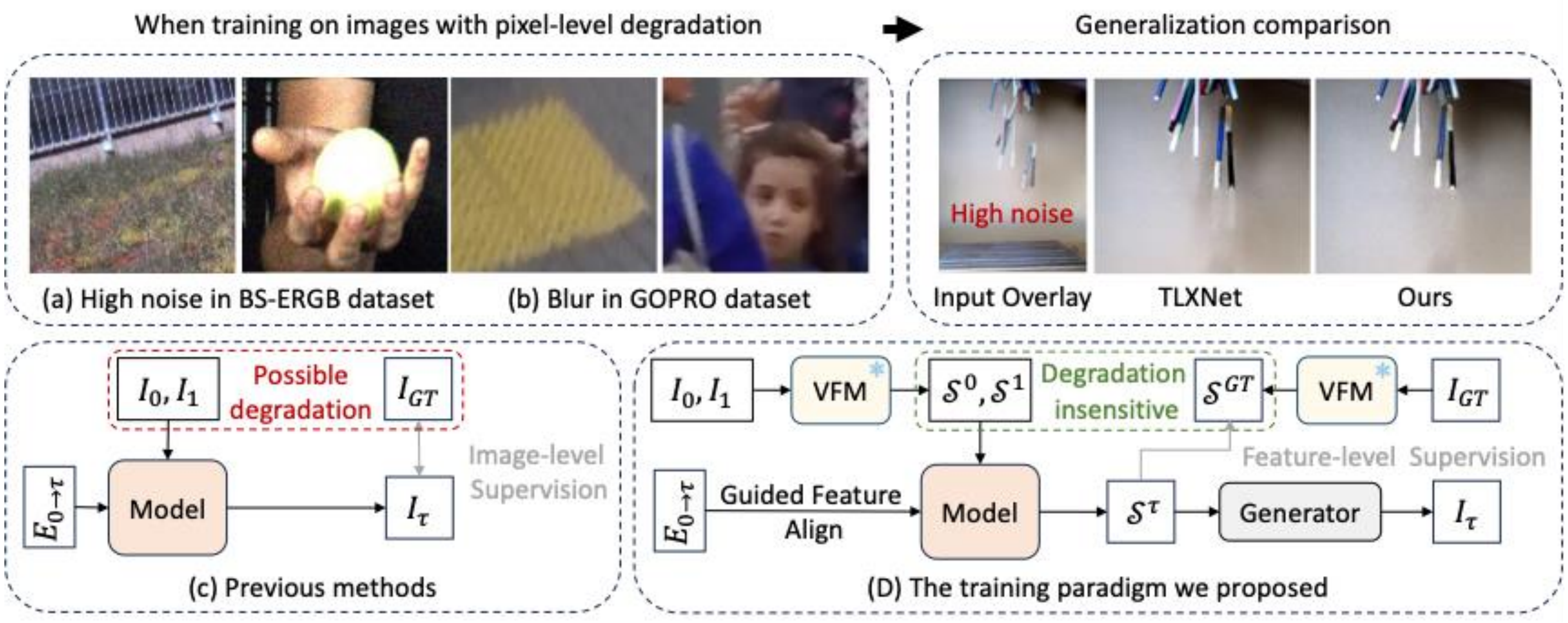
EPA: Boosting Event-based Video Frame Interpolation with Perceptually Aligned Learning

Problems:

- ➢ Existing Event-based Video Frame Interpolation (E-VFI) methods are severely limited by **motion blur and image degradation** commonly found in both the input keyframes and the ground truth supervision signals.
- ➢ Traditional approaches rely on **pixel-level supervision**, which forces the model to learn and amplify these visual artifacts, leading to perceptually unrealistic results and **poor generalization** to diverse, real-world scenes.

