CAUSAL-FLOW: DECONFOUNDING DARK OPTICAL FLOW ESTIMATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Dark optical flow estimation is fundamentally challenged by the entanglement of true motion and degradation-induced confounders in the observed information. Existing paradigms treat this entangled representation holistically, applying enhancement or suppression without explicitly identifying the degradation contribution, thereby incurring an inherent trade-off between preserving true motion and suppressing spurious responses. In this paper, we propose Causal-Flow, a causal decoding framework built on a dual causal design. First, a Causal Intervention Module (CIM) performs front-door adjustment, projecting motion features onto a low-rank subspace spanned by true motion patterns to remove high-dimensional stochastic confounders. Second, a Counterfactual Gating Unit (CGU) estimates degradation-induced spurious motion and derives adaptive spatial and channel gates from causal residuals to enhance true motion and suppress ambiguities. By embedding the CGU into each gate of the GRU-based decoder, Causal-Flow enables single-step decoding without iterative refinement, circumventing error accumulation and achieving state-of-the-art performance on dark-scene benchmarks.

1 INTRODUCTION

When optical flow encounters dark scenes, underlying degradation from noise and attenuated textures entangles with motion evidence, undermining reliable optical flow computationFeijoo et al. (2025); Wang et al. (2022); Zuo et al. (2026a). In view of this, the inter-frame intensity variation in such scenes, *i.e.*, the observed motion $\mathcal{M}_{obs}$, stems from two superimposed factors, formulated as:

$$\mathcal{M}_{obs} = \mathcal{M}_t + \mathcal{M}_d, \tag{1}$$

where $\mathcal{M}_t$ denotes the true motion component induced by the per-pixel displacement between adjacent frames; $\mathcal{M}_d$ denotes the degradation-induced component that is irrelevant to motion. Here we adopt the additive form as a tractable first-order approximation of the underlying entanglement, which renders the degradation component explicitly. Compared with Fig. 1(a), Fig. 1(b) exhibits severe entanglement between dark interference and motion vectors. Early methods proposed synthetic datasets with dark noise and color distortion, like FCDN and VBOF, to help well-lit methods narrow the gap, that is, dark-to-bright adaptation Zheng et al. (2020); Zhang et al. (2021). Subsequently, the dedicated DOFE method further boosts performance via targeted amplification of motion-induced pixel intensities. Representative methods, such as, CEDFlow Zuo et al. (2024) build illumination-invariant representations to extract primitives robust against dark degradation; FlowFM Zuo et al. (2026b) recasts its entire reasoning paradigm from reliance on brittle visual structures to the low-dimensional manifold topology of motion.

Despite the steady accuracy gains of the aforementioned methods Xu et al. (2022); Luo et al. (2023; 2021); Liu et al. (2021); Li et al. (2021), conventional paradigms harbor inherent discriminative limitations. As shown in Fig. 1(c), they fail to explicitly decouple the motion component from the degradation component, instead treating the two as a whole and refining motion evidences with heuristic enhancement and suppression schemes, which mistake amplified spurious responses for true motion and propagate them into the flow field. On one hand, amplifying response features to reinforce authentic motion signals inevitably magnifies prominent noise peaks and yields spurious flow. On the other hand, suppressing noisy-response signals eliminates subtle yet valid motion cues in dark areas. Such entangled processing induces an inherent trade-off: stringent noise suppression

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.

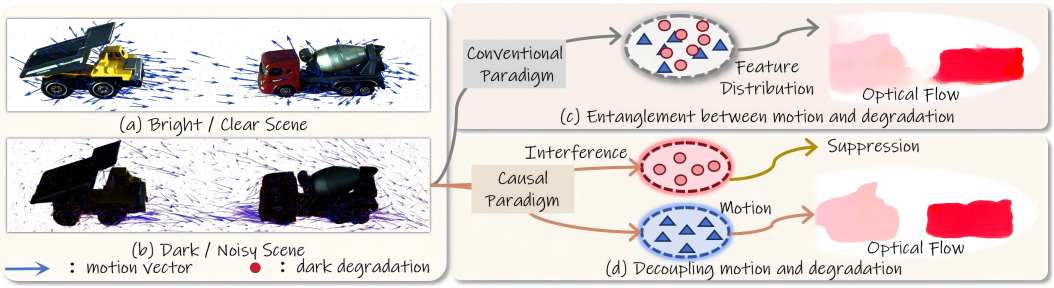

Figure 1: **Ideal illustration**. (a) Well-illuminated scene with clear motion vectors; (b) dark noisy scene, which induces substantial deviations in motion reasoning. The existing paradigm (c) applies enhancement directly to the coupled noise-motion feature space. In contrast, our idea (d) disentangles motion from confounding interference via a causal paradigm, yielding more accurate motion.

incurs the loss of fine-grained motion details, whereas excessive enhancement introduces spurious motion stemming purely from noise and other dark-scene degradations even within regions lacking genuine motion. Prevailing methods lack robust disentanglement capacity between true motion and dark-induced degradation components, underscoring the pressing need for dedicated decoupling methods tailored to dark scenes.

