practical multimodal systems.

B Training Details

Progressive Training. We adopt five training stages to first establish event-based prediction and then optimize its integration into the codec. First, we train EFP and EME using frame prediction and motion-compensated reconstruction losses, without rate losses (R). Second, we train RMC to balance motion-compensated reconstruction quality against the rate of the coded motion residual. Third, we train the context refinement modules and conditional frame codec while keeping the event and motion modules fixed. This stage combines frame reconstruction distortion, frame coding rate, and the decoupling loss of DCMVC (Tang et al., 2025). Fourth, we jointly optimize all inter-frame modules, including CME, under the complete RD objective with both motion and frame coding rates. The last stage extends this joint optimization to longer sequences (beyond the seven-frame septuplets) through cascaded training.

Both EME and CME are initialized with the SPyNet weights from DCMVC (Tang et al., 2025). We use the pretrained intra-frame codec of DCVC-DC (Li et al., 2023) for I-frame coding and keep it fixed throughout training.

Cascaded Training. We progressively increase the training sequence length from 8 to 16, 24, and 32 frames and minimize the RD loss averaged over all P-frames in each sequence. To reduce memory consumption, we use Forward Recomputation Backpropagation (FRB), as in DCMVC (Tang et al., 2025). This approach recomputes intermediate activations during backpropagation instead of retaining them all in memory (Chen et al., 2016), enabling optimization over longer prediction chains.

Training Data. We use Vimeo-90K septuplets (Xue et al., 2019) paired with synthetic events. For cascaded training, we use 9,220 training clips and 106 validation clips extracted from the original Vimeo videos[7], each containing 32 frames with paired synthetic events. Events are represented as signed voxel grids with 10 temporal bins. Training uses aligned random crops of 256×256 pixels, reduced to 128×128 for the final 32-frame stage.

Optimization. We train a single multi-rate model with $\lambda \in \{85, 170, 380, 840\}$ and periodic frame weights $(0.5, 1.2, 0.5, 0.9)$ following DCVC-DC (Li et al., 2023). Distortion combines YUV420 and RGB MSE with weights 0.8 and 0.2, respectively. We use AdamW (Loshchilov & Hutter, 2017) with weight decay 0.01 and train on four NVIDIA A100-SXM4-40GB GPUs.

[7] `https://github.com/anchen1011/toflow/blob/master/data/original_vimeo_links.txt`.

C SEQUENCE-LEVEL GAINS AND FAILURES

We examine the 53 sequences in the main benchmarks using the first 96 frames, and an intra period of 32 as in Tab. 2. ENVC achieves negative PSNR-RGB BD-rates relative to DCMVC on 50 of the 53 sequences, with a median of -44.9%. Tab. 4 reports representative gains and the three sequences with positive overall PSNR-RGB BD-rates, together with P-frame-only results and sequence statistics.

Success Cases. Larger savings are generally associated with denser events and greater temporal intensity changes. Across the 53 sequences, the Spearman correlation between event density and PSNR-RGB BD-rate is -0.47. The 13 sequences with the largest savings average -71.1% BD-rate, compared with -7.1% for the 13 with the smallest savings. Their mean event densities are 0.582 and 0.160 events per pixel, and their mean absolute luma changes are 11.34 and 5.39, respectively. MCL-JCV SRC22 and HEVC-C RaceHorses exemplify substantial savings on sequences with pronounced temporal changes. Event density alone, however, does not determine the gain. UVG Beauty achieves the largest saving (-82.4%) with only 0.146 events per pixel, showing that large gains also occur at relatively low event densities.

Table 4: **Sequence-level Results against DCMVC.** PSNR-RGB BD-rates (%) use all frames (All) or P-frames only (P). Events/px denotes mean event density per P-frame. $|\Delta Y|$ denotes mean absolute inter-frame BT.709 luma change on the 0–255 scale.

Dataset	Sequence	All	P	Events/px	$\lvert\Delta Y\rvert$
Representative gains					
UVG	Beauty	-82.4	-89.1	0.146	4.93
MCL-JCV	SRC15	-77.0	-94.7	0.340	7.06
MCL-JCV	SRC22	-73.5	-79.1	1.693	25.44
HEVC-B	Kimono1	-66.7	-78.2	0.435	9.52
HEVC-C	RaceHorses	-64.7	-74.9	0.820	21.15
Limited gains and losses					
UVG	HoneyBee	-14.7	-39.0	0.008	2.21
HEVC-E	Johnny	-6.7	-30.3	0.022	2.32
MCL-JCV	SRC13	$+1.2$	-4.8	0.034	3.36
UVG	Bosphorus	$+1.8$	-7.9	0.017	3.21
UVG	Jockey	$+56.6$	$+42.9$	0.282	11.94

Limited Gains and Failure Cases. Low-activity sequences such as HoneyBee, Bosphorus, and SRC13 generate few events and show limited gains or slight losses. Bosphorus and SRC13 have slightly positive overall BD-rates, but retain negative P-frame-only BD-rates of -7.9% and -4.8%. The largest loss occurs on Jockey, with $+56.6\%$ overall and $+42.9\%$ P-frame-only BD-rate despite an event density of 0.282 events per pixel. Its complex motion, occlusions, and repetitive textures are plausible sources of difficulty, although this analysis does not isolate their contributions. Jockey nevertheless achieves an LPIPS-based BD-rate of -52.4%, demonstrating that its relative performance depends on the quality metric.

