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

Yuhan Liu[1,2], **Linghui Fu**[2], **Zhen Yang**[2], **Hao Chen**[3], **Youfu Li**[4,5], **Yongjian Deng**[2,5*]
[1]Key Laboratory of Multimedia Trusted Perception and Efficient Computing, Ministry of Education of China, Xiamen University, 361005, P.R. China
[2]College of Computer Science, Beijing University of Technology
[3]Key Lab of Computer Network and Information Integration, Southeast University
[4]Department of Mechanical Engineering, City University of Hong Kong, Kowloon, Hong Kong SAR
[5]CityU Shenzhen Research Institute, Shenzhen, P.R. China
[1]yuhanliu@stu.xmu.edu.cn, {fulinghui@emails., yangzhen@, yjdeng@}bjut.edu.cn, haochen303@seu.edu.cn, meyfli@cityu.edu.hk

Abstract

Event cameras, with their capacity to provide high temporal resolution information between frames, are increasingly utilized for video frame interpolation (VFI) in challenging scenarios characterized by high-speed motion and significant occlusion. However, prevalent issues of blur and distortion within the keyframes and ground truth data used for training and inference in these demanding conditions are frequently overlooked. This oversight impedes the perceptual realism and multi-scene generalization capabilities of existing event-based VFI (E-VFI) methods when generating interpolated frames. Motivated by the observation that semantic-perceptual discrepancies between degraded and pristine images are considerably smaller than their image-level differences, we introduce EPA. This novel E-VFI framework diverges from approaches reliant on direct image-level supervision by constructing multilevel, degradation-insensitive semantic perceptual supervisory signals to enhance the perceptual realism and multi-scene generalization of the model's predictions. Specifically, EPA operates in two phases: it first employs a DINO-based perceptual extractor, a customized style adapter, and a reconstruction generator to derive multi-layered, degradation-insensitive semantic-perceptual features ($\mathcal{S}$). Second, a novel Bidirectional Event-Guided Alignment (BEGA) module utilizes deformable convolutions to align perceptual features from keyframes to ground truth with inter-frame temporal guidance extracted from event signals. By decoupling the learning process from direct image-level supervision, EPA enhances model robustness against degraded keyframes and unreliable ground truth information. Extensive experiments demonstrate that this approach yields interpolated frames more consistent with human perceptual preferences. Codes are available at https://github.com/yuhan0802/EPA.

1 Introduction

Recent years, Event-based Video Frame Interpolation (E-VFI) has attracted significant attention due to its outstanding performance under extreme conditions. Leveraging the high temporal resolution of event cameras [23, 48], E-VFI demonstrates clear advantages over frame-based VFI methods in challenging scenarios involving fast motion, severe occlusion, and non-rigid object deformation. It enables more accurate inter-frame motion estimation [40, 41, 21, 31], or even directly provides temporal priors at the interpolation location [28, 8], thereby facilitating the generation of higher-quality intermediate frames. However, previous works often overlook a critical issue exposed in such challenging environments: the presence of motion blur and image degradation caused by the inherent

* Corresponding author

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

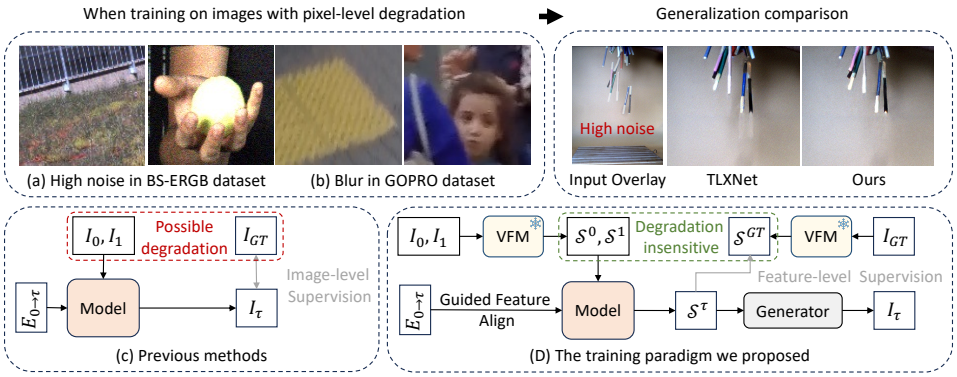


Figure 1: Visualization of image degradation in public datasets, along with a comparative illustration of our training paradigm versus prior methods, where VFM denotes the visual foundation model with frozen weights..

limitations of conventional imaging sensors, as shown in Fig. 1 (a)&(b). These methods typically assume that input frames are high quality and rely on image-level supervision during training (Fig. 1 (c)). *In fact, when images are degraded to a certain extent, defect messages conveyed by keyframes or Ground Truth data (GT) would hamper the capability of model to fit the distribution of real world, ultimately hindering the perceptual realism and generalization ability of E-VFI models across diverse scenes.*

Taking inspiration from [18] that semantic-perceptual features ($\mathcal{S}$) of images suffering less from image degradation, in this work, we propose a novel learning framework, EPA, to address above issues of **E**-VFI tasks through **P**ercepual-based feature **A**lignment. Unlike prior methods that rely on image-level supervision (Fig. 1(c)), our approach adopts feature-level training to mitigate the perceptual degradation caused by overfitting to low-quality inputs, as illustrated in Fig. 1(d). There are two stages contained in the EPA. The first stage aims to utilize the visual foundation model (M_f) to extract semantic-perceptual features ($\mathcal{S}^{GT}$) from GT and use the degradation-insensitive advantage of $\mathcal{S}^{GT}$ for model optimization. Here, one more thing has to be guaranteed that a carefully designed decoder can reconstruct the $\mathcal{S}^{GT}$ to its original image format. To this end, we introduce a reconstruction generator (G_r) equiped with a customized style adapter (A_s) for image reconstruction, where some low-level cues or non-saliency regions neglected by M_f can be supplemented via A_s. The second stage is imposed for aligning the distribution gap bettween semantic-perceptual features from keyframes and Ground Truth data. We achieve this by proposing a Bidirectional Event-Guided Alignment (BEGA) module, which performs alignment process under the guidance by inter-frame temporal messages of event data in a hierachical manner. Finally, the aligned perceptual feature is fed into the generator G_r for interpolating estimation.