Causal modeling has yielded substantial advances across a broad spectrum of research domainsLesci et al. (2025); Mencattini et al. (2026); Almodóvar et al. (2026); Shi et al. (2025); Duan et al. (2026); Hong et al. (2026). Motivated by these developments, we explore its potential to tackle the aforementioned entanglement challenge. See Fig. 1(d), the causal paradigm furnishes a principled framework for disentangling true motion $\mathcal{M}_t$ and degradation-induced interference $\mathcal{M}_d$. Its core insight is that the degradation confounding can be identified through a per-pixel counterfactual query, that is, for any given location, what would the observed motion response be if true target motion were absent? The answer constitutes a counterfactual nuisance baseline $\mathcal{M}_d$, which encapsulates all spurious responses arising from dark-scene degradations. This reasoning adapts to local noise levels, naturally accommodating spatially varying degradation across scenes. Isolating $\mathcal{M}_t$ from $\mathcal{M}_{obs}$ thus constitutes the core of our causal paradigm, achievable through two complementary paths. In detail, **i) Implicitly**, by seeking a representation subspace invariant to $\mathcal{M}_d$; **ii) Explicitly**, by estimating $\mathcal{M}_d$ counterfactually and subtracting it from $\mathcal{M}_{obs}$.

To efficiently decouple the two components, we devise a dual causal decoder framework. **i) Causal Intervention Module (CIM).** True motion $\mathcal{M}_t$ is a low-dimensional, structurally continuous, and cross-frame consistent signal governed by physical laws, exhibiting smooth boundaries and temporal coherence. In contrast, the confounding component $\mathcal{M}_d$ is a high-dimensional, stochastic, and unstructured signal arising from pixel-wise sensor noise and spurious peaks in the cost volume. Instead of explicitly inferring the complex degradation-induced $\mathcal{M}_d$, we seek a subspace of $\mathcal{M}_{obs}$ that is exclusively driven by $\mathcal{M}_t$ and invariant to $\mathcal{M}_d$. At the initial decoding stage, CIM projects the observed motion evidence onto a low-rank subspace spanned by learnable orthonormal bases, directly mitigating the deleterious effects of dark-scene confounders. **ii) Counterfactual Gating Unit (CGU).** Following the causal purification, we introduce counterfactual gating that predicts per-pixel interference baselines, *i.e.*, an estimate of $\mathcal{M}_d$, from contextual features and observed motion features. Continuous gating weights are derived from the causal residual $\Delta\mathcal{M} = \mathcal{M}_{obs} - \mathcal{M}_d$. Larger residuals dominated by true motion yield higher gate values, whereas smaller residuals corresponding to pure noise produce lower scores. CIM and CGU operate in tandem for causal motion decoding, enabling single-pass hidden-state updates and circumventing the iterative buildup of spurious flow errors. Our contributions are as follows:

- **A causal framework.** Causal-Flow emerges as a pioneering model designed to estimate dark optical flow, integrating the dual causal design reasoning into the neural network, offering a new perspective for better performance.
- **A two-tier causal reasoning.** Our flow decoder integrates a causal intervention module and a counterfactual gating unit. The former suppresses high-dimensional stochastic noise via low-rank subspace projection, whereas the latter estimates degradation-induced spurious responses to amplify true motion signals and remove artifacts.

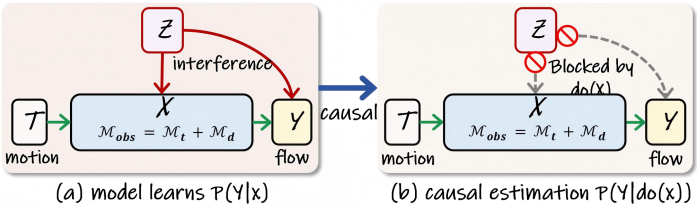

Figure 2: **Causal reasoning in the dark**. (a) Dark degradation Z confounds the causal chain $T \rightarrow X \rightarrow Y$ via $Z \rightarrow X$ and $Z \rightarrow Y$, so a discriminative model learns $P(Y|X)$ and mistakes degradation-induced responses for motion. (b) Our modules block the confounding path and estimate $P(Y|\mathrm{do}(X))$, so that the flow is determined only by the true motion $\mathcal{M}_t$.

- **State-of-the-art performance**. Extensive experiments on benchmark datasets demonstrate that our causal method achieves better performance, producing clean, edge-sharp flow fields even in dark high-noise scenes.

2 Core Challenges

To clarify how darkness degrades optical flow, we build a causal graph (Fig. 2) defined over four variables, namely true motion T, dark degradation Z, observed motion evidence X, and estimated flow Y. X corresponds to the observed motion evidence $\mathcal{M}_{obs}$ in Eq.1, covering both the encoded motion features and the update-stage features, since the superposition persists throughout the reasoning process. Z denotes the dark degradation imprinted in features as $\mathcal{M}_d$, and T denotes the true motion $\mathcal{M}_t$; the estimated flow Y is decoded from the motion together with the context features.

The chain $T \rightarrow X \rightarrow Y$ constitutes the desired estimation process. True motion produces matching evidence, that is, encoded into $\mathcal{M}_{obs}$, from which the decoder produces Y as an estimation of T. Importantly, the core challenge stems from the confounder Z, which opens the back-door path $X \leftarrow Z \rightarrow Y$ through two edges. Along $Z \rightarrow X$, dark degradation induces spurious responses that are propagated into the motion features, yielding the entangled superposition. Along $Z \rightarrow Y$, the estimator itself is biased, as the decoding update fuses degradation-corrupted context priors, where flows decoded in dark regions are systematically damped or over-smoothed. Consequently, discriminative models that learn $P(Y \mid X)$ cannot discern whether a feature response arises from true displacement or from degradation, and iterative decoders repeatedly amplify such spurious responses, accumulating them into flow field drift.