Limited Gains on HEVC-E. HEVC-E consists of video-conferencing scenes with low temporal activity, producing only 0.027 events per pixel on average. ENVC achieves a P-frame-only PSNR-RGB BD-rate of -32.0%, but the overall saving is limited to -7.8%. This difference is partly explained by the substantial contribution of intra coding. At Q_0, I-frames account for 56.2% of the total RGB rate of ENVC, compared with 30.1% for DCMVC, diluting the overall benefit of P-frame rate savings.

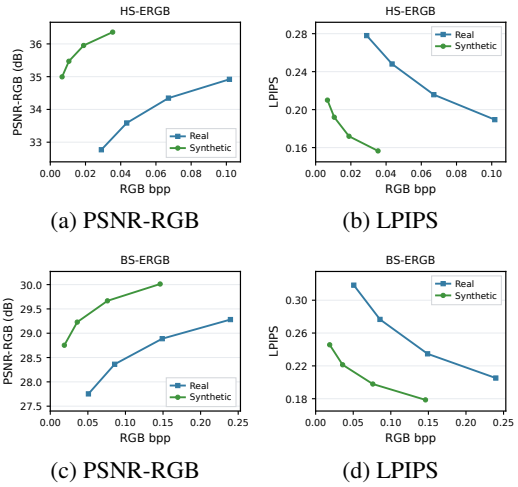
The allocation of bits to P-frames also differs between the two codecs. At Q_0, ENVC uses 66% fewer P-frame bits than DCMVC, with a mean P-frame PSNR of 37.71 dB versus 38.80 dB. Both codecs achieve the same mean I-frame PSNR of 38.02 dB. Thus, the mean P-frame quality of DCMVC exceeds its intra-frame quality, whereas that of ENVC remains lower. This pattern is consistent with further refinement of the intra reconstruction by DCMVC in these low-activity scenes. The results suggest that improving the reconstruction of largely static content, together with adapting event guidance to sparse-event inputs, is a promising direction for increasing coding efficiency on HEVC-E.

D GENERALIZATION TO REAL EVENTS

We evaluate whether ENVC, trained with synthetic events, can be applied to real sensor events without additional fine-tuning. By comparing real and synthetic events for the same RGB sequences, we examine how the event source affects RGB coding performance.

Evaluation Protocol. We use 15 test sequences from HS-ERGB (Tulyakov et al., 2021) and 17 validation sequences from BS-ERGB (Tulyakov et al., 2022). Each sequence is evaluated over 32 frames with an intra period of 32. The real-event condition uses sensor recordings, while the synthetic-event condition uses events generated from the same RGB frames following Sec. 3.

RD Performance. At corresponding quality settings, real events yield higher RGB rates, lower PSNR-RGB, and higher LPIPS than synthetic events on both datasets (Fig. 10). The degradation in both rate and distortion suggests that the synthetic-trained codec uses real observations less effectively for compression. Thus, ENVC can operate on real sensor events without fine-tuning, but its performance under synthetic guidance does not fully transfer to real recordings. This comparison demonstrates applicability to real inputs while highlighting a remaining gap in generalization across event domains.



(a) PSNR-RGB (b) LPIPS

(c) PSNR-RGB (d) LPIPS

Figure 10: **Generalization to Real Events.** RD curves with real and synthetic events on HS-ERGB (top) and BS-ERGB (bottom).

Understanding the Domain Gap. Synthetic events are denser on average ($\times 1.54$ and $\times 1.59$ in Tab. 1). However, App. C shows that event density alone does not determine RGB coding gains. Differences in spatiotemporal event statistics and compatibility with the training distribution may also contribute to the real–synthetic performance gap.

Implications. ENVC can be applied to real sensor events without fine-tuning. The observed performance difference suggests room to improve transfer from synthetic training data to real sensor recordings. Training with more diverse event statistics and adaptation to real data are promising directions for improving this transfer.

E Analysis of Learned Event Gating

Analysis Setup. We extend the analysis in Sec. 5.3 using the ENVC model at quality factor Q_1. The gate and feature analyses use 12 sequences, with two from each of six dataset groups, as listed in Tab. 5. The constant-gate interventions use the four sequences reported in Tab. 6. With an intra period of 32, we analyze the first 15 P-frames following the initial I-frame of each sequence. All gates are initialized to $\sigma(4) \approx 0.982$, while the adapters initially output zero. RMS ratios normalize adapter-feature magnitudes by those of the temporal context before injection.

Spatial and Temporal Variation. For each frame, we compute the spatial standard deviation and spatial mean of the gate activations. For each sequence, we average the spatial standard deviations over the first 15 P-frames and compute the temporal standard deviation of the frame-wise means. Fig. 11 shows the averages across 12 sequences as bars and individual sequences as circles. Spatial variation is larger at all three resolutions, indicating spatially nonuniform gating with comparatively stable frame-averaged activation.

Table 5: **Selected Sequences for the Gate Activation Analyses.** Each sequence contributes its first 15 P-frames.