Our contributions can be summarized as follows: (1) We propose EPA, a novel E-VFI framework that learns in the semantic-perceptual feature space using degradation-insensitive supervision, enabling more perceptually realistic frame synthesis. (2) We introduce a Style Adapter to enhance low-level details and non-salient regions overlooked by vision foundation models, improving reconstruction quality. (3) We design a Bidirectional Event-Guided Alignment (BEGA) module, which utilizes fine-grained motion cues from events to guide hierarchical feature alignment. (4) Extensive experiments on synthetic and real-world datasets demonstrate that our method consistently outperforms prior approaches in perceptual quality and generalization.

2 Related Work

2.1 Video Frame Interpolation

Motion-based frame interpolation methods dominate traditional VFI, typically relying on optical flow to warp keyframes. Enhancements such as bidirectional flow [34, 17], coarse-to-fine refinement [37], correlation-based updates [24, 35], and motion-synthesis coupling [16, 12] have improved performance, yet these methods remain vulnerable to severe occlusions. To address this, synthesis-based approaches [38, 20] avoid warping errors but demand more temporal information, increasing complexity. Both paradigms struggle under large motions due to substantial inter-frame gaps.

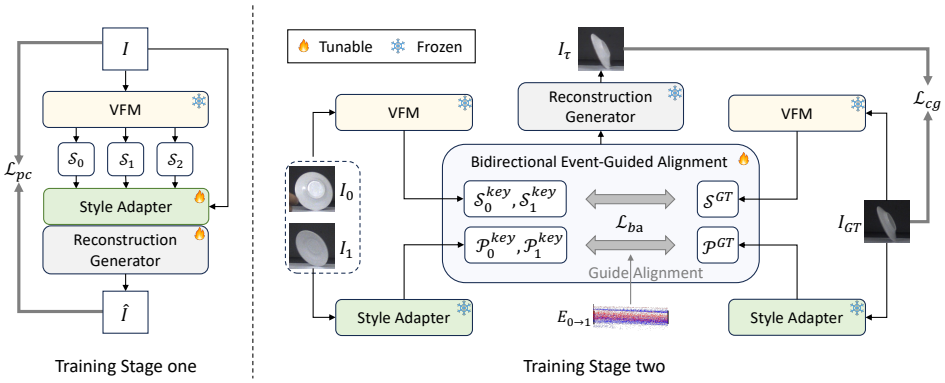


Figure 2: The pipline of the proposed EPA. Modules with the same name share weights.

Recently, generative methods [15, 45, 4] have emerged to produce perceptually convincing frames, albeit often sacrificing pixel-level accuracy due to their stochastic nature.

2.2 Event-based Video Frame Interpolation

Event cameras, with microsecond-level temporal resolution, enable dense inter-frame reconstruction and have driven rapid progress in event-based video frame interpolation (E-VFI). Most methods exploit events to improve optical flow estimation [40, 46, 25, 49, 51, 43, 13, 29, 27]. Advances such as spline-based modeling [41], unsupervised consistency [10], recurrent architectures [39], multi-cost volumes [21], and piecewise flow fitting [31] have improved motion estimation, yet non-rigid motion and occlusion remain challenging. To mitigate these issues, synthesis-based approaches [8, 28] have emerged, avoiding flow-related errors. However, the issue of real semantic perception deviation caused by frame degradation remains largely unaddressed. Although some methods [39] jointly optimize denoising and interpolation, this multi-task formulation tends to compromise task-specific performance and introduces unnecessary complexity.

3 Method

This section aims to describe the E-VFI framework, EPA, from problem formulation to detailed architectural designs. Given two input keyframes I_0 and I_1 along with the events $\mathcal{E}_{0\to1}$ between them, our objective is to generate the intermediate frame I_τ at a specific timestamp $\tau \in [0, 1]$. In detail, as shown in Fig. 2, EPA is composed of two stages:

The first stage contains three components, *i.e.,* a Vision Foundation Model M_f, a Style Adapter A_s, and a reconstruction generator G_r, where M_f is for extracting semantic-perceptual features $\mathcal{S}$ from images and A_s & G_r are introduced for reconstruct high-fedility images from corresponding $\mathcal{S}$. In the second stage, $\mathcal{E}_{0\to1}$ is split into two subsets $E_{0\to\tau}$ and $E_{1\to\tau}$, which are then converted into voxel grids $V_{0\to\tau}$ and $V_{1\to\tau}$. These event voxel representations, together with the semantic features $\{\mathcal{S}_i^{key} | i \in \{0, 1\}\}$ extracted from I_0 and I_1 are fed into the Bidirectional Event-Guided Alignment (BEGA) module, which performs hierarchical alignment under the temporal guidance of event data to fit the semantics of interpolated frame. Finally, G_r synthesizes the interpolated frame I_τ directly from the fitted semantic perceptual features. In the remainder of this section, we first motivate our choice of feature-level supervision in Sec. 3.1, then describe our image reconstruction strategy in Sec. 3.2, detail the design of feature alignment in Sec. 3.3, and conclude with the architectural specifics of each component in Sec. 3.4.