The Core Challenge:

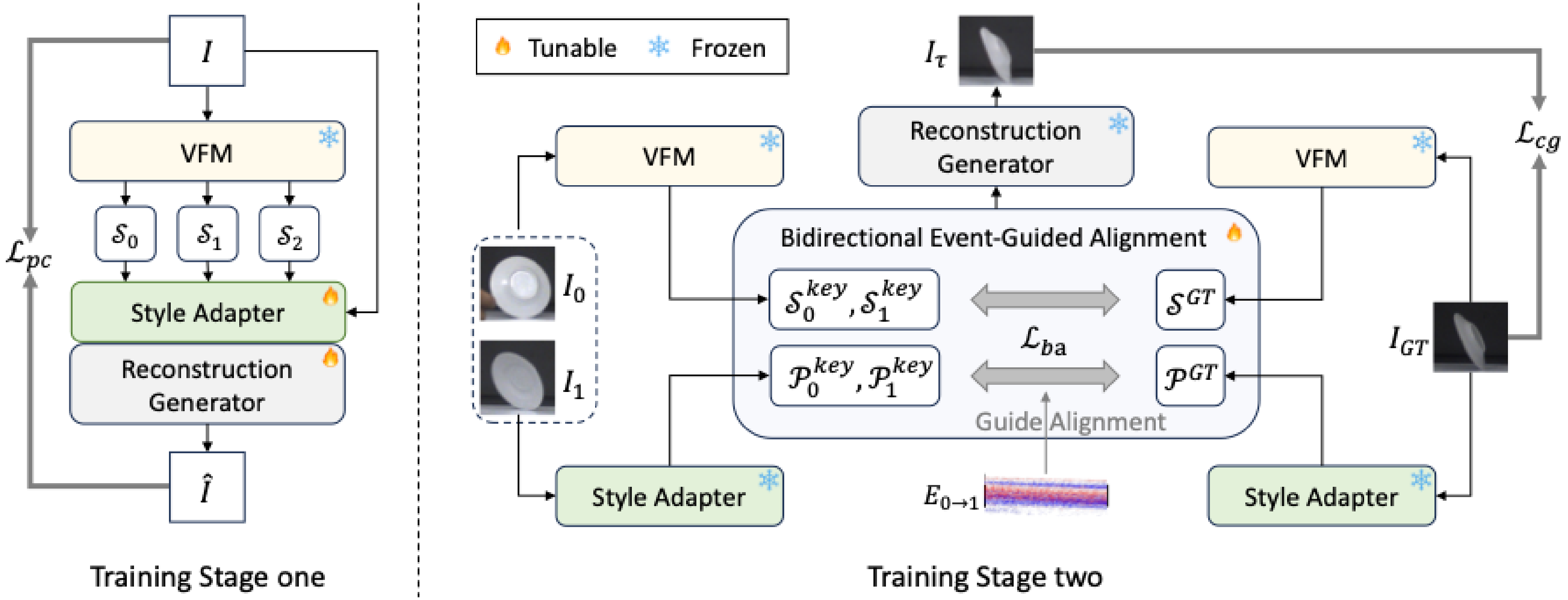
➢How can we effectively learn from degraded data without propagating these flaws into the final interpolated frames, thereby achieving higher perceptual quality?



We propose a novel framework, **EPA**, which shifts the learning paradigm from the unstable pixel space to a **degradation-insensitive semantic-perceptual feature space**. Our method consists of two core stages:

- **Robust Feature Extraction & Reconstruction**: We leverage a powerful Vision Foundation Model (VFM) to extract robust semantic features that are insensitive to degradation. A custom Style Adapter complements these features with low-level details, ensuring they can be reconstructed into high-fidelity images.

- **Bidirectional Event-Guided Alignment (BEGA)**: We introduce a novel BEGA module that uses the high temporal resolution of event streams as precise motion guidance to align and fuse perceptual features from the keyframes at the feature level.