Dataset	Sequences
UVG	Beauty, Jockey
MCL-JCV	SRC02, SRC16
HEVC-B	BasketballDrive, ParkScene
HEVC-C	BasketballDrill, PartyScene
HEVC-D	BQSquare, BasketballPass
HEVC-E	FourPeople, Johnny

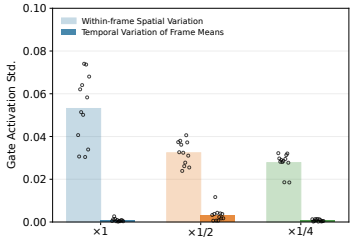


Figure 11: **Spatial and Temporal Variation of Event Gates.** Light and dark bars denote spatial and temporal variation, respectively.

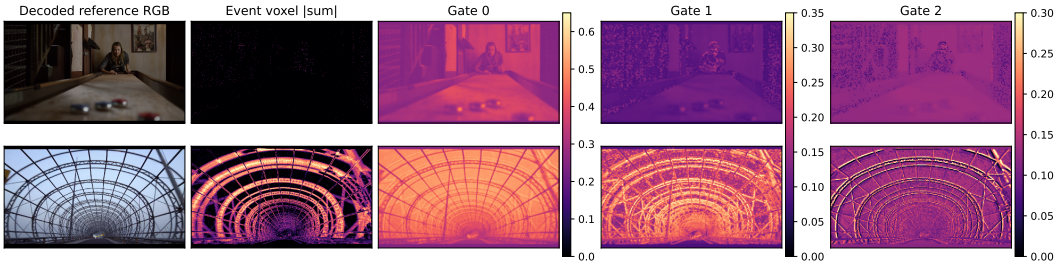


Figure 12: **Spatial Patterns of Learned Event Gates.** MCL-JCV SRC01 (top) and SRC10 (bottom). Columns show the decoded reference, event voxel visualization, and gate maps at three scales.

Spatial Gate Patterns. Fig. 12 visualizes the learned gates on two additional MCL-JCV sequences with contrasting event activity. The maps exhibit different spatial patterns across the three scales.

In SRC10, Gates 1 ($\times 1/2$) and Gate 2 ($\times 1/4$) show pronounced responses along curved scene structures, while Gate 0 ($\times 1$) exhibits a broader response. These examples complement the variation statistics by illustrating how the learned modulation is spatially organized.

Constant-gate Interventions. Tab. 6 reports the results of replacing all learned gates with $G = 0$ or $G = 1$ at inference, without retraining. With $G = 0$, PSNR drops by 2.42–2.94 dB and estimated bpp increases on all four sequences; on Jockey, bpp rises from 0.005 to 0.021. With $G = 1$, PSNR drops by 2.30–3.89 dB, while bpp increases on three sequences. BQSquare is the exception: its bpp decreases from 0.030 to 0.021, alongside a PSNR decrease from 31.779 to 28.937 dB. For BasketballDrill, $G = 1$ yields nearly the same bpp as the learned gates (0.017), but PSNR falls by 2.30 dB. These interventions show that fixed gates change the measured operating points. They do not isolate the effect of spatial variation, since replacing the gates also changes injection magnitude, and moreover, $G = 0$ disables only the gated injection branch, leaving other event pathways active.

Table 6: **Constant-gate Interventions.** PSNR (dB) and estimated codec bpp over the first 15 P-frames. Bpp rates exclude the initial I-frame and event transmission.

	Learned Gates		$G = 0$		$G = 1$	
Sequence	PSNR	bpp	PSNR	bpp	PSNR	bpp
BQSquare	31.779	0.030	28.841 (-2.94)	0.073	28.937 (-2.84)	0.021
Jockey	36.351	0.005	33.801 (-2.55)	0.021	32.461 (-3.89)	0.014
BasketballDrill	33.102	0.017	30.679 (-2.42)	0.032	30.799 (-2.30)	0.017
HoneyBee	36.715	0.007	33.772 (-2.94)	0.013	33.729 (-2.99)	0.009

F Learned Use of Spatiotemporally Aligned Events

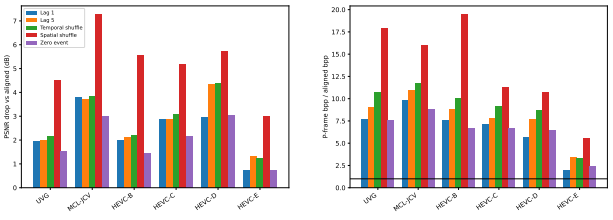
We investigate whether ENVC learns to exploit the spatiotemporal correspondence between events and RGB frames. By removing or disrupting this correspondence at inference, we examine how aligned event observations contribute to the trained codec's reconstruction quality and coding efficiency.

Evaluation Protocol. We evaluate a fixed ENVC checkpoint at Q_1 without retraining on 13 selected sequences from HEVC-B/C/D/E, UVG, and MCL-JCV. Each sequence is paired with synthetic events and contains one I-frame followed by 31 P-frames. All conditions use aligned events for the first five P-frames, after which we intervene on the remaining 26 P-frames. Each condition propagates its own reconstructed state. We average PSNR-RGB and entropy-estimated P-frame bpp over these 26 frames and then equally across sequences.