Our core insight is to block the confounding path by causal reasoning, *i.e.*, to estimate $P(Y \mid \mathrm{do}(X))$, where the $\mathrm{do}(\cdot)$ operator Pearl et al. (2003) denotes an intervention that exogenously fixes X and severs its incoming influences, so that the flow is determined only by the true motion $\mathcal{M}_t$. Since Z is unobserved and spatially dynamic, explicitly modeling it is infeasible; we thus realize the intervention with two modules integrated into the decoder. $\mathcal{M}_t$ is low-dimensional and structured, whereas $\mathcal{M}_d$ is high-dimensional and stochastic. CIM projects $\mathcal{M}_{obs}$ onto a learnable low-rank subspace, thereby filtering confounding components at the source without ever modeling Z. CGU predicts a per-pixel counterfactual baseline, *i.e.*, an estimate of $\mathcal{M}_d$ that captures spurious features induced solely by degradation in the absence of true motion, and gates features via the causal residual $\Delta\mathcal{M} = \mathcal{M}_{obs} - \mathcal{M}_d$, amplifying true motion while suppressing degradation-induced responses.

3 Method

3.1 Overview

Fig. 3 depicts the overall architecture of our proposed Causal-Flow, which integrates causal reasoning into the decoding pipeline. Given dark frames T and $\mathrm{T}+1$, a feature encoder and a context encoder extract base features $\mathrm{x}_1, \mathrm{x}_2$ and contextual features f_c, respectively. A cost volume cv is constructed based on x_1 and x_2. Following the flow-matching paradigm adopted in FlowFM Zuo et al. (2026b), the noisy flow f_n is synthesized from Gaussian noise ϵ, ground-truth flow f^{gt}, and initial flow f_0. The cost volume cv and noisy flow f_n are further encoded into motion features f_m with a motion head, which undergo two sequential purification stages. The CIM eliminates source-side

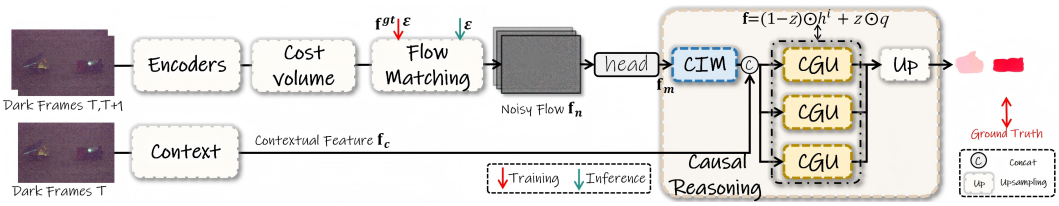

Figure 3: **The overall architecture**. During training, the noisy flow f_n is produced by corrupting the ground truth f^{gt} with noise ε via flow matching, while at inference it is obtained by adding ε to the initial flow f_0. The noisy flow and context feature f_c are fused and fed into the causal reasoning decoder, where CIM purifies the motion evidence and CGU calibrates each gate of the update unit.

feature corruption via front-door causal adjustment, and the CGU refines the purified features with contextual feature f_c through counterfactual residual gating to yield the final flow f. $\mathcal{L}_1$ serves as the supervision loss that quantifies the discrepancy between the final flow f and ground-truth f^{gt}.

3.2 Causal Intervention Module

Conventional causal methods rely on explicitly constructed confounders or mediators tailored to semantic factors, and are ill-suited for dark flow, where Z is non-semantic, continuous, and spatially dynamic. Hence, we devise a degradation-oriented front-door method termed the causal intervention module. The front-door adjustment Pearl et al. (2003) identifies the causal effect through a mediator feature F, *i.e.*, an intermediate representation on the causal path from X to Y: $P(Y \mid \mathrm{do}(X)) = \mathbb{E}_{F\sim P(F|X)}[\mathbb{E}_{X'\sim P(X)}[P(Y|X',F)]]$, where X' denotes values of X sampled from its marginal distribution $P(X)$, and the nested expectations are intractable during training. The normalized weighted geometric mean Xu et al. (2015) approximation exchanges the order of the expectations and the prediction function, reducing the adjustment to a deterministic estimate of two expectations:

$$P(Y \mid \mathrm{do}(X)) \approx P(Y|\mathbb{E}[X'] + \mathbb{E}[F]). \tag{2}$$

where $\mathbb{E}[X']$ aggregates generalizable information from the input, and $\mathbb{E}[F]$ summarizes the mediator that transmits the causal effect; together with the mediator feature F, they constitute the intervened feature. Since $\mathcal{M}_t$ is low-dimensional and structured while $\mathcal{M}_d$ is high-dimensional and stochastic, the two expectations can be captured by a few low-rank basis vectors while filtering confounding components, without ever modeling Z. Thus, CIM estimates them by query-guided weighting and low-rank basis projection, as detailed below.