3.1 The Motivation of Introducing Feature-Level Supervision

State-of-the-art E-VFI methods [40, 41, 31, 21] typically estimate optical flow from events and warp keyframes to synthesize intermediate frames. However, optical flow mainly captures pixel-level motion, normally ignores object semantics and scene structure, making these methods heavily reliant on high-quality keyframes and GT. When keyframes suffer degradation, their discrepancy from human perception of the real world grows, and training on degraded images propagates these defects, undermining perceptual fidelity and generalization across diverse scenes. Inspired by [18], we find that the perceptual discrepancy between degraded and clean images is reduced in the feature space

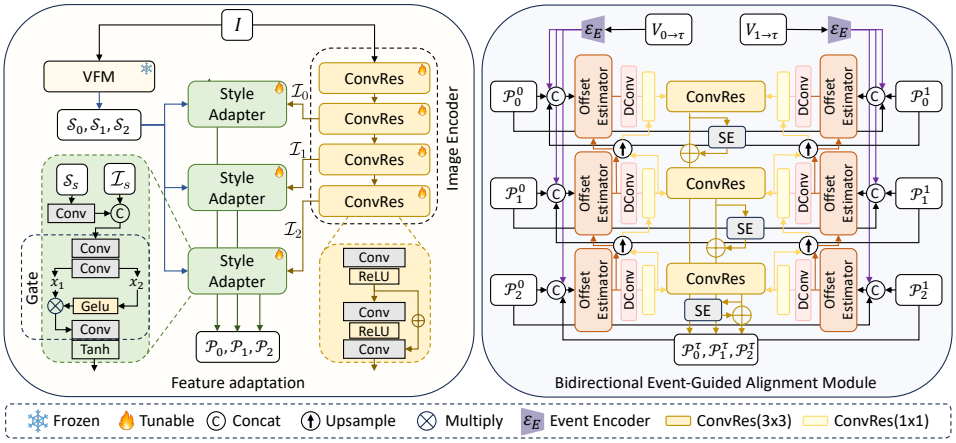


Figure 3: An overview of the proposed modules. Upsample refers to up-sampling using bilinear interpolation.

compared to the image space (Fig. 8). This motivates us proposing the feature-level supervision strategy, which reduces reliance on the quaility of input frames.

While prior works [40, 8, 28] attempt to achieve this goal by applying image-level perceptual losses [19] on the predictions, their interpolation processes still operate in the pixel domain. As a result, the models lack a true understanding of image semantics and tend to overfit to low-level details, which limits their robustness under degraded conditions. In contrast, our method enforces alignment with the ground truth at the feature level, promoting high-level semantic consistency and enhancing resilience to variations in input quality. In specific, we incorporate a vision foundation model M_f that is designed to comprehensively capture object-level semantics in visual scenes [1], which has been shown to be degradation-insensitive [26]. From M_f, we extract semantic-perceptual features $\mathcal{S}$, which play a pivotal role in guiding the learning process.

3.2 Image Reconstruction from Semantic-Perceptual Features

After obtaining the semantic representation $\mathcal{S}$ from M_f, the network that is able to reconstruct a high-quality image from this high-level abstraction is required. To this end, we employ an off-the-shelf reconstruction generator G_r that is compatible to multi-level feature input for restoring the image from $\mathcal{S}$.

Reconstruction with Style Adapter. From our observation, while feature representations from M_f has already offer strong semantic expressiveness and generalization, they are typically optimized for high-level tasks such as segmentation and detection. Consequently, these features tend to overlook low-level cues and non-salient regions. As illustrated in the feature visualizations in Fig. 8, attention to areas like the sky is significantly reduced, which adversely affects reconstruction quality. To compensate for this deficiency, we introduce a customized style adapter A_s, which complements the high-level semantic features with essential low-level details. As shown in Fig. 3, we first employ a lightweight image encoder to extract low-level features $\mathcal{I}_s$, which are subsequently passed to A_s to enrich the semantic representation. Specifically, the proposed style adapter A_s consists of convolutional layers and a gating mechanism, which performs a weighted fusion of $\mathcal{S}_s$ and $\mathcal{I}_s$ to produce the final adapted features $\mathcal{P}_s$, where $s \in 0, 1, 2$. These adapted features enable the G_r to recover high-fidelity images with both semantic consistency and fine-grained visual quality through the adapted feature $\mathcal{P}$.

3.3 Feature-Level Alignment with Event-Based Assistance

In the E-VFI setting, the objective is to synthesize the intermediate frame I_τ from the keyframes I_0 and I_1 along with the inter-frame events $\mathcal{E}_{0\to1}$. Unlike methods that operate at the image-level, EPA interpolates in the semantic-perceptual feature space by estimating the semantic representation of the intermediate frame and generating the final frame using a pretrained reconstruction generator G_r.

Bidirectional Event-Guided Alignment (BEGA) module. In order to obtain the semantic-perceptual features of the interpolations, we introduce the Bidirectional Event-Guided Alignment (BEGA) module. This module semantically aligns the two keyframes by leveraging the fine-grained temporal cues embedded in the event stream. Intuitively, BEGA should be directly applied to align

$\mathcal{S}^\tau$ from $\mathcal{S}_i^{key}$. However, to compensate for the fine-grained low-level semantics (*e.g.,* color and texture) that are missing in $\mathcal{S}^\tau$ during reconstruction, alignment of the $\mathcal{I}^\tau$ is also required. Our empirical findings reveal that this process is non-trivial due to the inherent limitations of event data: events are insensitive to smooth regions and lack color information. This leads to unstable training and sub-optimal reconstruction quality. Motivated by this observation, we instead apply BEGA to jointly align both semantic and low-level features, *i.e.,* direct alignment $\mathcal{P}^\tau$. This approach preserves richer low-level cues, resulting in more stable training and improved synthesis performance.