Event Interventions. The `aligned` condition supplies events from the interval immediately preceding the target frame. For `lag 1` and `lag 5`, we instead use the event interval from one or five frame periods earlier. For `temporal shuffle`, a fixed random permutation reassigns the 26 event intervals to target frames, with none remaining at its original target. This preserves the set of observations while disrupting their temporal correspondence. For `spatial shuffle`, one fixed permutation of the $H \times W$ locations is applied to every temporal bin and intervened frame within a sequence. This preserves voxel values and temporal-bin membership while disrupting spatial struc-

Table 7: **Contribution of Aligned Event Guidance.** Averages over 13 selected sequences. ΔPSNR denotes the drop relative to aligned inputs. Rate ratios are averaged over sequences.

Input	PSNR (dB)	P-frame bpp	ΔPSNR (dB)	Rate ratio
Aligned	34.794	0.01085	0.000	1.00×
Zero event	32.844	0.07611	1.950	6.52×
Lag 1	32.446	0.07752	2.348	6.75×
Lag 5	32.122	0.09238	2.672	8.06×
Temporal shuffle	32.027	0.10400	2.767	9.08×
Spatial shuffle	29.634	0.14133	5.159	13.83×



(a) PSNR-RGB Drop (b) Relative Bpp

Figure 13: **Effects of Disrupting Event Correspondence.** Selected-sequence averages within each dataset. (a) PSNR-RGB drop relative to aligned events. (b) Mean per-sequence P-frame bpp ratio relative to aligned events.

ture and correspondence with RGB. The `zero-event` condition sets the input event voxel grid to zero.

Learned Use of Event Correspondence. Tab. 7 and Fig. 13 show that aligned events enable higher reconstruction quality at lower coding rates. Shifting the input by one frame reduces PSNR-RGB by 0.72–3.79 dB across the selected dataset groups, with mean per-sequence bpp ratios of 1.99–9.86. Both shuffle conditions perform worse than zero-event inputs in PSNR and bpp on every evaluated sequence. These results support that ENVC has learned to exploit event observations that correspond to the target frame, with benefits extending to both reconstruction and coding efficiency.

Content-dependent Contribution. The contribution of aligned events varies with video content. Removing or temporally reassigning events has smaller effects on the selected HEVC-E sequences than on the MCL-JCV examples. Increasing the delay also produces sequence-dependent changes. These observations are consistent with different degrees of reliance on event guidance across video content.

G Bit Allocation Analysis

Evaluation Protocol. We evaluate ENVC and DCMVC on seven UVG sequences using the first 96 frames and an intra period of 32.

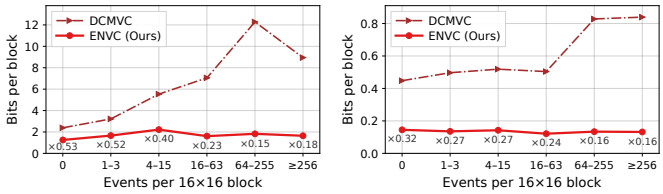
For Tab. 8, we linearly interpolate each latent stream's log-bpp against dataset-average PSNR-RGB using Q_0–Q_3 and average over 50 equally spaced points within the common range of 35.445–37.394 dB. Motion and frame coding rates each include their corresponding hyperlatents. For the spatial analysis, we group 16×16 P-frame blocks by event count at Q_1, where ENVC and DCMVC achieve mean PSNR-RGB values of 35.71 and 36.40 dB, respectively. This block comparison therefore does not match reconstruction quality.

Component-wise Rate Reduction. Over the common quality range, ENVC reduces motion and conditional frame coding rates by 67.9% and 39.8%, respectively (Tab. 8). Frame coding accounts for 86.4% of the P-frame bit savings, showing that the larger absolute reduction occurs beyond the motion stream. These results are consistent with event guidance reducing coding costs through both motion estimation and temporal context construction. Matching average quality selects higher quality settings for ENVC, increasing the shared intra codec's rate and partially offsetting the P-frame savings.

Spatial Distribution of Savings. In Fig. 14, the frame-latent rate of ENVC is 53% of DCMVC's in blocks without events and 15–18% in blocks with at least 64 events. Blocks with at least 16 events constitute 35.4% of all evaluated blocks but account for 69.6% of the frame-latent bit savings. ENVC's motion rate remains within 0.12–0.15 bits per block, while DCMVC uses 0.83–0.84 bits in the two highest event-count groups. The larger relative reductions in regions with more event activity are consistent with ENVC exploiting the supplied temporal observations.

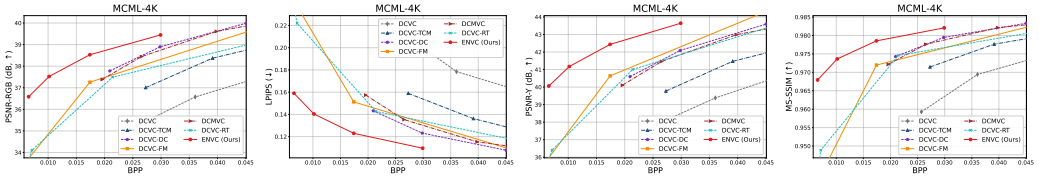
Table 8: **RGB Coding Rate Contributions.** Rates (10^{-3} bpp) on UVG are averaged over the common PSNR-RGB range after interpolation and normalized over all coded frames. Motion and frame coding include their respective hyperlatents. Changes are relative to DCMVC.