Given the input motion feature $\mathrm{f}_m \in \mathbb{R}^{B\times C\times N}$ (flattened to $N{=}H\cdot W$), CIM first constructs the mediator feature with a convolutional block $\sigma(\cdot)$, $F = \sigma(\mathrm{f}_m)$, and then estimates the two expectations with a generic estimator applied to both f_m and F. For a generic variable $\mathrm{V} \in \{\mathrm{f}_m, F\}$, a set of sample response maps aggregates cross-sample generalizable information:

$$A = \sum_{s=1}^{S} \mathrm{softmax}(R)_s \cdot R_s, \quad \mathrm{V}_q = A \odot \mathrm{V}, \tag{3}$$

where $R = \sigma(\mathrm{V}), R \in \mathbb{R}^{B\times S\times N}$ contains S sample response maps and A is the aggregated spatial weighting. The query-guided feature V_q is then projected onto a learnable low-rank basis $\mathbf{B} \in \mathbb{R}^{K\times C}$ ($K < C$) initialized as an orthonormal basis, with the rank set to $K = C/2$ (64 for 128-dimensional features), as given by

$$\hat{\mathbb{E}}[\mathrm{V}] = \mathrm{V}_q\mathbf{B}^\top(\mathbf{B}\mathbf{B}^\top + \epsilon\mathbf{I})^{-1}\mathbf{B}, \tag{4}$$

where $(\mathbf{B}\mathbf{B}^\top + \epsilon\mathbf{I})^{-1}$ normalizes the basis that may drift from orthogonality during training. This projection reconstructs V_q within the low-rank causal subspace, filtering out-of-subspace confounding components. The intervened feature is finally composed as in Eq.5:

$$\mathrm{f}_m^{\mathrm{CIM}} = \mathrm{f}_m + \gamma \cdot \sigma(F + \hat{\mathbb{E}}[\mathrm{f}_m] + \hat{\mathbb{E}}[F]), \tag{5}$$

where γ is zero-initialized, so that CIM degenerates to an identity mapping at initialization. As γ is learned during training, it progressively evolves into an adaptive parameter. In this composition, the mediator feature F retains task-relevant information, while the expectation terms carry out the front-door adjustment.

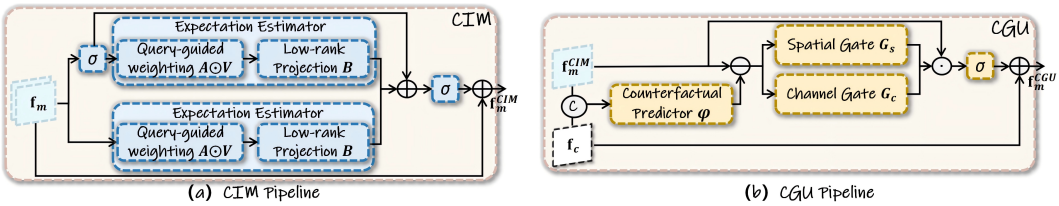

Figure 4: **The CIM and CGU architecture**. (a) CIM estimates the two expectations through query-guided weighting and low-rank basis projection, and composes the intervened feature with the mediator via a residual branch. (b) CGU predicts a counterfactual baseline from $\mathrm{f}_m^{\mathrm{CIM}}$ and the context feature f_c, derives spatial and channel gates from the causal residual, and calibrates residually.

3.3 Counterfactual Gating Unit

Within the path $X \leftarrow Z \rightarrow Y$, CIM mitigates the corruption along the $Z \rightarrow X$ edge through implicit low-rank subspace projection, while the estimator bias along $Z \rightarrow Y$ remains unaddressed. To mitigate this estimation bias, we propose the counterfactual gating unit, which implements the causal intervention $P(Y \mid \mathrm{do}(X))$ via counterfactual inference during decoding. CGU circumvents this limitation by exploiting the additive property of degradation-induced feature disturbances. Rather than marginalizing over Z, it detects and removes the confounder's footprint directly in feature space, thereby realizing an equivalent intervention.

Specifically, CGU takes the purified motion feature $\mathrm{f}_m^{\mathrm{CIM}}$ and the context feature f_c as inputs, and first predicts a per-pixel counterfactual baseline f_d, which corresponds to the feature estimate of the confounding component $\mathcal{M}_d$. Subtracting f_d from $\mathrm{f}_m^{\mathrm{CIM}}$ yields the causal residual $\Delta\mathrm{f}_m$, which explicitly characterizes the true motion:

$$\mathrm{f}_d = \varphi([\mathrm{f}_m^{\mathrm{CIM}}; \mathrm{f}_c]), \qquad \Delta\mathrm{f}_m = \mathrm{f}_m^{\mathrm{CIM}} - \mathrm{f}_d, \tag{6}$$

where $[\cdot;\cdot]$ denotes channel-wise concatenation and φ is a convolutional predictor. The context feature f_c provides structural cues for inferring degradation-induced spurious responses. The residual then drives two continuous gates for spatial and channel modulation, respectively:

$$\begin{aligned} \mathbf{G}_s &= \mathrm{sigmoid}(\mathrm{mean}_c(\Delta\mathrm{f}_m)), \\ \mathbf{G}_c &= \mathrm{sigmoid}(\mathrm{SE}(\Delta\mathrm{f}_m)), \end{aligned} \tag{7}$$

where mean_c denotes averaging along the channel dimension, $\mathrm{SE}(\cdot)$ is the squeeze-and-excitation operation Hu et al. (2018). Both gates produce continuous values rather than binary masks, enabling adaptive suppression strength for diverse degradation scenarios without hard thresholds. The calibrated feature is then obtained through residual gating fusion:

$$\mathrm{f}_m^{\mathrm{CGU}} = \mathrm{f}_m^{\mathrm{CIM}} + \gamma \cdot \sigma(\mathrm{f}_m^{\mathrm{CIM}} \odot \mathbf{G}_s \odot \mathbf{G}_c), \tag{8}$$

where $\odot$ denotes element-wise multiplication and γ is a zero-initialized scalar, such that the CGU reduces to an identity mapping at initialization. We further integrate CGU as the core unit of all three gates in the standard GRU, yielding the causal recurrent decoder (CRD):