As illustrated in Fig.3, BEGA takes the event voxel grids $V_{0\to\tau}$ and $V_{1\to\tau}$ as inputs and processes them through a shared-weight event encoder $\mathcal{E}_E$ to extract hierarchical features E_s^0 and E_s^1. To retain fine motion cues inherent in the event streams, $\mathcal{E}_E$ is deliberately kept lightweight, comprising only three convolutional layers to emphasize low-level feature representation. Subsequently, at multiple spatial scales $s \in \{0, 1, 2\}$, we synthesize the intermediate frame features via one-way warping using deformable convolutions [2], followed by occlusion-aware refinement with a lightweight SE layer [11]. This design enables robust bidirectional alignment and effective fusion of intermediate semantic features, as formulated in Eq.1.

$$\begin{aligned} \epsilon_s^o &= \mathcal{F}_{DC}(\mathcal{F}_{OE}(\mathcal{C}(\mathcal{P}_s^o, E_s^o))), o \in \{0, 1\} \\ \epsilon_s &= \mathcal{C}(\epsilon_s^0, \epsilon_s^1), \mathcal{P}_s^\tau = \epsilon_s + \mathcal{F}_{SE}(\epsilon_s) \end{aligned} \tag{1}$$

where $\mathcal{F}_{DC}$ denotes the deformable convolution, $\mathcal{F}_{OE}$ represents the offset estimator, which computes the offset and mask required by $\mathcal{F}_{DC}$, following a structure similar to that of [42]. $\mathcal{F}_{SE}$ refers to the squeeze-and-excitation operation.

3.4 Network Structure

Specifically, we use a DINO-pretrained ResNet backbone as the vision foundation model (M_f) to extract perceptual features, and adopt a normalized flow-based generator [45] as the reconstruction network (G_r). During the process, keyframes are fed into M_f to extract multi-scale semantic features $\mathcal{S}_s^{key}$, which are enhanced by three independent style adapters A_s using low-level features $\mathcal{I}_s^{key}$ from an image encoder, producing adapted features $\mathcal{P}_s$ for each scale $s \in \{0, 1, 2\}$. These are then passed to G_r to synthesize the intermediate frame I_τ.

3.5 Loss

As shown in Fig. 2, in the first stage, we utilize $\mathcal{L}_{pc}$ to ensure the reconstruction consistency, as depicted in Eq.2.

$$\mathcal{L}_{pc} = \mathcal{L}_{Lpips}(I, \hat{I}) + \mathcal{L}_{Lap}(I, \hat{I}) + \mathcal{L}_{nll}, \tag{2}$$

where the $\hat{I}$ denoted the the reconstructed image. The $\mathcal{L}_{Lpips}$ is perceptual loss introduced in [49], which excels at measuring the structural similarity between images. The $\mathcal{L}_{Lap}$ is a variant of the $L1$ loss proposed in [14], where the $L1$ loss is computed on the Laplacian pyramid representations of two images. $\mathcal{L}_{nll}$ is the negative log-likelihood loss, used in the optimization of normalized flow-based generators by [45].

In the second stage, we utilize a hybrid loss to emphasize feature consistency, as depicted in Eq. (3).

$$\begin{aligned} \mathcal{L}_{bf} &= \sum_{s=0}^{2} \mathcal{L}_2(\mathcal{P}_s, \mathcal{P}_s^{GT}) + \lambda_2 * (\mathcal{L}_2(\epsilon_0^s, \mathcal{P}_s^{GT}) + \mathcal{L}_2(\epsilon_1^s, \mathcal{P}_s^{GT})) \\ \mathcal{L}_{cg} &= \mathcal{L}_{Lpips}(I_\tau, I_{GT}) + \lambda_{lap} * \mathcal{L}_{Lap}(I_\tau, I_{GT}) \end{aligned} \tag{3}$$

where λ_{lap} and λ_2 are set as 0.2, 0.1 respectively. The function $\mathcal{L}_2$ denotes the $L2$ loss that used to align two features.

Table 1: Performance comparison on synthetic datasets. The best results are marked in **Bold** while the second ones are marked with underlines. We reconstructed all skipped frames for GOPRO.

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**

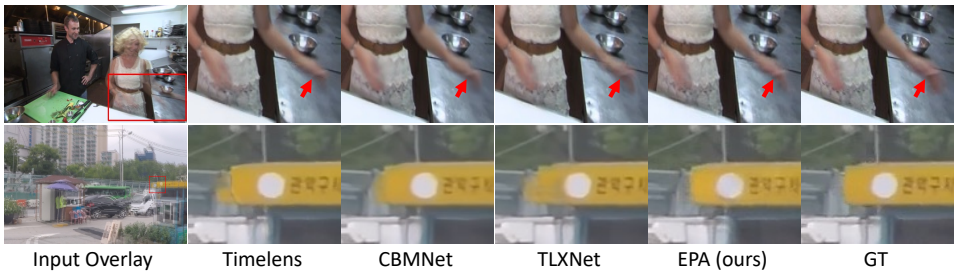


Figure 4: Visual comparison among different methods on synthesis datasets.

4 Experiments

4.1 Setup

Training Settings. For the proposed EPA, we first optimize the style adapter and reconstruction generator modules, after which their weights are frozen to train the bidirectional feature alginement module. In the first training stage, our method is optimized using AdamW [30] for 100 epochs within the PyTorch [36]. The initial learning rate is set to 1×10^{-4} and is gradually decreased to 1×10^{-6} via cosine annealing. The batch size is set to 40 for each training step. In the second stage, the bidirectional adaptation module is trained for 40 epochs under the same configuration. The entire model is trained on GOPRO [33] following [31], where synthetic event data is generated using the v2e simulator [9]. Note that the training of the generator is not restricted to any specific dataset. To ensure reliable generation quality, this work applies NIQE [32] to filter out degraded images, retaining only high-quality samples. For data augmentation, both input frames and their corresponding event voxel grids are cropped to 256×256 and randomly augmented with rotation and flipping. Our normalized flow generation module follows the setup used in [45].

