THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY (GUANGZHOU)

NeurIPS

LaSe-E2V: Towards Language-guided Semantic-Aware Event-to-Video Reconstruction

Kanghao Chen, Hangyu Li, Jiazhou Zhou, Zeyu Wang, Lin Wang

Artificial Intelligence Thrust, HKUST(GZ)

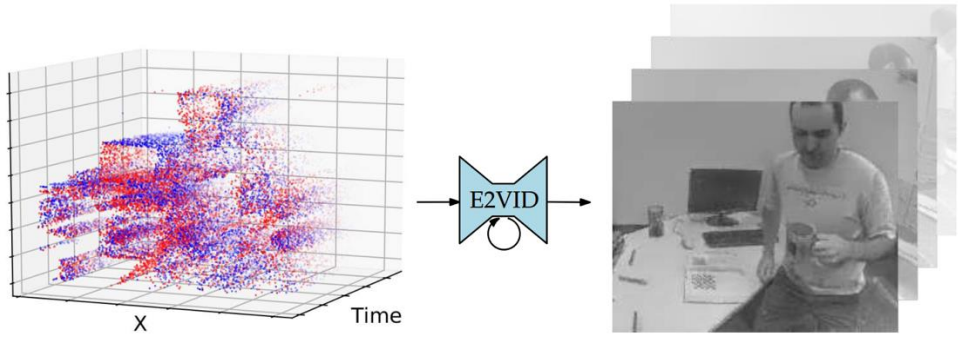
Background

How to take the advantages of both worlds?

Event Camera

- High temporal resolution
- High dynamic range
- No absolute intensity
- Not easy to develop algorithms