$$\begin{aligned} z &= \mathrm{sigmoid}(\mathrm{CGU}_z([h^i; \mathrm{f}_m^{\mathrm{CIM}}])), \\ r &= \mathrm{sigmoid}(\mathrm{CGU}_r([h^i; \mathrm{f}_{\mathrm{m}}^{\mathrm{CIM}}])), \\ q &= \tanh(\mathrm{CGU}_q([\mathbf{r} \odot h^i; \mathrm{f}_m^{\mathrm{CIM}}])), \\ h^{i+1} &= (1 - z) \odot h^i + z \odot q, \end{aligned} \tag{9}$$

where h^i denotes the i-th hidden state, which is initialized as h^0 from the context feature f_c, that is $h^0 = \mathrm{f}_c$. In CRD, each gate predicts its own counterfactual baseline f_d from its respective inputs, and each gating operation distinguishes valid true-motion signals from degradation-induced spurious features, preventing misleading responses from contaminating the hidden state. Lastly, a flow head decodes the flow increment from h^{i+1}, which is added to the initial flow f_0 and upsampled to the full resolution as the output f.

Table 1: EPE comparison of different optical flow evaluations on FCDN and VBOF datasets (trained on FCDN dataset). Bold font indicates the best performance.

Method	Reference	EPE(trained on FCDN)↓									
		FCDN	Canon	Fuji	Fuji2	Nikon	Nikon2	Sony	Sony2	Sony3	VBOF
RAFT	ECCV-20	1.23	20.25	39.52	14.20	26.40	25.45	15.61	32.43	19.16	21.84
GMA	CVPR-21	1.18	20.37	39.64	14.40	27.54	26.07	15.83	30.23	19.04	21.77
KPAFlow	CVPR-22	1.24	22.19	47.21	14.71	29.97	27.58	16.15	36.65	21.61	23.10
FlowFormer++	CVPR-23	1.12	20.09	39.34	14.03	25.49	25.46	15.10	30.32	18.87	20.94
Flow*Diffuser*	CVPR-24	1.09	19.83	36.74	14.00	25.32	24.26	15.17	30.57	19.14	20.77
CEDFlow	AAAI-24	1.08	19.69	39.21	13.94	25.24	25.09	15.16	30.30	18.73	20.89
DPFlow	CVPR-25	1.17	19.88	37.20	14.32	26.07	25.56	15.20	27.73	19.29	20.92
CEDFlow++	IJCV-25	1.01	20.07	34.82	13.93	25.85	24.34	15.05	28.58	19.00	20.56
PsFE	TCSVT-26	0.85	19.51	33.74	13.94	25.17	24.57	15.03	26.98	18.77	20.31
FlowFM	CVPR-26	0.87	12.09	23.96	8.99	17.38	16.59	9.82	16.65	11.97	13.28
Causal-Flow	-	**0.81**	**11.67**	**22.61**	**8.97**	**17.03**	**16.16**	**9.80**	**14.83**	**11.82**	**12.96**

Table 2: EPE comparison of different optical flow evaluations on FCDN and VBOF datasets (trained on Mix dataset). Bold font indicates the best performance.

Method	Reference	EPE(trained on Mix)↓									
		FCDN	Canon	Fuji	Fuji2	Nikon	Nikon2	Sony	Sony2	Sony3	VBOF
RAFT	ECCV-20	1.38	8.25	16.77	7.34	11.43	12.06	6.00	11.56	11.29	8.89
GMA	CVPR-21	1.26	6.33	10.11	4.90	11.56	11.96	5.16	7.04	6.63	6.81
KPAFlow	CVPR-22	1.39	7.36	12.70	6.11	10.14	10.51	6.22	8.71	8.37	7.47
FlowFormer++	CVPR-23	1.24	6.22	9.73	4.77	8.38	9.72	4.64	6.74	6.82	6.46
Flow*Diffuser*	CVPR-24	1.20	6.38	9.74	4.40	9.32	9.63	5.11	6.57	6.89	6.57
CEDFlow	AAAI-24	1.23	6.22	9.41	4.69	8.23	9.87	4.66	6.50	6.10	6.52
DPFlow	CVPR-25	1.24	6.31	9.55	4.56	7.80	9.50	4.74	6.23	6.43	6.21
CEDFlow++	IJCV-25	1.14	5.85	9.66	4.38	7.45	9.29	4.63	6.18	6.77	6.25
PsFE	TCSVT-26	1.08	**4.72**	8.59	4.32	7.31	9.11	4.48	5.44	6.39	5.43
FlowFM	CVPR-26	1.06	5.83	7.03	3.30	6.75	7.85	3.74	4.37	5.98	5.24
Causal-Flow	-	**1.02**	5.23	**6.75**	**3.15**	**6.54**	**7.57**	**3.62**	**4.32**	**5.06**	**4.94**

4 EXPERIMENTS

4.1 IMPLEMENTATION DETAILS

Following prior works Zuo et al. (2024; 2025), Causal-Flow was trained with AdamW with gradient clipping [-1, 1], one-cycle learning rate, random initialization, and input 368×496. The proposed Causal-Flow was trained on full FCDN (350k iterations) and full Mix (400k) datasets, with a batch size of 4 and a learning rate of 2.5e-4 for both.