State-of-the-Art Methods. We compare our approach against several state-of-the-art VFI&E-VFI methods, including RIFE (ECCV'2022) [14], UPR-Net (CVPR'2023) [16], TimeLens (CVPR'2021) [40], CBMNet (CVPR'2023) [22], and TLXNet (ECCV'2024) [31], using their publicly available implementations. Additionally, to ensure a fair comparison, we re-train all competing methods under the same configuration. As TLXNet does not support 6-skip training, we train this method using only GOPRO. For works that have not released official code but may demonstrate promising performance [46, 28], we attempted to reproduce them. However, due to unsatisfactory results, we do not include them in our comparisons.

Datasets. We evaluate our method on both synthetic and real-world event datasets. The synthetic benchmarks include Vimeo90k-Triplet [47] and GOPRO [33]. For real-world evaluation, we use HS-ERGB [40], comprising 15 scenes and various motion types; BS-ERGB [41], characterized by noise and complex non-rigid deformations; and EventAid-F [7], containing various motion scenarios. *Note that due to space constraints, full evaluation results on EventAid-F are provided in the supplementary material.*

Metrics. Our goal is to generate images that align with human perceptual preferences, which are not fully captured by traditional image quality metrics as evidenced in Fig. 6. Thus, We primarily adopt the following metrics for performance evaluation: LPIPS [50], DISTS [6], and FloLPIPS [3], as they

Table 2: Performance comparison on real datasets. The best results are marked in **Bold** while the second ones are marked with underlines.

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**

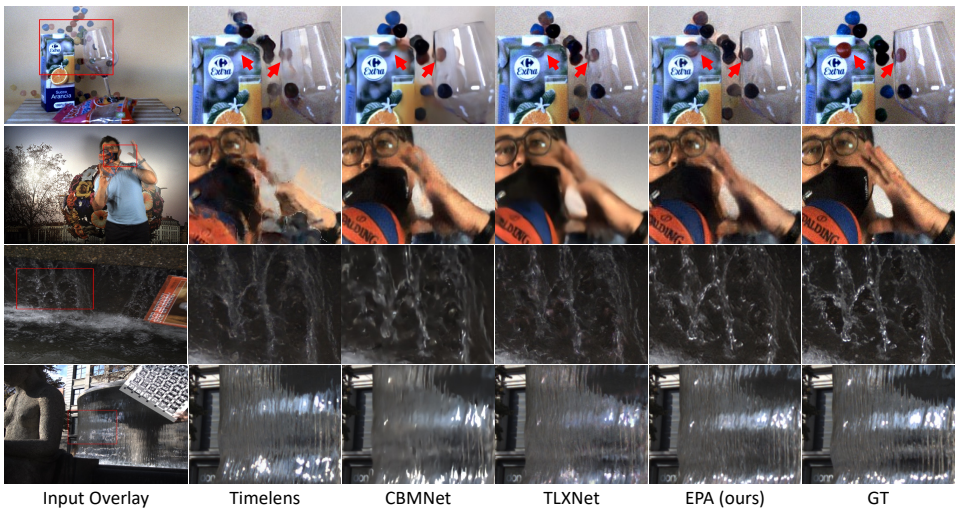


Figure 5: Visual comparison among different methods on real datasets.

better reflect human perceptual judgments of interpolation quality. For completeness, we also include results using conventional metrics such as PSNR and SSIM [44].

4.2 Evaluations on Synthesis Datasets

Quantitative Comparison Tab. 1 presents a comparative analysis between EPA and existing methods. Thanks to our training strategy and the precise guidance from events, our method consistently achieves the highest perceptual quality on both the Vimeo90k and GOPRO datasets. Notably, on the GOPRO dataset, EPA demonstrates superior alignment between scene content and human perception. This advantage stems from EPA's semantic feature-based fitting, which enables more effective capture of scene semantics and leads to significantly better performance in terms of the DISTS metric compared to other approaches.

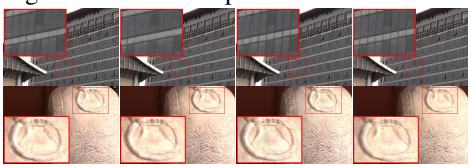
Qualitative Comparison Fig. 4 compares E-VFI methods on a synthetic dataset. Our method outperforms TimeLens, CBMNet, and TLXNet under blurred keyframes, producing clearer fingers and roofs. This superiority is attributed to our model's enhanced capability to capture scene semantics, enabling the synthesis of images that are better aligned with human perception. These results validate the effectiveness of our framework designs and perceptual feature fitting approach.

Figure 6: Visualization of effects and scores of inserted frames for each method in extreme motion scenes.

Table 3: Performance Comparison on EventAid-F.

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)	<u>31.43</u>	**0.015**	<u>35.15</u>	**0.011**

Figure 7: Visual comparison on EventAid-F.



4.3 Evaluations on Real Datasets

Quantitative Comparison Tab. 2 and Tab. 3 presents comparisons on real-world datasets, highlighting the practical effectiveness of our method. Our approach consistently leads in perceptual metrics across three datasets. Notably, on the HS-ERGB dataset, EPA also surpasses other methods in traditional image quality metrics such as PSNR and SSIM. We attribute this to the limitations of compared methods in handling the widespread irregular object deformations present in the data, while our method, leveraging semantic-level features, addresses these challenges more effectively.