Event-based Video Reconstruction

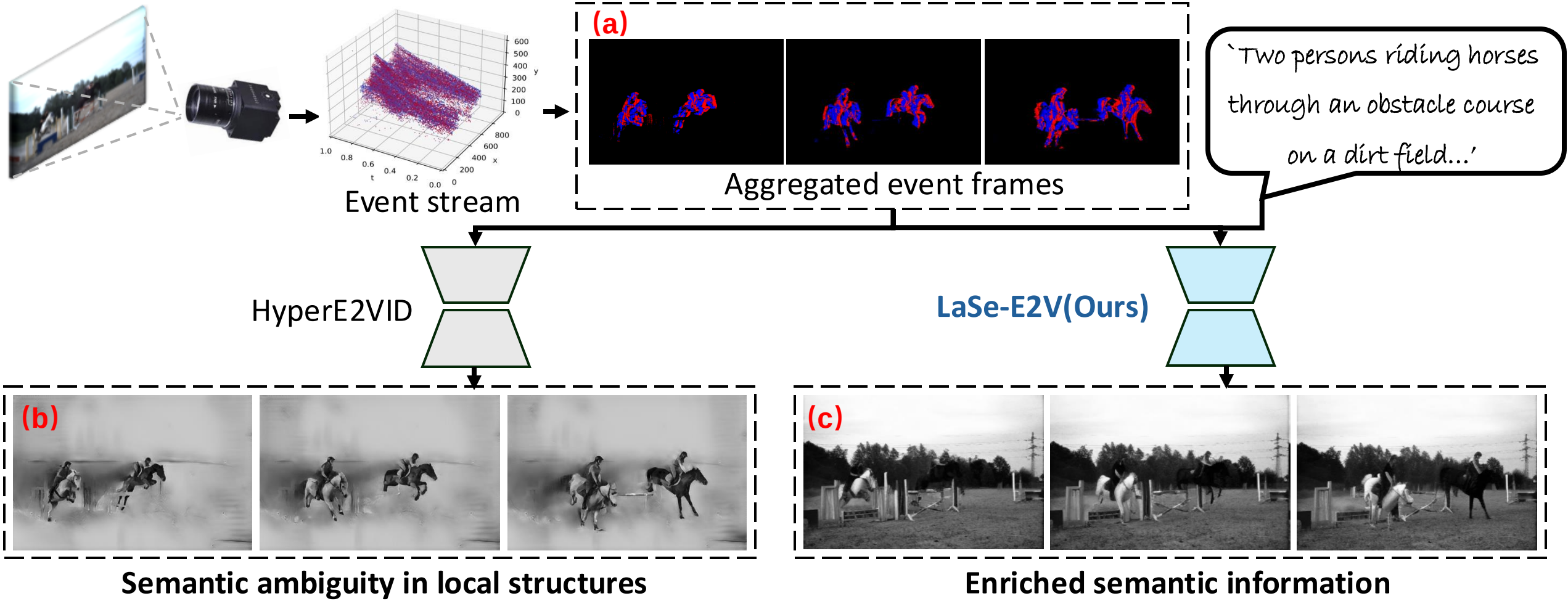


Standard Camera

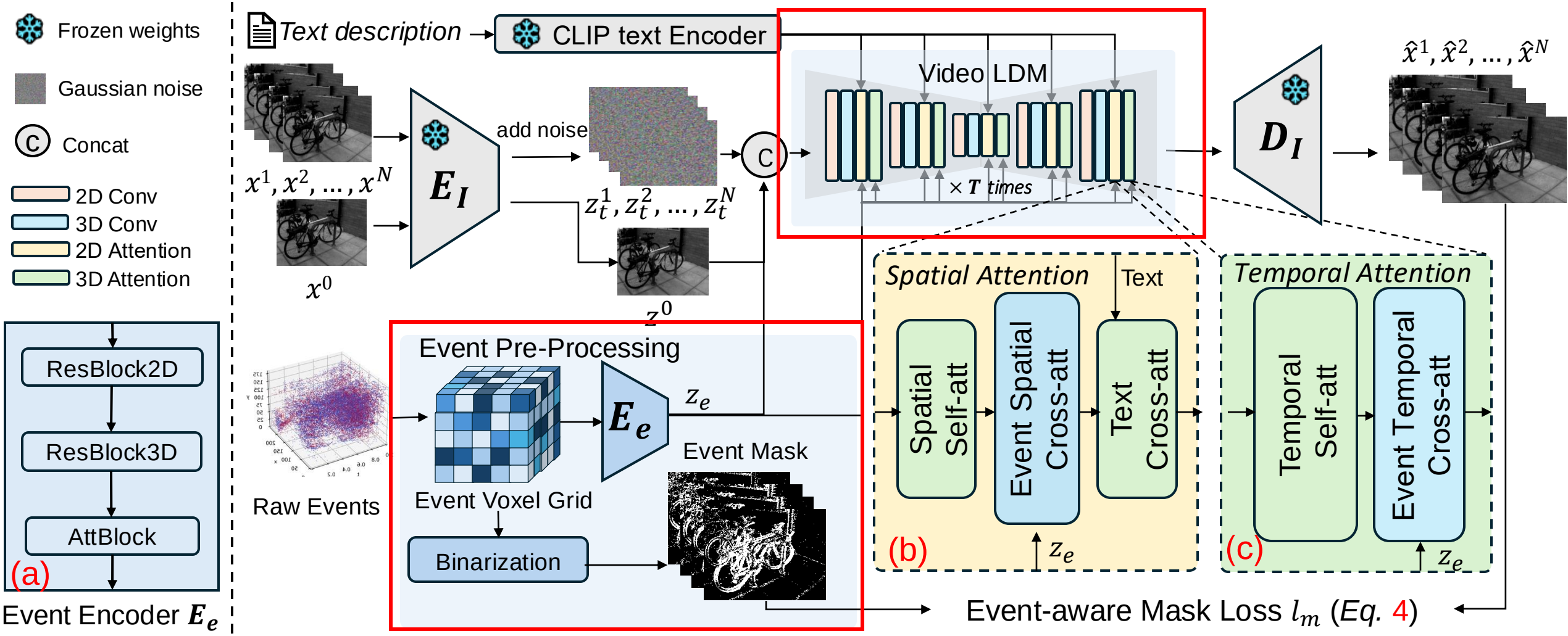
- Low temporal resolution
- Low dynamic range
- Absolute intensity
- A lot of algorithms