Component	DCMVC	ENVC	Change
Motion	2.038	0.655	−67.9%
Frame coding	22.093	13.292	−39.8%
I-frame	5.390	7.878	+46.2%
Total	29.521	21.825	−26.1%



(a) Frame Latent Bits (b) Motion Bits

Figure 14: **Spatial Distribution of Coding Rates at Q_1.** Mean bits per 16×16 P-frame block on UVG, grouped by event count within the frame interval. (a) Frame latent y. (b) Motion latent and hyperlatent. Annotations show ENVC/DCMVC ratios. Motion hyperlatent bits are distributed uniformly across the corresponding blocks.



(a) PSNR-RGB (b) LPIPS (c) PSNR-Y (d) MS-SSIM

Figure 15: **Rate–Distortion Performance on MCML-4K (Cheon & Lee, 2018).** Results are reported using PSNR-RGB, LPIPS, PSNR-Y, and MS-SSIM.

H Additional Datasets and Quality Metrics

Evaluation on MCML-4K. To assess whether ENVC's compression benefits extend to higher spatial resolutions, we additionally evaluate it on MCML-4K (Cheon & Lee, 2018), which contains 10 source sequences at a resolution of 3840×2160 pixels. As shown in Fig. 15, ENVC achieves higher PSNR-RGB, PSNR-Y, and MS-SSIM, together with lower LPIPS, than the compared codecs over the overlapping bitrate range. These results support the effectiveness of event-guided compression for 4K video in terms of both reconstruction fidelity and perceptual quality.

Additional Quality Metrics. We additionally evaluate the compared methods using PSNR-Y and MS-SSIM (Wang et al., 2003). Figs. 16 and 17 show the corresponding rate–distortion curves, while Tab. 9 summarizes the BD-rate results relative to DCMVC. ENVC achieves average BD-rate savings of 53.51% in PSNR-Y and 40.00% in MS-SSIM across the six benchmarks. These results show that the gains of ENVC are not limited to the PSNR-RGB and LPIPS metrics used in the main evaluation.

I Additional Qualitative Results

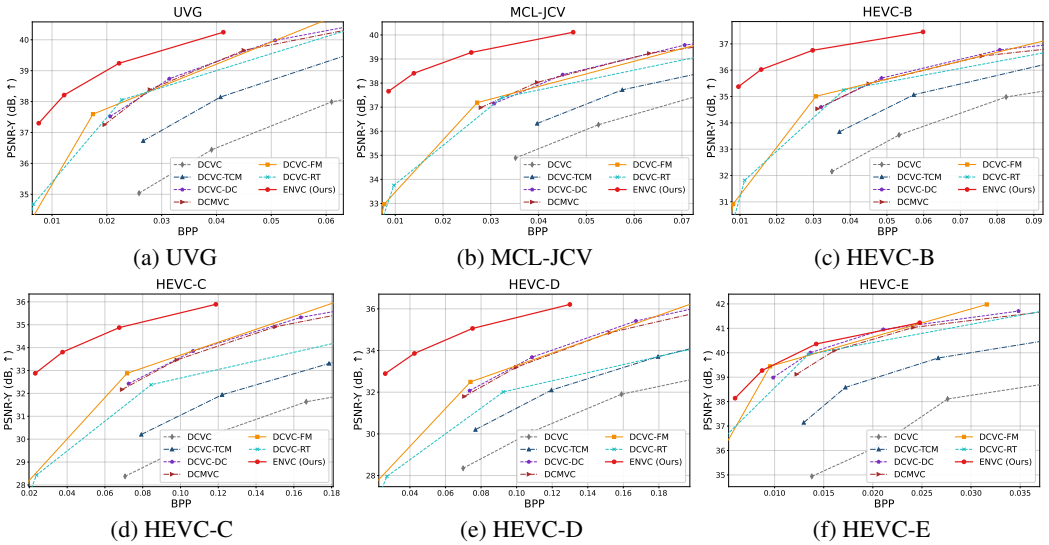
Real and Synthetic Event Comparison. Figs. 18 and 19 provide qualitative comparisons on HS-ERGB (Tulyakov et al., 2021) and BS-ERGB (Tulyakov et al., 2022), respectively. For each sequence, we visualize the real and synthetic events, the corresponding VFPSIE (Zhu et al., 2024a) predictions obtained from the same reference frame, and their difference and prediction-error maps. These examples complement the qualitative comparison in Fig. 3 by showing the spatial discrepancies between real and synthetic events and how they propagate to frame prediction.

Qualitative Reconstruction Comparison. Fig. 20 compares ENVC and DCMVC at similar RGB coding rates. In these selected examples, ENVC better preserves local textures and object boundaries, complementing the quantitative RD results.

Frame Changes and Event Activity. Fig. 21 visualizes the RGB frame pair, their frame-difference map, and the intervening events for the same sequences used in Fig. 20. The examples illustrate how event activity is concentrated around spatio-temporal changes occurring between consecutive RGB frames.

