

EvDiff: Event-Based Video Reconstruction using One-Step Diffusion Models

Weilun Li^{1,2,*}, Lei Sun^{2,*†}, Ruixi Gao¹, Qi Jiang¹, Yuqin Ma¹,
Kaiwei Wang^{1†}, Ming-Hsuan Yang^{3,4}, Luc Van Gool², Danda Pani Paudel²
¹Zhejiang University ²INSAT, Sofia University “St. Kliment Ohridski” ³UC Merced ⁴Google DeepMind



Project Repository
github.com/LiWeilun/EvDiff

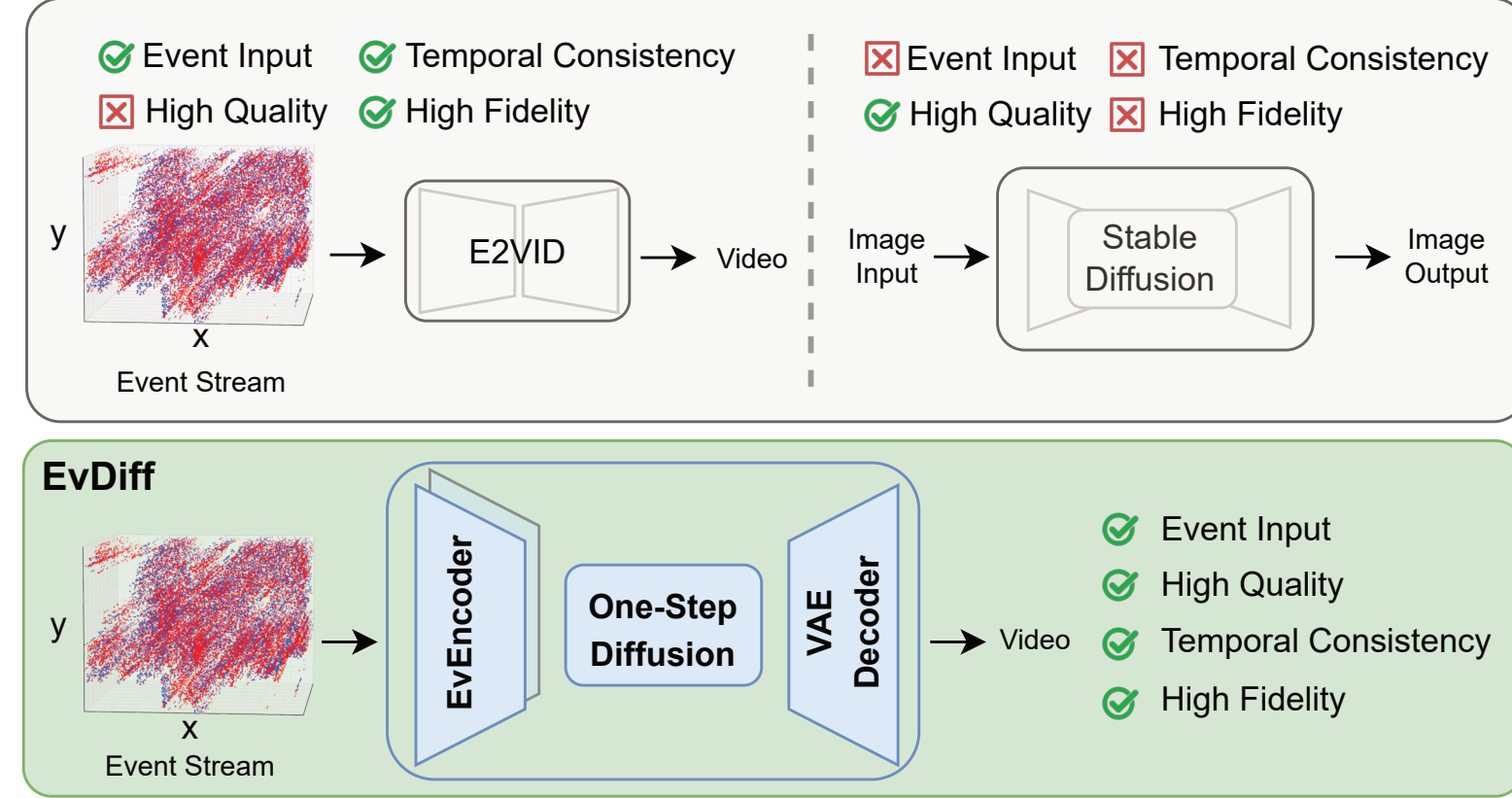
ECCV 2026 | MALMÖ | SEPT 8-12

1. MOTIVATION

Event cameras offer microsecond latency and high dynamic range, but record only **brightness changes**. Recovering photorealistic video is severely under-determined.