Qualitative Comparison Fig. 5 and Fig. 7 presents a visual comparison of different E-VFI methods on a real-world event dataset. A direct comparison with the outputs of TimeLens, CBMNet, and TLXNet reveals the advantages of our method. Specifically, in Fig. 5, our approach ensures better visual quality in various extreme motion scenarios, as demonstrated by more complete candies, fingers, and better-preserved water flow details. From the images produced by CBMNet, we observe that when meet noisy keyframes or occlusion situations, the generated interpolated frames exhibit sever blur ($e.g.$, the candy) and poor-quality ($e.g.$, the water flow). Instead, our method can handle such scenerios more effectively on dynamic objects. Similarly, Fig. 7 shows that our method generates sharper and more perceptually aligned results for structures such as building and sculptures. We attribute to that benefiting from the perceptual features alignment with visual foundation models, our approach is able to synthesize more realistic images by comprehending the semantics of interpolated scenes to a certain extent. Interestingly, as illustrated in Fig. 6, although CBMNet and TLXNet achieve higher PSNR scores, our method still produces visually superior results in extreme scenarios, demonstrating the efficacy of using perceptual metrics ($e.g.$, lpips/dists) to measure the VFI task.

Table 4: Comparison results of different feature extraction settings in training stage one.

Variants	Norm.	A_s	Vimeo90k			
			PSNR	SSIM	LPIPS	DISTS
A			36.121	0.983	0.0038	0.0218
B	✓		40.793	0.991	0.0016	0.0086
C (ours)	✓	✓	**47.732**	**0.996**	**0.0002**	**0.0012**

Table 5: Comparison results of different $\mathcal{S}^\tau$ synthesis method settings in training stage two. $\mathcal{E}$ denotes events.

Variants	$\mathcal{E}$	SCA	BEGA	Vimeo90k				
				PSNR	SSIM	LPIPS	FloLPIPS	DISTS
D				28.834	0.857	0.0215	0.0481	0.0755
E	✓	✓		36.900	0.961	0.0078	0.0129	0.0375
F (ours)	✓		✓	**37.011**	**0.964**	**0.0074**	**0.0124**	**0.0361**

5 Model Analysis

In this section, we conduct experiments on the Vimeo90k dataset to analyze the effectiveness of the two core components in our proposed method: the Style Adapter and the Bidirectional Event-Guided Alignment module.

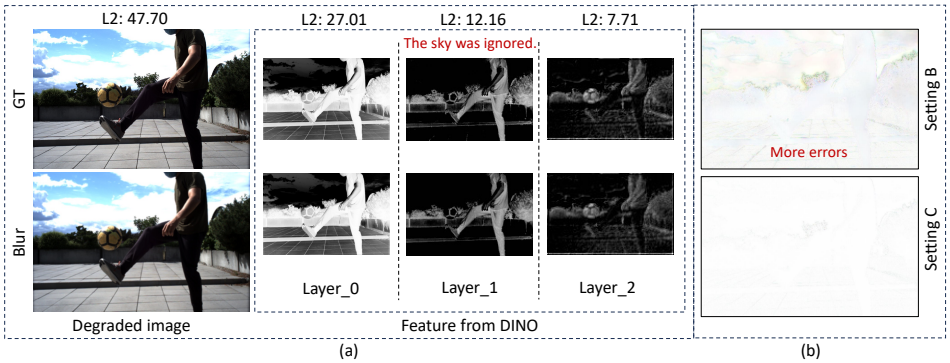


Figure 8: Visualization of comparison results. (a) compares L2 distances between image-level and feature-level supervision, alongside DINO feature visualizations. (b) shows difference maps between the generated results from settings B and C in Tab. 4 and the ground truth.

Rationality and effectiveness of feature-level supervision. In Fig. 8(a), we present a comparison between the differences induced by image-level supervision and those induced by feature-level supervision when an image undergoes image-level degradation, exemplified here by blurring. We use the L2 distance as an evaluation metric. It can be observed that the feature-level supervision signals extracted from DINO exhibit higher robustness to blur degradation. Furthermore, as the network depth increases, the robustness of the feature supervision becomes more pronounced. This improvement is attributed to the VFM's capability to capture image representations from a perceptual perspective, which helps mitigate the impact of image-level distortions, suggesting the reasonableness of our design philosophy to a certain extent.

Effectiveness of the Style Adapter (A_s). As presented in Tab. 4, to evaluate the proposed A_s, we introduce baseline model A, which utilizes the first three neural blocks of features from DINO, without incorporating shallow image details. Given that DINO is pretrained on ImageNet [5] with associated normalization, we introduce model B, which applies both normalization and denormalization processes. Model B normalizes input images using ImageNet statistics and denormalizes the outputs accordingly. Model C is built upon Model B and trained in combination with the proposed A_s. The results of settings A and B in Tab. 4 highlight the importance of the adaptation process. Furthermore, the comparison between Settings B and C demonstrates that our choice substantially improves the model's ability in reconstructing images from perceptual features, thereby enhancing the quality of the synthesized interpolated frames, as illustrated in Fig. 8(b).

Effectiveness of the Bidirectional Event-Guided Alignment (BEGA) Module. As shown in Tab. 5, to verify the importance of event guidance, we construct a baseline model D by removing all event data and performing feature fitting purely based on frames. Additionally, to assess the effectiveness of our specific design, we introduce a comparison model E, which employs a fusion module—referred to as SCA—comprising a self-attention layer followed by a cross-attention layer, similar to the structure proposed in [21]. The comparison between models D and E highlights the significance of event guidance. The precise inter-frame motion priors provided by events significantly enhance the model's ability to fit features of the interpolated frame. Furthermore, the performance gap between models E and F demonstrates the superiority of our proposed module.

6 Conclusion

In this work, we present EPA, a novel E-VFI framework that addresses the limitations of image-level degradation in modeling real-world data distributions. By leveraging degradation-insensitive semantic supervision, EPA enhances both the perceptual fidelity and generalization capability of frame synthesis under challenging conditions. EPA's two-stage training paradigm first extracts robust semantic-perceptual features and ensures reconstruction quality via a dedicated style adapter. In the second stage, the proposed Bidirectional Event-Guided Alignment module effectively aligns semantic features from keyframes to the ground truth, guided by fine-grained event cues, and synthesizes the interpolation via a pretrained reconstruction generator. Extensive experiments demonstrate that our method outperforms existing approaches in both perceptual quality and generalization, offering a more reliable solution for real-world scenarios affected by camera noise and motion blur.