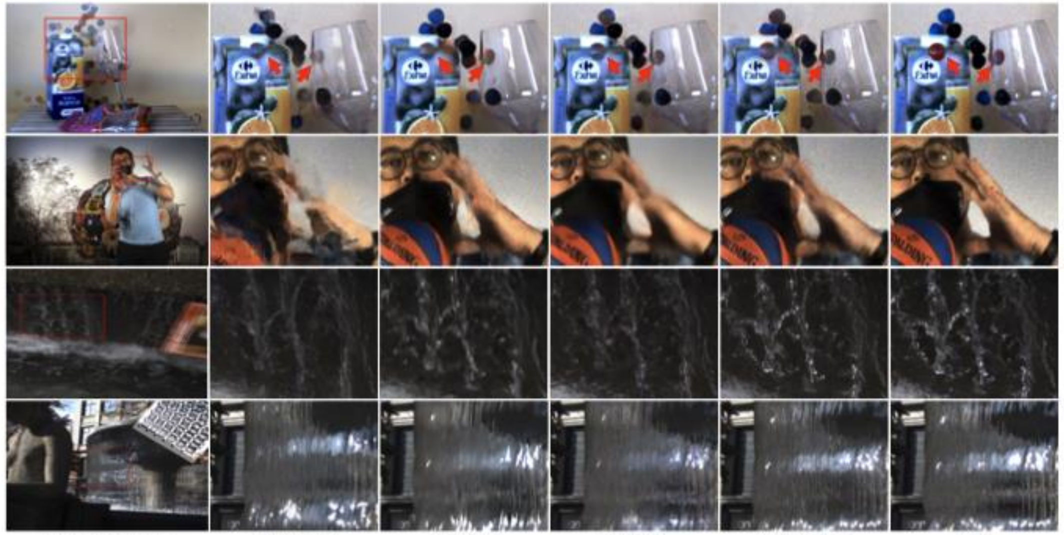
Experiment （Synthetic Dataset）

Methods	Vimeo90k			GOPRO					
	1 skip			7 skip			15 skip		
	LPIPS↓	FloLPIPS↓	DISTS↓	LPIPS↓	FloLIPIS↓	DISTS↓	LPIPS↓	FloLIPIS↓	DISTS↓
RIFE	0.021	0.062	0.048	0.029	0.100	0.060	0.051	0.168	0.082
UPR-Net	0.015	0.039	0.037	0.024	0.077	0.052	0.042	0.140	0.067
Timelens	0.022	0.040	0.052	0.009	0.033	0.031	0.012	0.047	0.036
CBMNet	0.012	0.021	0.039	0.012	0.050	0.046	0.013	0.058	0.050
TLXNet	0.089	0.142	0.116	0.028	0.052	0.049	0.031	0.063	0.053
EPA (ours)	**0.007**	**0.012**	**0.036**	**0.006**	**0.021**	**0.019**	**0.008**	**0.031**	**0.023**