- Events encode brightness changes, not absolute intensity or color.
- Large-scale paired event-RGB training data are scarce.
- Multi-step diffusion is impractical for high-frame-rate video.

From sparse events to one-step reconstruction



Qualitative evidence across real-world datasets



Core hypothesis

Learn an image restoration prior from E2VID-style surrogate pairs, then align sparse event latents to it.

Design targets

Faithful reconstruction
consistent with events

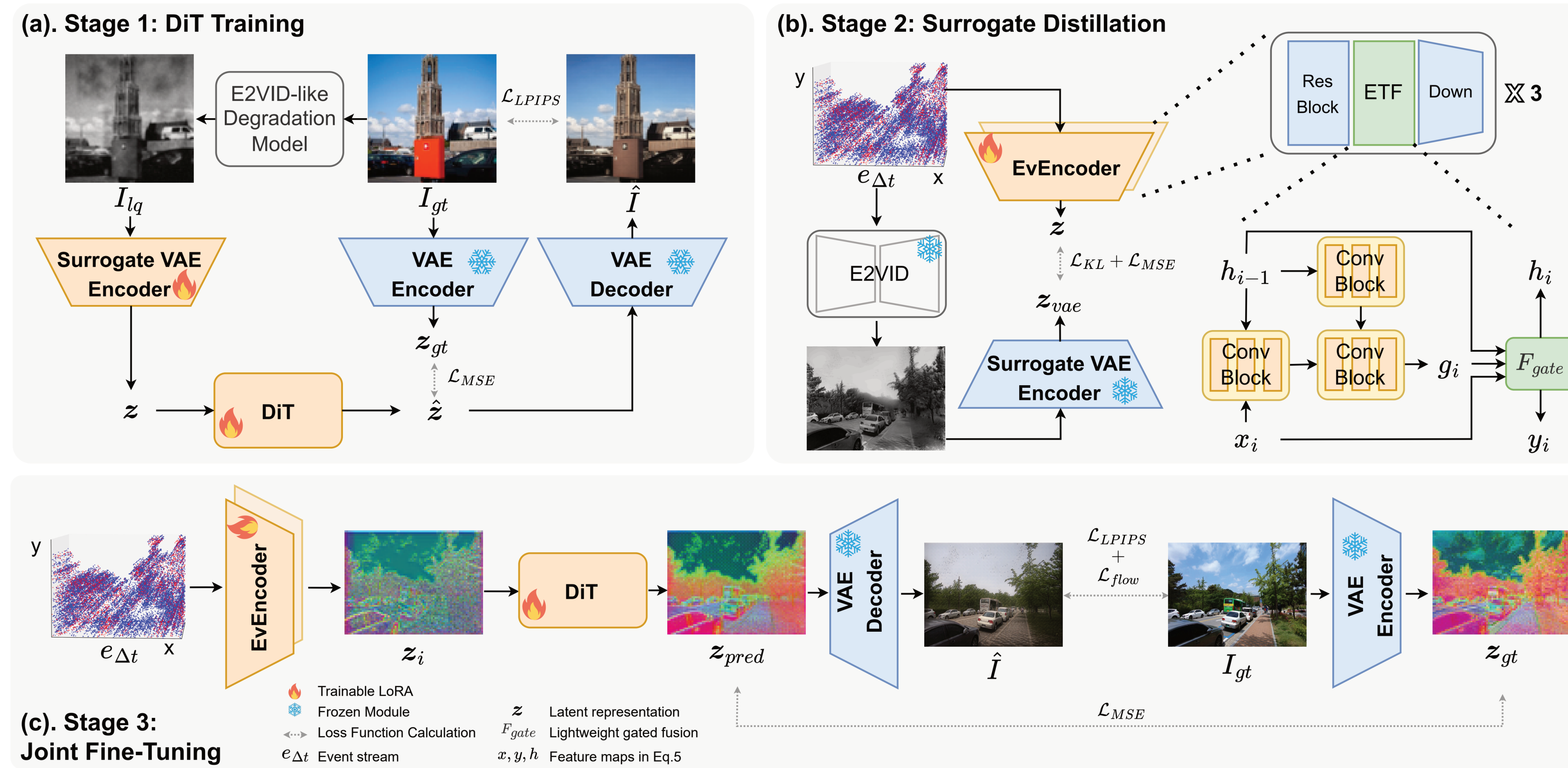
Perceptual realism
natural texture & color

Efficient inference
one diffusion step

2. KEY IDEAS & CONTRIBUTIONS

- Surrogate degradation bridge**
E2VID-style degradation turns 1.8M Places365 images into scalable LQ-HQ supervision. This bridge transfers natural-image restoration priors to event reconstruction without requiring large-scale paired event-RGB data.
- Latent-space transfer**
The Surrogate VAE Encoder provides target latents for EvEncoder. Distillation aligns event representations with the learned restoration latent space, giving OSDiff a clean and stable input.
- Efficient temporal generation**
ETF recurrently fuses history while OSDiff reconstructs each frame in one step. Its lightweight gate preserves temporal information flow without the cost of full spatiotemporal attention.