Similar to prior work Zuo et al. (2024; 2025), we evaluate Causal-Flow on FCDN Zheng et al. (2020), Mix (FCDN + VBOF Zhang et al. (2021)), and real-world dark image/video datasets (SMID Chen et al. (2019), LOM Cui et al. (2024), GOF Li et al. (2021) and LLRS Sun et al. (2025)), with end-point error (EPE) Dosovitskiy et al. (2015) as the metric, against 10 advanced flow methods (RAFTTeed & Deng (2020), GMAJiang et al. (2021), KPAFlow Luo et al. (2022), FlowFormer++ Shi et al. (2023), Flow*Diffuser* Luo et al. (2024), DPFlow Morimitsu et al. (2025)) and DOFE-oriented methods (CEDFlow Zuo et al. (2024), CEDFlow++ Zuo et al. (2025), PsFEZuo et al. (2026a) and FlowFMZuo et al. (2026b).

4.2 COMPARISON WITH STATE-OF-THE-ARTS

Trained on FCDN. As shown in **Tab. 1**, the proposed Causal-Flow achieves the best EPE score of 0.81 on the FCDN dataset, which outperforms the recent PsFE by 4.7% (0.85 → 0.81) and the strong baseline FlowFM by 6.9% (0.87 → 0.81). Remarkably, Causal-Flow maintains consistent advantages on VBOF and its subsets. Specifically, columns 4 to 12 report the generalization evaluation on the VBOF dataset, where Causal-Flow obtains a 12.96 EPE value on the full VBOF sample. This result surpasses the recent PsFE by 36.2% (20.31 → 12.96) and further improves over FlowFM

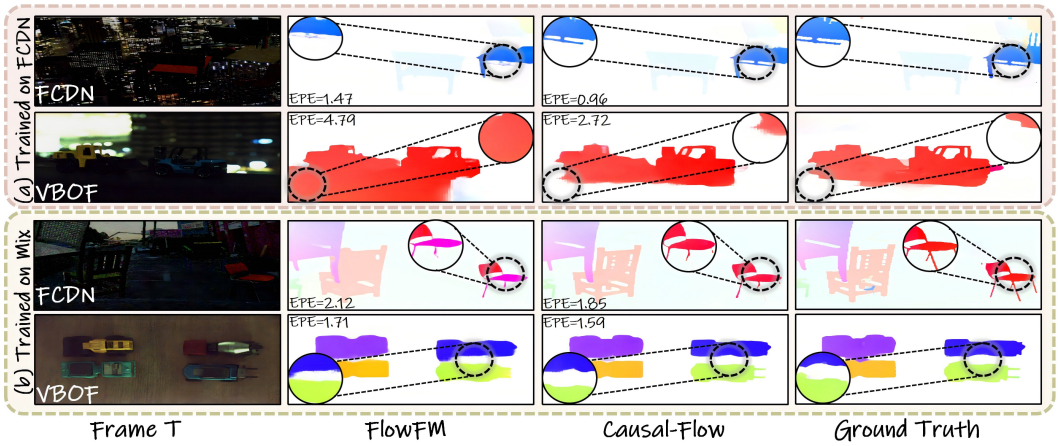

Figure 5: Qualitative comparison between FlowFM and the proposed Causal-Flow. (a) shows the results of models trained on FCDN and (b) shows those trained on Mix. The EPE value calculated using the ground truth is annotated on each prediction. Best viewed zoomed in.

by 2.4% (13.28 $\rightarrow$ 12.96), with gains observed on every camera subset. We attribute this strong generalization to the causal disentanglement of motion and degradation. CIM purifies the observed motion evidence at the source, while CGU suppresses the degradation-induced estimation bias during decoding. The model thus learns the interventional $P(Y \mid \mathrm{do}(X))$ rather than the correlational $P(Y \mid X)$, which yields robust performance across dark, high-noise scenes.

Trained on Mix. Tab. 2 reports the quantitative evaluation of models trained on the more complex Mix dataset. VBOF's more complex camera and synthetic noise patterns induce a domain shift that slightly elevates all methods' FCDN scores relative to Tab. 1, while VBOF performance benefits substantially from broader training data. Even under this harder training setting, the proposed Causal-Flow achieves the best EPE of 1.02 on the FCDN dataset, followed by FlowFM with 1.06 and PsFE with 1.08. Causal-Flow also secures the top EPE scores on nine of the ten evaluations. The only exception is the Canon subset, which exhibits the mildest degradation in VBOF. Since the entanglement of motion and degradation is weak in such scenes, the benefit of causal disentanglement diminishes. On the full VBOF, it reaches 4.94, outperforming the strong baseline FlowFM by 5.7% (5.24 $\rightarrow$ 4.94) and the recent PsFE by 9.0% (5.43 $\rightarrow$ 4.94). While advanced methods such as DPFlow and Flow*Diffuser* perform competitively, they still struggle under low illumination variations, whereas Causal-Flow maintains its advantage across diverse cameras and noise types.

4.3 Visual Comparison with State-of-the-Arts

Evaluation on FCDN and VBOF. Fig. 5 provides a qualitative comparison between the strong baseline FlowFM and the proposed Causal-Flow. Each visualization result incorporates the EPE value calculated using the ground truth. Clearly, our Causal-Flow delivers more precise dark optical flow predictions, underscoring its exceptional generalization and robustness. Take the 2nd row as an example, a cross-dataset case where the model trained on FCDN is directly evaluated on VBOF. FlowFM produces spurious flow responses in the background above the vehicle and reaches an EPE of 4.79, while Causal-Flow suppresses these artifacts and reduces the error to 2.72. A similar trend appears in the 4th row with the multi-vehicle scene, where Causal-Flow delineates sharper contours and finer details around the circled regions.