Table 9: **BD-Rate (%) against DCMVC using Additional Quality Metrics.** Each comparison uses PSNR-Y (left) and MS-SSIM (right). Negative values indicate bitrate savings.

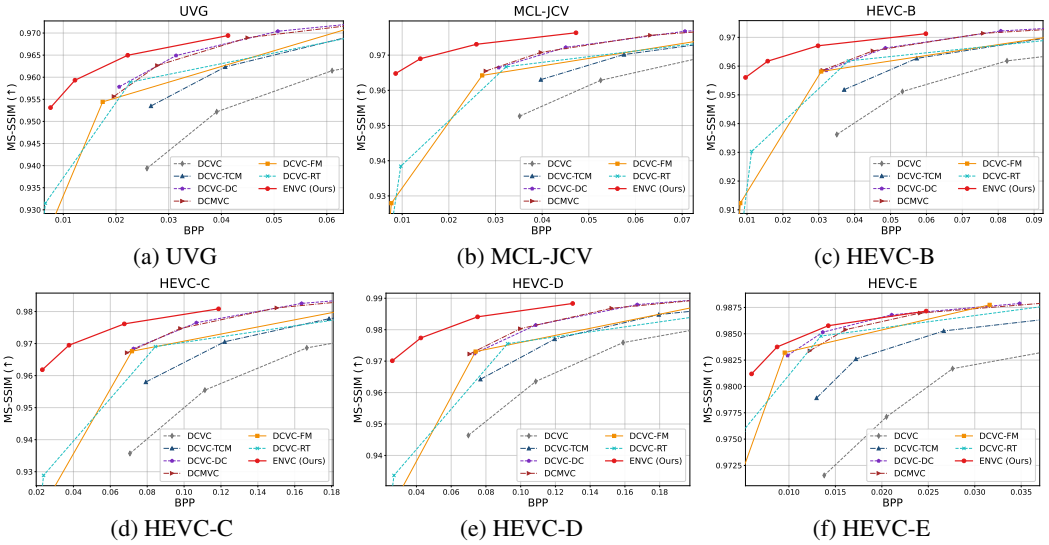
Method	UVG	MCL-JCV	HEVC-B	HEVC-C	HEVC-D	HEVC-E	Average
DCMVC	0.00 / 0.00	0.00 / 0.00	0.00 / 0.00	0.00 / 0.00	0.00 / 0.00	0.00 / 0.00	0.00 / 0.00
DCVC	+150.24 / +128.29	+117.41 / +95.52	+123.41 / +122.59	+191.51 / +125.20	+122.41 / +95.49	+254.76 / +202.22	+159.96 / +128.22
DCVC-TCM	+55.06 / +41.91	+56.62 / +44.08	+45.82 / +43.67	+88.70 / +51.40	+56.26 / +37.91	+84.95 / +68.74	+64.57 / +47.95
DCVC-DC	−1.66 / −3.19	−0.59 / −0.14	−4.01 / −0.60	−3.25 / −2.29	−4.30 / +1.07	−11.13 / −9.72	−4.16 / −2.48
DCVC-FM	−17.14 / +6.74	−12.00 / +8.28	−21.34 / +7.49	−16.64 / +8.31	−12.98 / +4.07	−24.72 / −7.61	−17.47 / +4.55
DCVC-RT	−7.22 / +9.95	+1.75 / +12.27	−8.32 / +12.97	+26.85 / +25.06	+36.53 / +24.50	−10.10 / −1.59	+6.58 / +13.86
ENVC (Ours)	**−47.47 / −37.10**	**−64.49 / −52.60**	**−69.16 / −46.46**	**−59.86 / −39.46**	**−58.34 / −43.74**	**−21.75 / −20.65**	**−53.51 / −40.00**



(a) UVG (b) MCL-JCV (c) HEVC-B

(d) HEVC-C (e) HEVC-D (f) HEVC-E

Figure 16: **Rate–Distortion Performance in PSNR-Y.** PSNR-Y (↑) versus bitrate on six benchmarks.



(a) UVG (b) MCL-JCV (c) HEVC-B

(d) HEVC-C (e) HEVC-D (f) HEVC-E

Figure 17: **Rate–Distortion Performance in MS-SSIM.** MS-SSIM (↑) versus bitrate on six benchmarks.