3. METHOD OVERVIEW



Stage 1 learns a surrogate image prior; Stage 2 aligns event latents; Stage 3 jointly adapts on simulated event-RGB sequences.

1. DiT Pretraining

Train an SD3-based one-step refiner and Surrogate VAE Encoder on 1.8M synthetic LQ-HQ pairs using latent MSE and LPIPS.

2. Surrogate Distillation

Use E2VID reconstructions to define surrogate image latents, then distill EvEncoder with KL and MSE losses.

3. Joint Fine-Tuning

Jointly fine-tune EvEncoder, ETF and OSDiff on REDS sequences with v2e-simulated events using perceptual, flow and latent losses.

E2VID-style surrogate degradation

Event ambiguity: blotchy textures
Sensor: persistence, edge corruption, noise
Reconstruction: motion blur and detail loss

Efficient Temporal Fusion

$y_i = g_i \odot x_i + (1 - g_i) \odot h_{i-1}$
Adaptive recurrent fusion carries history without full spatiotemporal attention; cost scales linearly with frame count.



8. TAKEAWAYS

MSE

32.2%
LOWER MSE
BS-ERGB vs. second-best

FID

44.9%
LOWER FID
DSEC vs. second-best

FLOPs

12.2 ×
FEWER FLOPs
than 10-step ControlNet at 512²

RGB

COLORVIDEO
FROM MONO EVENTS
in one diffusion step

EvDiff uses surrogate training to transfer image restoration priors to events, delivering efficient one-step color reconstruction with strong fidelity and perceptual quality.

4. MAIN RESULTS — BS-ERGB

Method	MSE↓	SSIM↑	LPIPS↓	FID↓	FVD↓
E2VID	0.1175	0.3313	0.5518	276	1688
FireNet	0.0857	0.3302	0.5328	271	1909
E2VID+	0.0717	0.3710	0.4397	238	1652
FireNet+	0.0827	0.3080	0.4903	313	1562
SPADE-E2VID	0.0891	0.3235	0.6715	347	2343
SSL-E2VID	0.0798	0.3448	0.6169	312	1907
ET-Net	0.0683	0.3557	0.4566	270	1569
HyperE2VID	0.0756	0.3489	0.4598	275	1673
EvDiff	0.0463	0.3394	0.4023	148	984

EvDiff leads four of five metrics: 32.2% lower MSE and 37.8% lower FID.

5. MAIN RESULTS — DSEC

Method	MSE↓	SSIM↑	LPIPS↓	FID↓	FVD↓
E2VID	0.1042	0.3393	0.5427	239	1541
FireNet	0.0927	0.3478	0.5670	262	1492
E2VID+	0.0881	0.2812	0.5138	248	1397
FireNet+	0.0913	0.2268	0.5655	291	1592
SPADE-E2VID	0.0601	0.4578	0.4911	234	1277
SSL-E2VID	0.1055	0.3303	0.5565	326	1521
ET-Net	0.0737	0.2953	0.5232	259	1496
HyperE2VID	0.0691	0.2977	0.5297	272	1445
EvDiff	0.0476	0.3677	0.4226	129	1491

Best MSE, LPIPS and FID; second-best SSIM. Among the compared baselines, only EvDiff reconstructs color from monochromatic events. MSE, SSIM and LPIPS use grayscale outputs for fair comparison on both datasets.

6. EFFICIENCY

Model	Steps	Params	FLOPs 512 ²	FLOPs 1024 ²
EvDiff	1	2.19G	2.18T	9.56T
ControlNet	10	8.23G	26.62T	87.15T

3.8× fewer parameters and 12.2× fewer FLOPs than 10-step SD3-ControlNet at 512² (single frame).

7. SUMMARY & LIMITATIONS

- Surrogate training transfers large-scale image restoration priors to events.
 - On BS-ERGB and DSEC, EvDiff leads MSE, LPIPS and FID while generating color in one step.
- Residual temporal flicker.** Despite recurrent conditioning, slight appearance flicker remains in dynamic scenes.
- Backbone-induced grid artifacts.** The patch-based DiT can leave faint periodic structures.
- Next: temporal diffusion priors and artifact-aware backbone design.