Input Overlay Timelens CBMNet TLXNet EPA (ours) GT

Experiment （Real Dataset）

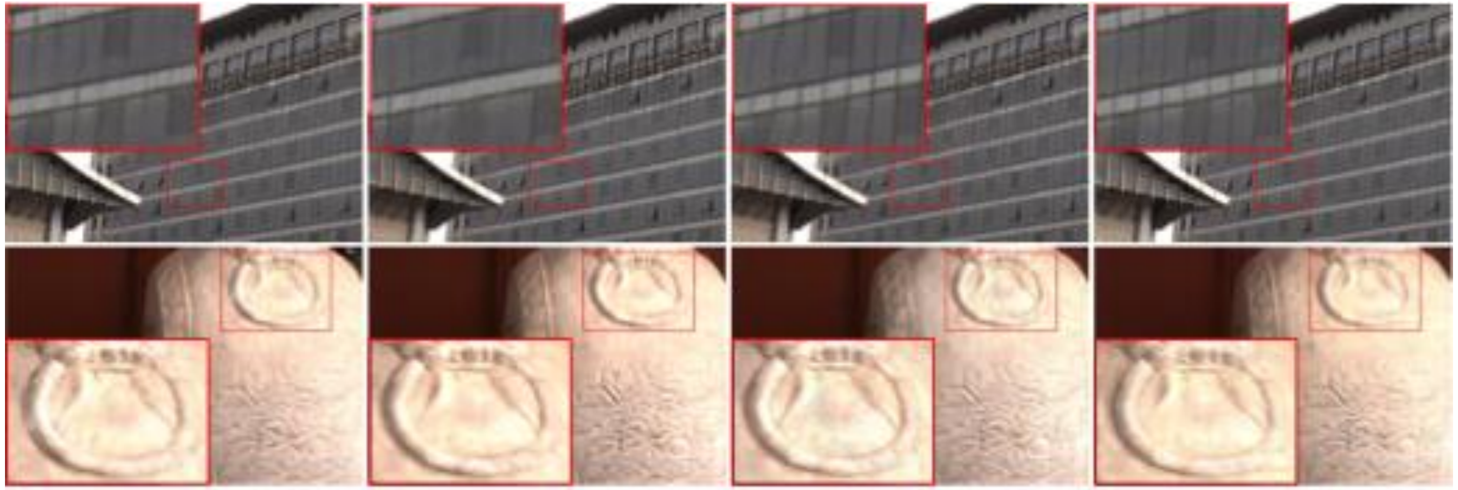
Method	HS-ERGB									
	5 skip					7 skip				
	PSNR↑	SSIM↑	LPIPS↓	FloLPIPS↓	DISTS↓	PSNR↑	SSIM↑	LPIPS↓	FloLPIPS↓	DISTS↓
RIFE	32.624	0.857	0.032	0.192	0.083	31.150	0.836	0.037	0.212	0.092
UPR-Net	32.235	0.857	0.075	0.170	0.075	30.689	0.834	0.085	0.188	0.081
Timelens	32.760	0.861	0.046	0.112	0.059	31.871	0.851	0.053	0.126	0.065
CBMNet	32.206	0.842	0.098	0.212	0.108	31.876	0.837	0.101	0.218	0.110
TLXNet	-	-	-	-	-	31.578	0.827	0.046	0.105	0.054
EPA (ours)	**33.842**	**0.872**	**0.014**	**0.057**	**0.045**	**33.402**	**0.867**	**0.015**	**0.062**	**0.048**
	BS-ERGB									
	1 skip					3 skip				
RIFE	25.616	0.765	0.098	0.310	0.067	23.435	0.728	0.114	0.357	0.073
UPR-Net	25.621	0.779	0.104	0.308	0.083	23.081	0.736	0.108	0.335	0.082
Timelens	27.164	0.783	0.052	0.153	0.065	25.855	0.765	0.064	0.202	0.075
CBMNet	29.257	**0.814**	0.060	0.203	0.087	28.446	**0.807**	0.063	0.221	0.090
TLXNet	**29.298**	0.813	0.047	0.088	0.052	**28.720**	**0.807**	0.046	0.090	0.058
EPA (ours)	27.943	0.791	**0.024**	**0.068**	**0.051**	27.221	0.782	**0.028**	**0.082**	**0.057**