Overview

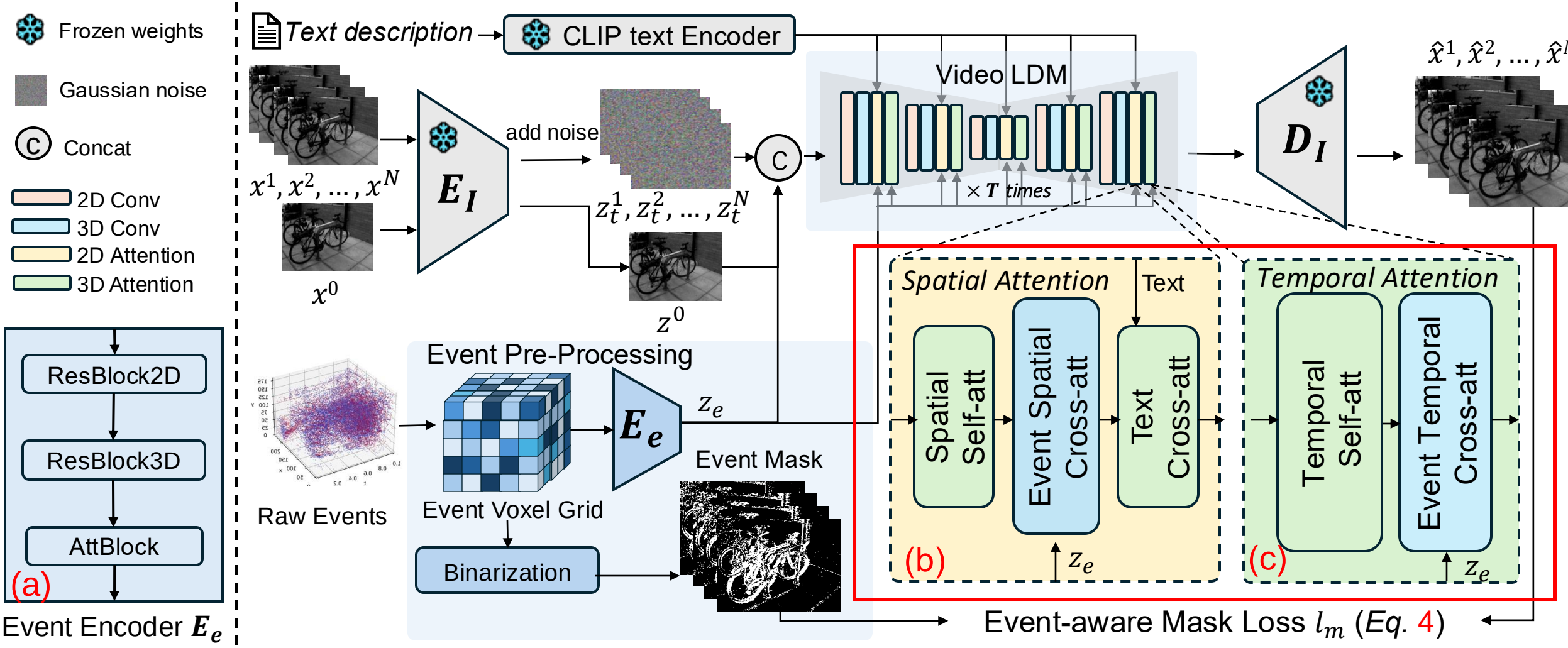
- **Key Idea:** incorporate text-conditional diffusion prior to facilitate a language-guided semantic-aware E2V reconstruction framework