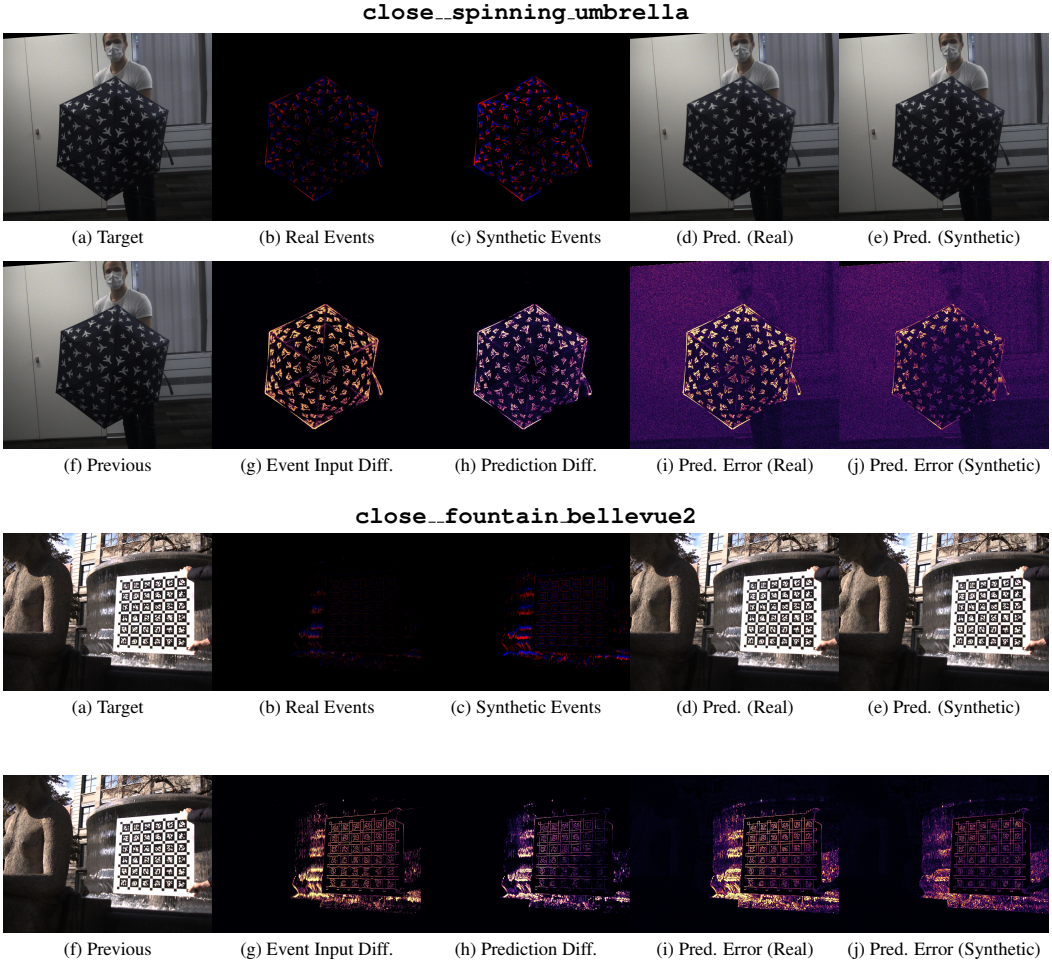


Figure 18: **Additional Qualitative Comparisons on HS-ERGB (Tulyakov et al., 2021).** For each sequence, we compare the target frame, real and synthetic events, predictions obtained from each event source, the previous frame, differences between the event inputs and predictions, and prediction errors with respect to the target.