7 Acknowledgments

This work is jointly supported by Beijing Natural Science Foundation (4252026), National Natural Science Foundation of China (62203024, 62573369), Research and Development Program of Beijing Municipal Education Commission (KM202310005027).

References

[1] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021.

[2] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 764–773, 2017.

[3] Duolikun Danier, Fan Zhang, and David Bull. Flolpips: A bespoke video quality metric for frame interpolation. In *2022 Picture Coding Symposium (PCS)*, pages 283–287. IEEE, 2022.

[4] Duolikun Danier, Fan Zhang, and David Bull. Ldmvfi: Video frame interpolation with latent diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1472–1480, 2024.

[5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[6] Keyan Ding, Kede Ma, Shiqi Wang, and Eero P Simoncelli. Image quality assessment: Unifying structure and texture similarity. *IEEE transactions on pattern analysis and machine intelligence*, 44(5):2567–2581, 2020.

[7] Peiqi Duan, Boyu Li, Yixin Yang, Hanyue Lou, Minggui Teng, Yi Ma, and Boxin Shi. Eventaid: Benchmarking event-aided image/video enhancement algorithms with real-captured hybrid dataset. *arXiv preprint arXiv:2312.08220*, 2023.

[8] Yue Gao, Siqi Li, Yipeng Li, Yandong Guo, and Qionghai Dai. Superfast: 200× video frame interpolation via event camera. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(6):7764–7780, 2023.

[9] Daniel Gehrig, Mathias Gehrig, Javier Hidalgo-Carrió, and Davide Scaramuzza. Video to events: Recycling video datasets for event cameras. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3586–3595, 2020.

[10] Weihua He, Kaichao You, Zhendong Qiao, Xu Jia, Ziyang Zhang, Wenhui Wang, Huchuan Lu, Yaoyuan Wang, and Jianxing Liao. Timereplayer: Unlocking the potential of event cameras for video interpolation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17804–17813, 2022.

[11] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7132–7141, 2018.

[12] Mengshun Hu, Kui Jiang, Zhihang Zhong, Zheng Wang, and Yinqiang Zheng. Iq-vfi: Implicit quadratic motion estimation for video frame interpolation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6410–6419, 2024.

[13] Xueyu Hu, Tao Xiong, Biao Yi, Zishu Wei, Ruixuan Xiao, Yurun Chen, Jiasheng Ye, Meiling Tao, Xiangxin Zhou, Ziyu Zhao, et al. Os agents: A survey on mllm-based agents for computer, phone and browser use, 2024.

[14] Zhewei Huang, Tianyuan Zhang, Wen Heng, Boxin Shi, and Shuchang Zhou. Real-time intermediate flow estimation for video frame interpolation. In *Proceedings of the European conference on computer vision*, pages 624–642, 2022.

[15] Siddhant Jain, Daniel Watson, Eric Tabellion, Ben Poole, Janne Kontkanen, et al. Video interpolation with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7341–7351, 2024.

[16] Xin Jin, Longhai Wu, Jie Chen, Youxin Chen, Jayoon Koo, and Cheul-hee Hahm. A unified pyramid recurrent network for video frame interpolation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1578–1587, 2023.

[17] Xin Jin, Longhai Wu, Guotao Shen, Youxin Chen, Jie Chen, Jayoon Koo, and Cheul-hee Hahm. Enhanced bi-directional motion estimation for video frame interpolation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5049–5057, 2023.

[18] Younghyun Jo, Sejong Yang, and Seon Joo Kim. Investigating loss functions for extreme super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 424–425, 2020.

[19] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *Proceedings of the European conference on computer vision*, pages 694–711, 2016.

[20] Tarun Kalluri, Deepak Pathak, Manmohan Chandraker, and Du Tran. Flavr: Flow-agnostic video representations for fast frame interpolation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2071–2082, 2023.

[21] Taewoo Kim, Yujeong Chae, Hyun-Kurl Jang, and Kuk-Jin Yoon. Event-based video frame interpolation with cross-modal asymmetric bidirectional motion fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18032–18042, 2023.

[22] Taewoo Kim, Yujeong Chae, Hyun-Kurl Jang, and Kuk-Jin Yoon. Event-based video frame interpolation with cross-modal asymmetric bidirectional motion fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18032–18042, 2023.

[23] Ke Li, Gengyu Lyu, Hao Chen, Bochen Xie, Zhen Yang, Youfu Li, and Yongjian Deng. Know where you are from: Event-based segmentation via spatio-temporal propagation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 4806–4814, 2025.

[24] Zhen Li, Zuo-Liang Zhu, Ling-Hao Han, Qibin Hou, Chun-Le Guo, and Ming-Ming Cheng. Amt: All-pairs multi-field transforms for efficient frame interpolation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9801–9810, 2023.

[25] Songnan Lin, Jiawei Zhang, Jinshan Pan, Zhe Jiang, Dongqing Zou, Yongtian Wang, Jing Chen, and Jimmy Ren. Learning event-driven video deblurring and interpolation. In *Proceedings of the European conference on computer vision*, pages 695–710, 2020.

[26] Xin Lin, Jingtong Yue, Kelvin C. K. Chan, Lu Qi, Chao Ren, Jinshan Pan, and Ming-Hsuan Yang. Multi-task image restoration guided by robust dino features, 2024.

[27] Yuhan Liu, Yongjian Deng, Hao Chen, Bochen Xie, Youfu Li, and Zhen Yang. Event-based video frame interpolation with edge guided motion refinement. *arXiv preprint arXiv:2404.18156*, 2024.