Evaluation on Real-world Scenes. To investigate Causal-Flow's performance in real-world dark scenes, **Fig. 6** presents a generalization evaluation on the SMID, LOM, GOF, and LLRS datasets. Notably, these cases lack optical flow ground truth. Obviously, our Causal-Flow exhibits superior performance in these real-world conditions. In the SMID case, FlowFM nearly misses the moving pedestrian in the circled region, whereas Causal-Flow recovers a complete and sharp silhouette. In the GOF case with a fast-moving vehicle at night, Causal-Flow produces a more coherent and distinctive motion estimate for the car, while FlowFM's prediction is diffuse. Causal-Flow also delineates cleaner contours for the indoor object in LOM and recovers the building structure in the dark region of LLRS, where FlowFM produces smoothed and washed-out predictions.

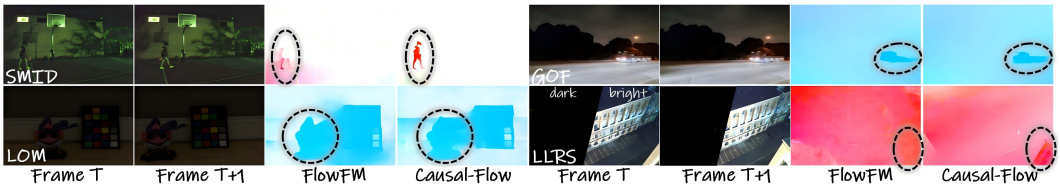

Figure 6: Real-world dark optical flow results. Best viewed zoomed in.

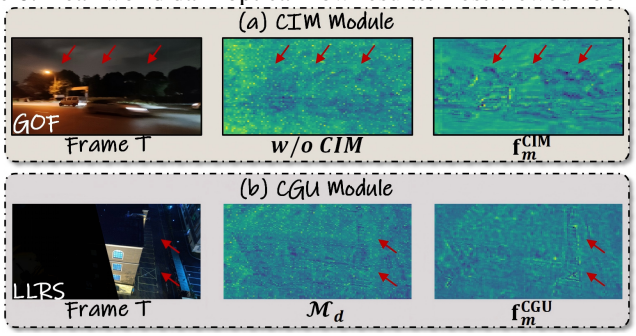

Figure 7: Feature visualization of CIM and CGU. Best viewed zoomed in.

4.4 Ablation Study

Ablation on Causal Intervention Module. Tab. 3 (2nd-8th rows) examines the contribution of each design within CIM. Removing the entire module raises the VBOF error to 13.25, close to the baseline, which verifies that disentangling the degraded motion evidence is the foundation of the subsequent decoding. Replacing CIM with the recent causal hypergraph (STCH) Wang et al. (2026) enlarges the model to 15.77 M parameters yet still lags behind on both benchmarks, confirming that the gains of CIM stem from the causal design rather than the model capacity. Among the internal components, the query-guided weighting and the low-rank basis have the greatest impact, as they are responsible for aggregating reliable cross-sample responses and suppressing out-of-subspace confounding components, respectively. The absence of $\hat{\mathbb{E}}[F]$ mainly hurts the accuracy on FCDN, suggesting that the mediator expectation provides complementary causal information beyond the input expectation. Both the zero-initialized γ and the refinement block $\sigma(\cdot)$ bring further gains.

Ablation on Counterfactual Gating Unit. As shown in **Tab. 3** (9th-16th rows), removing the entire CGU and replacing it with a plain convolutional block raises the EPE from 0.81 to 1.00 on FCDN and from 12.96 to 13.18 on VBOF, confirming that calibrating the estimator bias during decoding is indispensable. Next, we replace CGU with the interactive de-confounding module (IDM) Tang et al. (2026). Despite more parameters, IDM yields even worse results than removing CGU, indicating that generic causal modules cannot capture the degradation-induced error in the dark. Removing $\varphi(\cdot)$ raises the EPE to 0.89 on FCDN and 13.14 on VBOF, which verifies that explicitly estimating the degradation baseline is critical, while discarding f_c also degrades both benchmarks, as contextual cues guide the baseline prediction. The spatial gate $\mathbf{G}_s$ and the channel gate $\mathbf{G}_c$ each contribute to the final performance, and removing $\mathbf{G}_s$ causes a larger drop since it directly localizes spurious responses. Both the zero-initialized γ and the $\mathrm{SE}(\cdot)$ operation bring further gains.

Analysis on Causal Recurrent Decoder. As shown in **Tab. 3** (last three rows), replacing the CRD with a vanilla one yields the worst EPE despite the fewest parameters, confirming that recurrent causal calibration is essential. Moreover, increasing the iteration number i from 1 to 2 brings extra parameters but degrades both benchmarks, so we adopt a single-pass update.

Ablation Visualization. Fig. 7 visualizes the feature behaviors of the two modules. In (a), the motion features without CIM are overwhelmed by dense strong-response interference, as indicated by the arrows, whereas the purified $\mathrm{f}_m^{\mathrm{CIM}}$ exhibits structured patterns concentrated on moving objects, confirming that CIM filters out-of-subspace confounding components. In (b), the predicted f_d responds mainly to degradation-dominated regions such as the illuminated edges, while the calibrated $\mathrm{f}_m^{\mathrm{CGU}}$ preserves the structural motion evidence and suppresses these degradation responses, which verifies that the causal residual drives effective counterfactual gating.

Table 3: Ablation analysis for different components.