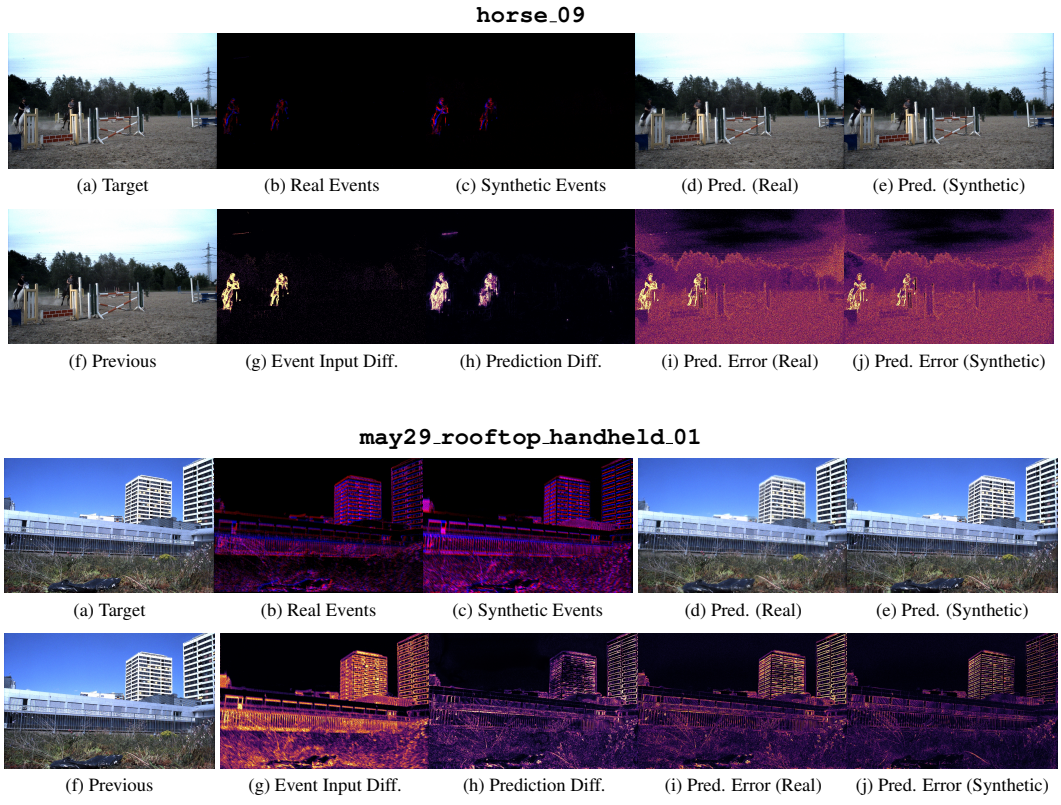
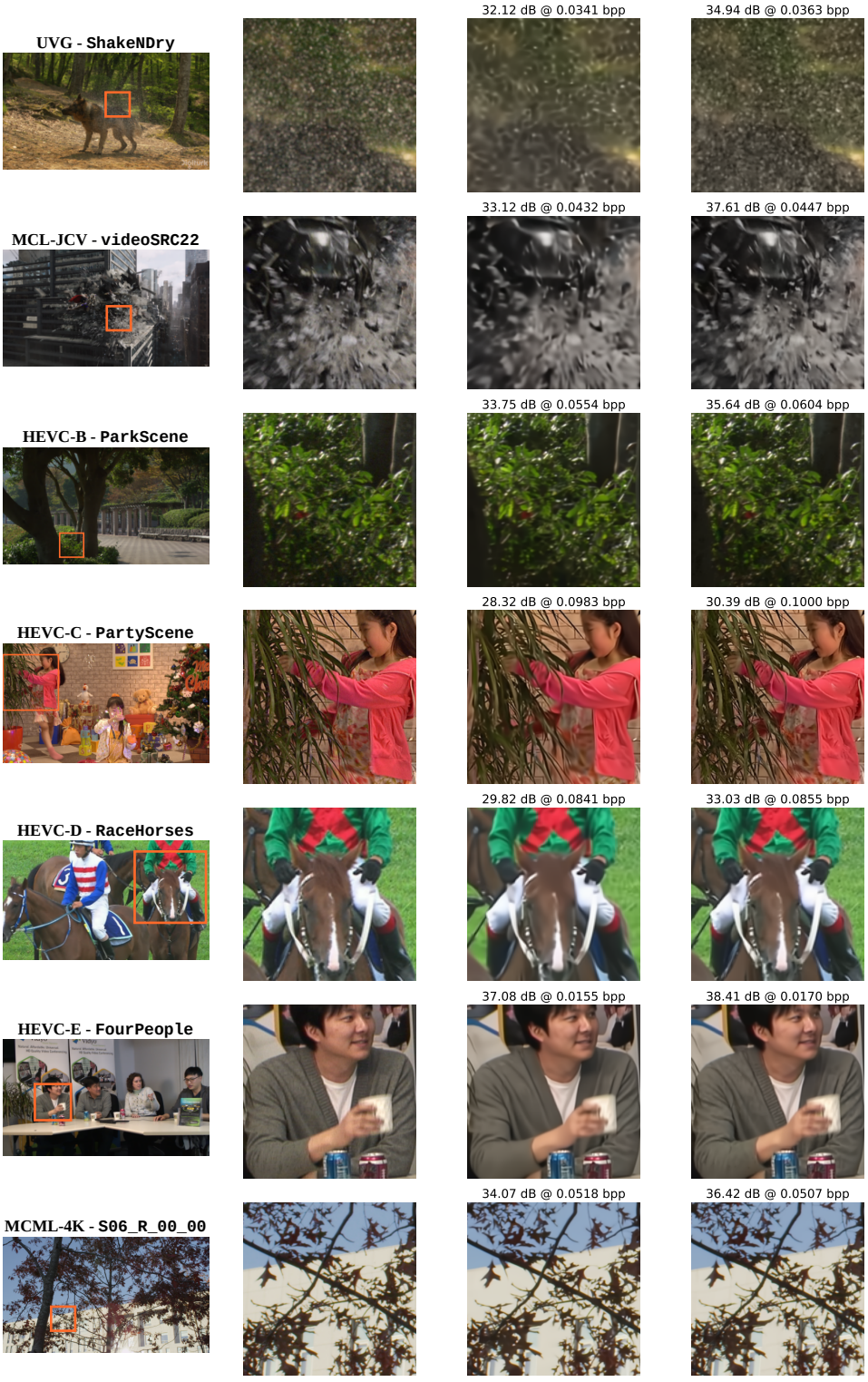


Figure 19: **Additional Qualitative Comparisons on BS-ERGB (Tulyakov et al., 2022).** For each sequence, we compare the target frame, real and synthetic events, predictions obtained from each event source, the previous frame, differences between the event inputs and predictions, and prediction errors with respect to the target.



(a) Crop location (b) Original (c) DCMVC (d) ENVC (Ours)

Figure 20: **Closest-rate Qualitative Comparisons.** From top to bottom, examples are shown for UVG, MCL-JCV, HEVC-B, HEVC-C, HEVC-D, HEVC-E, and MCML-4K, respectively. The orange boxes in (a) indicate the regions displayed in (b)–(d). Note that reported PSNR and bpp are both measured over the full frame.



(a) Previous frame (b) Target frame (c) Frame difference (d) Events

Figure 21: **Frame Changes and Corresponding Events.** From top to bottom, examples are shown for UVG, MCL-JCV, HEVC-B, HEVC-C, HEVC-D, HEVC-E, and MCML-4K respectively. For each sequence, we show the previous frame $\mathbf{X}_{t-1}$, target frame $\mathbf{X}_t$, their frame-difference map, and the synthetic events $\mathbf{E}_{t-1:t}$ generated for the corresponding interval.