Input Overlay Timelens CBMNet TLXNet EPA (ours) GT

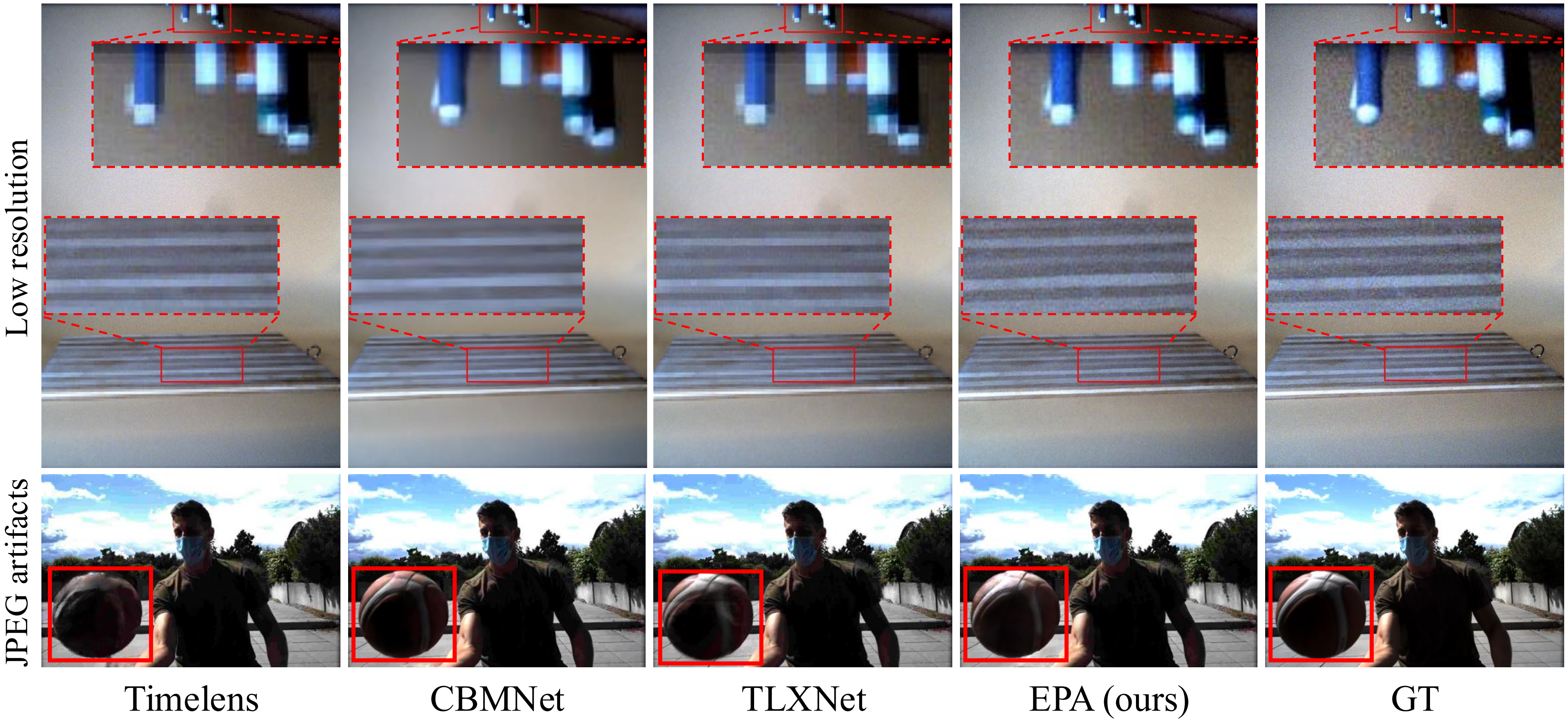
Experiment （Real Dataset）

Method	Building		Sculpture	
	PSNR↑	LPIPS↓	PSNR↑	LPIPS↓
Timelens	**31.78**	0.037	**36.52**	0.020
CBMNet	31.42	0.054	34.79	0.052
TLXNet	29.07	0.035	33.85	0.026
EPA (ours)	31.43	**0.015**	35.15	**0.011**



Timelens CBMNet TLXNet EPA (ours)

Experiment (Robustness to Other Degradation Types)



Conclusion

- We introduced EPA, a novel E-VFI framework that tackles the critical issue of input degradation by learning in a robust perceptual feature space.

- Our proposed BEGA module effectively leverages event data to guide feature alignment, achieving more robust and higher-quality interpolation.