[28] Yuhan Liu, Yongjian Deng, Hao Chen, and Zhen Yang. Video frame interpolation via direct synthesis with the event-based reference. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8477–8487, 2024.

[29] Yuhan Liu, Linghui Fu, Hao Chen, Zhen Yang, Youfu Li, and Yongjian Deng. Event-based video interpolation via complementary motion information. *Engineering Applications of Artificial Intelligence*, 162:112606, 2025.

[30] Ilya Loshchilov, Frank Hutter, et al. Fixing weight decay regularization in adam. *arXiv preprint arXiv:1711.05101*, 5:5, 2017.

[31] Yongrui Ma, Shi Guo, Yutian Chen, Tianfan Xue, and Jinwei Gu. Timelens-xl: Real-time event-based video frame interpolation with large motion. In *Proceedings of the European conference on computer vision*, pages 178–194, Cham, 2025.

[32] Anish Mittal, Rajiv Soundararajan, and Alan C Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal processing letters*, 20(3):209–212, 2012.

[33] Seungjun Nah, Tae Hyun Kim, and Kyoung Mu Lee. Deep multi-scale convolutional neural network for dynamic scene deblurring. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3883–3891, 2017.

[34] Junheum Park, Keunsoo Ko, Chul Lee, and Chang-Su Kim. Bmbc: Bilateral motion estimation with bilateral cost volume for video interpolation. In *Proceedings of the European conference on computer vision*, pages 109–125, 2020.

[35] Junheum Park, Jintae Kim, and Chang-Su Kim. Biformer: Learning bilateral motion estimation via bilateral transformer for 4k video frame interpolation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1568–1577, 2023.

[36] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

[37] Fitsum Reda, Janne Kontkanen, Eric Tabellion, Deqing Sun, Caroline Pantofaru, and Brian Curless. Film: Frame interpolation for large motion. In *Proceedings of the European conference on computer vision*, pages 250–266, 2022.

[38] Zhihao Shi, Xiangyu Xu, Xiaohong Liu, Jun Chen, and Ming-Hsuan Yang. Video frame interpolation transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17482–17491, 2022.

[39] Lei Sun, Christos Sakaridis, Jingyun Liang, Peng Sun, Kai Zhang, Jiezhang Cao, Qi Jiang, Kaiwei Wang, and Luc Van Gool. Event-based frame interpolation with ad-hoc deblurring. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18043–18052, 2023.

[40] Stepan Tulyakov, Daniel Gehrig, Stamatios Georgoulis, Julius Erbach, Mathias Gehrig, Yuanyou Li, and Davide Scaramuzza. Time lens: Event-based video frame interpolation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16155–16164, 2021.

[41] Stepan Tulyakov, Alfredo Bochicchio, Daniel Gehrig, Stamatios Georgoulis, Yuanyou Li, and Davide Scaramuzza. Time lens++: Event-based frame interpolation with parametric non-linear flow and multi-scale fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17755–17764, 2022.

[42] Xintao Wang, Kelvin CK Chan, Ke Yu, Chao Dong, and Chen Change Loy. Edvr: Video restoration with enhanced deformable convolutional networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition workshops*, 2019.

[43] Xiao Wang, Zongzhen Wu, Yao Rong, Lin Zhu, Bo Jiang, Jin Tang, and Yonghong Tian. Sstformer: Bridging spiking neural network and memory support transformer for frame-event based recognition. *arXiv preprint arXiv:2308.04369*, 2023.

[44] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

[45] Guangyang Wu, Xin Tao, Changlin Li, Wenyi Wang, Xiaohong Liu, and Qingqing Zheng. Perception-oriented video frame interpolation via asymmetric blending. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2753–2762, 2024.

[46] Song Wu, Kaichao You, Weihua He, Chen Yang, Yang Tian, Yaoyuan Wang, Ziyang Zhang, and Jianxing Liao. Video interpolation by event-driven anisotropic adjustment of optical flow. In *Proceedings of the European conference on computer vision*, pages 267–283, 2022.

[47] Tianfan Xue, Baian Chen, Jiajun Wu, Donglai Wei, and William T Freeman. Video enhancement with task-oriented flow. *International Journal of Computer Vision*, 127(8):1106–1125, 2019.

[48] Bowen Yao, Yongjian Deng, Yuhan Liu, Hao Chen, Youfu Li, and Zhen Yang. Sam-event-adapter: Adapting segment anything model for event-rgb semantic segmentation. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 9093–9100. IEEE, 2024.

[49] Zhiyang Yu, Yu Zhang, Deyuan Liu, Dongqing Zou, Xijun Chen, Yebin Liu, and Jimmy S Ren. Training weakly supervised video frame interpolation with events. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14589–14598, 2021.

[50] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018.

[51] Xiang Zhang and Lei Yu. Unifying motion deblurring and frame interpolation with events. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17765–17774, 2022.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main contributions and scope are introduced in the abstract and the introduction.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper discusses the limitations of the work in the supplementary.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: All the theorems, formulas, and proofs in the paper are numbered and cross-referenced.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The paper fully discloses all the necessary information needed to reproduce the main experimental results, ensuring that the main claims and conclusions are verifiable. The methods, datasets, and evaluation metrics are described in sufficient detail to allow independent replication.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: All datasets used in this work are publicly available. The source code will be released upon acceptance.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We describe all the training and test details in the section Experiment.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: The paper reports statistical significance that is suitably and correctly defined, ensuring the reliability of the findings.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The paper indicates the type of computer workers, relevant memory, and storage.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The paper conforms, in every respect, with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The paper discusses both positive societal impacts and negative societal impacts of the work performed.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.

- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The paper poses no such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All assets used in this work, including datasets and models, are properly cited in the paper.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: This study does not introduce any new assets. All resources used are publicly available and therefore require no additional documentation.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: the paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.