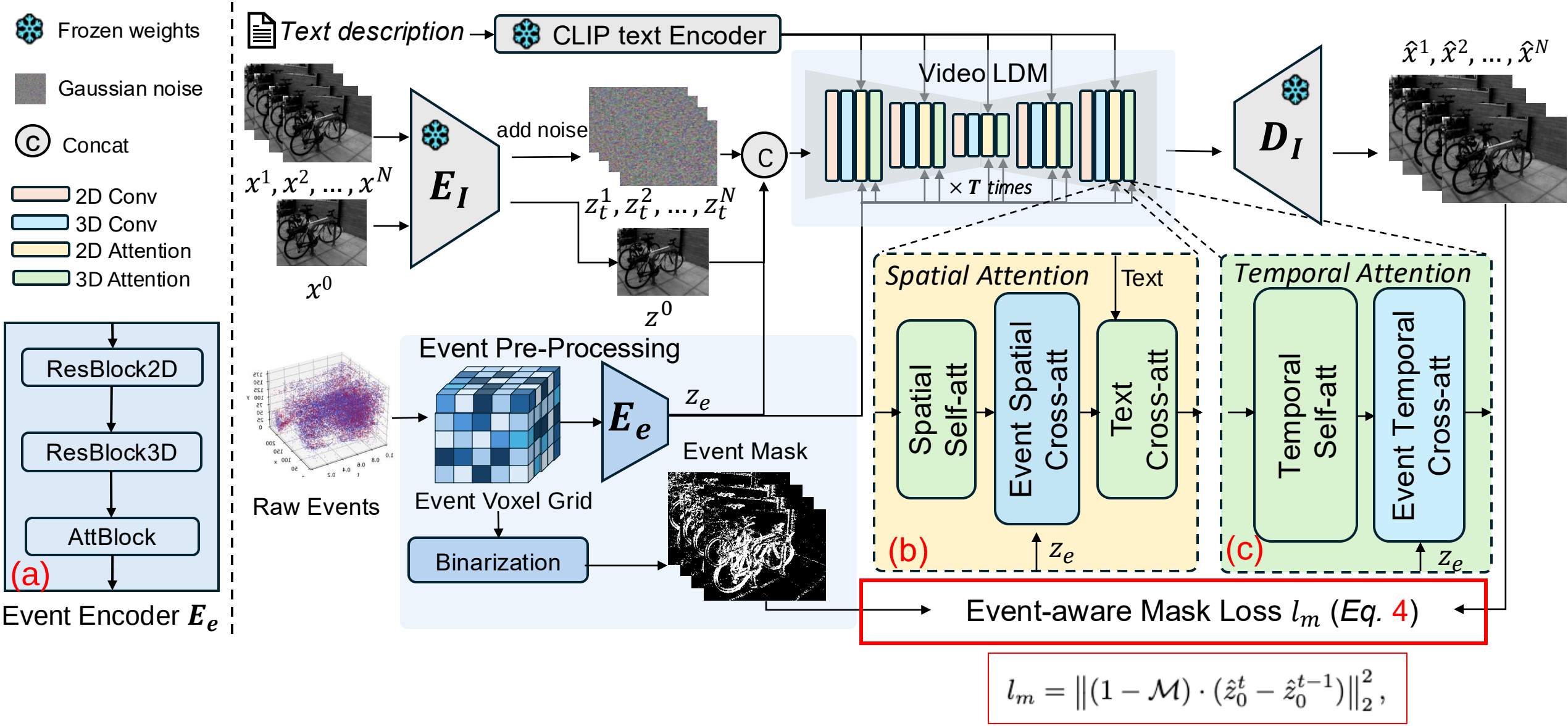
Method


Frozen weights
Gaussian noise
Concat
2D Conv
3D Conv
2D Attention
3D Attention
Text description
CLIP text Encoder
$x^1, x^2, \dots, x^N$
x^0
E_I
add noise
$z_t^1, z_t^2, \dots, z_t^N$
z^0
C
Video LDM
$\times T$ times
D_I
$\hat{x}^1, \hat{x}^2, \dots, \hat{x}^N$
Raw Events
Event Pre-Processing
Event Voxel Grid
E_e
z_e
Binarization
Event Mask
Spatial Attention
Text
Spatial Self-att
Event Spatial Cross-att
Text Cross-att
(b)
Temporal Attention
Temporal Self-att
Event Temporal Cross-att
(c)
ResBlock2D
ResBlock3D
AttBlock
(a)
Event Encoder E_e
Event-aware Mask Loss l_m (Eq. 4)

Method


Frozen weights
Gaussian noise
Concat
2D Conv
3D Conv
2D Attention
3D Attention
Text description
CLIP text Encoder
$x^1, x^2, \dots, x^N$
E_I
add noise
$z_t^1, z_t^2, \dots, z_t^N$
C
Video LDM
$\times T$ times
D_I
$\hat{x}^1, \hat{x}^2, \dots, \hat{x}^N$
x^0
z^0
Event Pre-Processing
Raw Events
Event Voxel Grid
E_e
z_e
Binarization
Event Mask
Spatial Attention
Text
Spatial Self-att
Event Spatial Cross-att
Text Cross-att
(b)
Temporal Attention
Temporal Self-att
Event Temporal Cross-att
(c)
ResBlock2D
ResBlock3D
AttBlock
(a)
Event Encoder E_e
Event-aware Mask Loss l_m (Eq. 4)

Method


Frozen weights
Gaussian noise
Concat
2D Conv
3D Conv
2D Attention
3D Attention
Text description
CLIP text Encoder
$x^1, x^2, \dots, x^N$
x^0
E_I
add noise
$z_t^1, z_t^2, \dots, z_t^N$
z^0
C
Video LDM
$\times T$ times
D_I
$\hat{x}^1, \hat{x}^2, \dots, \hat{x}^N$
Event Pre-Processing
Raw Events
Event Voxel Grid
E_e
z_e
Event Mask
Binarization
Spatial Attention
Text
Spatial Self-att
Event Spatial Cross-att
Text Cross-att
(b)
Temporal Attention
Temporal Self-att
Event Temporal Cross-att
(c)
ResBlock2D
ResBlock3D
AttBlock
(a)
Event Encoder E_e
Event-aware Mask Loss l_m (Eq. 4)
$l_m = \left\| (1 - \mathcal{M}) \cdot (\hat{z}_0^t - \hat{z}_0^{t-1}) \right\|_2^2,$

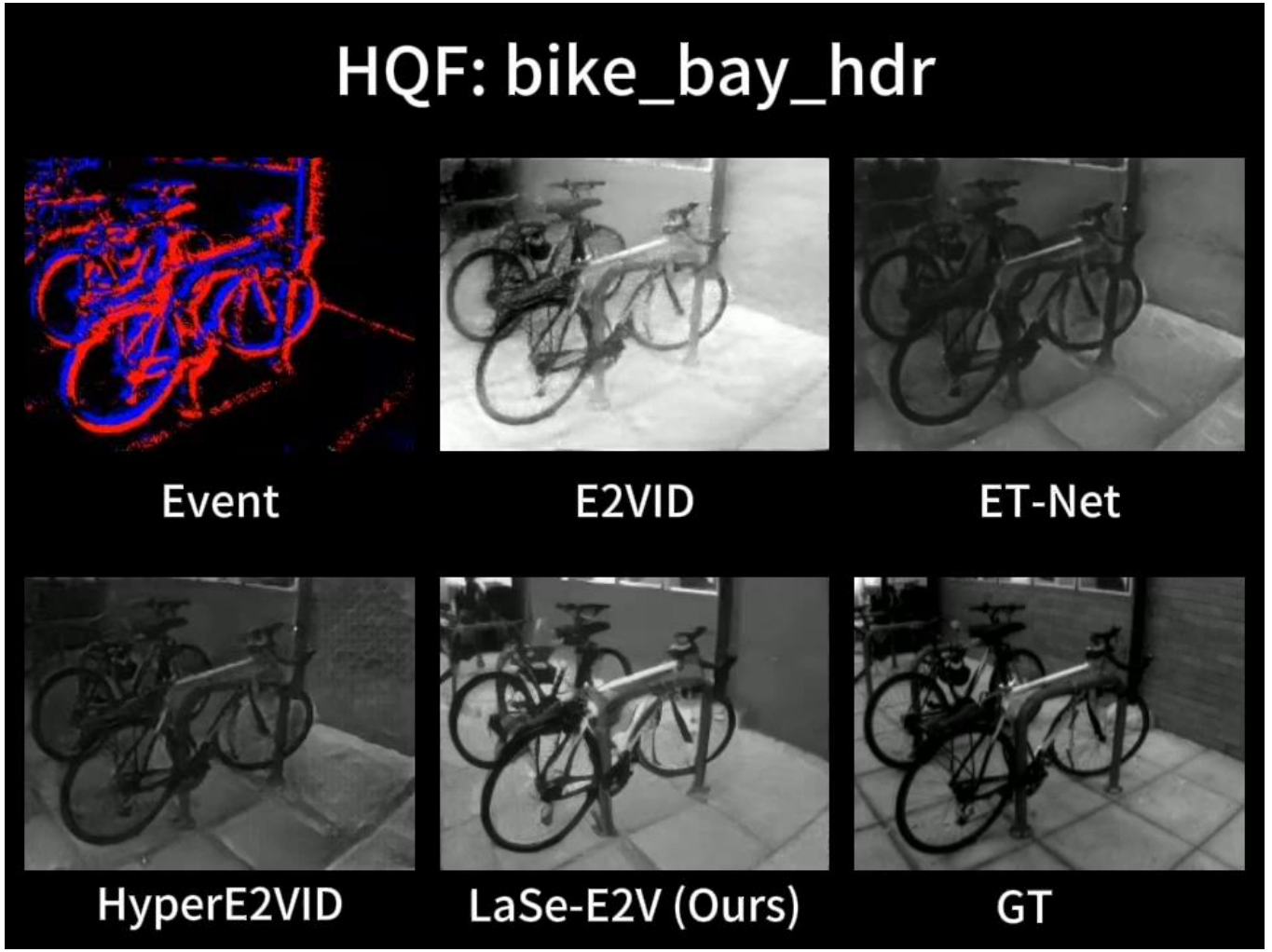
Experiments

- Our LaSe-E2V can reconstruct videos with higher quality compared to other SOTA methods

Datasets	Metrics	E2VID [50]	FireNet [54]	E2VID+ [57]	FireNet+ [57]	SPADE-E2VID [7]	SSL-E2VID [46]	ET-Net [64]	HyperE2VID [13]	**LaSe-E2V (Ours)**
ECD	MSE↓	0.212	0.131	0.070	0.063	0.091	0.046	0.047	0.033	**0.023**
	SSIM↑	0.424	0.502	0.560	0.555	0.517	0.364	0.617	0.655	-
	SSIM* ↑	0.450	0.459	0.503	0.452	0.461	0.415	0.552	0.576	**0.629**
	LPIPS↓	0.350	0.320	0.236	0.290	0.337	0.425	0.224	0.212	**0.194**
MVSEC	MSE↓	0.337	0.292	0.132	0.218	0.138	0.062	0.107	0.076	**0.055**
	SSIM↑	0.206	0.261	0.345	0.297	0.342	0.345	0.380	0.419	-
	SSIM* ↑	0.241	0.198	0.262	0.212	0.266	0.264	0.288	0.315	**0.342**
	LPIPS↓	0.705	0.700	0.514	0.570	0.589	0.593	0.489	0.476	**0.461**
HQF	MSE↓	0.127	0.094	0.036	0.040	0.077	0.126	0.032	**0.031**	0.034
	SSIM↑	0.540	0.533	0.643	0.614	0.521	0.295	0.658	0.658	-
	SSIM* ↑	0.462	0.422	0.536	0.474	0.405	0.407	0.534	0.531	**0.548**
	LPIPS↓	0.382	0.441	**0.252**	0.314	0.502	0.498	0.260	0.257	0.254

Experiments

- Our LaSe-E2V can reconstruct videos with higher quality compared to other SOTA methods



Take-away Message

We novelly explore E2V reconstruction from a language-guided perspective, utilizing the text-conditioned diffusion model to effectively address the semantic ambiguities inherent in event data.