Method	Par.(M)	EPE(trained on FCDN)	
		FCDN	VBOF
Baseline	12.22	0.87	13.28
# Causal Intervention Module			
w/o CIM	12.44	0.98	13.25
w/o CIM(+STCH)	15.77	0.87	13.01
w/o query-guided $A \odot \mathrm{V}$	12.06	0.93	13.08
w/o low-rank basis $\mathbf{B}$	12.07	0.95	13.15
w/o $\hat{\mathbb{E}}[F]$	12.81	0.85	12.96
w/o zero-init γ	13.09	0.89	12.97
w/o refinement $\sigma(\cdot)$	12.73	0.90	13.00
# Counterfactual Gating Unit			
w/o CGU	11.16	1.00	13.18
w/o CGU(+IDM)	13.65	1.03	13.39
w/o $\varphi(\cdot)$	12.68	0.89	13.14
w/o f_c	12.83	0.85	13.03
w/o spatial gate $\mathbf{G}_s$	12.61	0.92	13.12
w/o channel gate $\mathbf{G}_c$	12.70	0.89	13.10
w/o zero-init γ	13.09	0.91	12.98
w/o $\mathrm{SE}(\cdot)$	12.69	0.88	13.11
# Causal Recurrent Decoder			
w/o CRD	9.57	1.17	13.74
i=1	13.09	0.81	12.96
i=2	13.58	0.85	13.02
Causal-Flow	13.09	0.81	12.96

Table 4: Comparison of computational overhead.

Method	Par.(M)	Tim.(ms)	Mem.(GB)	Fl.(G)
FlowFormer++	16.11	122	4.3	55.6
Flow*Diffuser*	16.32	133	5.1	87.1
CEDFlow	7.78	76	2.1	32.1
CEDFlow++	**6.78**	57	1.9	26.3
FlowFM	12.22	**28**	**1.3**	**19.6**
Causal-Flow	13.09	35	1.4	21.7

4.5 Computational Analysis

Tab. 4 summarizes the parameters, runtime, memory, and FLOPs requirements of Causal-Flow alongside SOTA methods for 736×480 images. Compared with FlowFM, the causal designs introduce only 0.87 M additional parameters and 2.1 G FLOPs, while the memory footprint and runtime remain nearly identical, yet the VBOF error drops from 13.28 to 12.96. Causal-Flow is still considerably lighter than FlowFormer++ and Flow*Diffuser*.

5 Conclusion

The core insight is that the performance bottleneck stems from the entanglement of true motion and dark degradation in the observed motion evidence, which biases existing estimators toward spurious responses. Therefore, we present Causal-Flow, a causality-driven framework that resolves this entanglement in two successive steps. CIM purifies motion evidence via query-guided weighting and low-rank basis projection, while CGU estimates degradation-induced spurious motion and calibrates the flow update through causal residual gating. Embedded into each gate of the GRU-based decoder, Causal-Flow enables single-step decoding without iterative refinement. Experiments on FCDN, VBOF, and real-world dark scenes show that Causal-Flow outperforms state-of-the-art methods in both accuracy and generalization. We hope this work inspires further exploration of causal reasoning in adverse vision scenarios.

AI USE STATEMENT

In this work, we have not used generative AI tools for research ideation, method design, data analysis, or code generation, and these tasks are not applicable to this work. Additionally, we used generative AI tools for grammar correction and language polishing of the manuscript. We have reviewed all AI-assisted text, and all technical content, claims, experimental results, and conclusions were independently developed and verified by the authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Alejandro Almodóvar, Adrián Javaloy, Juan Parras, Santiago Zazo, and Isabel Valera. Decaflow: A deconfounding causal generative model. *NeurIPS*, 38:136612–136666, 2026.

Chen Chen, Qifeng Chen, Minh N Do, and Vladlen Koltun. Seeing motion in the dark. In *ICCV*, pp. 3185–3194, 2019.

Ziteng Cui, Lin Gu, Xiao Sun, Xianzheng Ma, Yu Qiao, and Tatsuya Harada. Aleth-nerf: Illumination adaptive nerf with concealing field assumption. In *AAAI*, volume 38, pp. 1435–1444, 2024.

Alexey Dosovitskiy, Philipp Fischer, Eddy Ilg, Philip Hausser, Caner Hazirbas, Vladimir Golkov, Patrick Van Der Smagt, Daniel Cremers, and Thomas Brox. Flownet: Learning optical flow with convolutional networks. In *ICCV*, pp. 2758–2766, 2015.

Weiwei Duan, Luping Ji, Shipeng Lei, Sicheng Zhu, Jianghong Huang, and Mao Ye. Chal: Causal-guided hierarchical anomaly-aware learning for moving infrared small target detection. In *CVPR*, pp. 21357–21366, 2026.

Daniel Feijoo, Juan C Benito, Alvaro Garcia, and Marcos V Conde. Darkir: Robust low-light image restoration. In *CVPR*, pp. 10879–10889, 2025.

Mingbo Hong, Feng Liu, Caroline Gevaert, George Vosselman, and Hao Cheng. Bridge: Basis-driven causal inference marries vfms for domain generalization. In *CVPR*, pp. 31962–31973, 2026.

Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *CVPR*, pp. 7132–7141, 2018.

Shihao Jiang, Dylan Campbell, Yao Lu, Hongdong Li, and Richard Hartley. Learning to estimate hidden motions with global motion aggregation. In *ICCV*, pp. 9772–9781, 2021.

Pietro Lesci, Clara Meister, Thomas Hofmann, Andreas Vlachos, and Tiago Pimentel. Causal estimation of tokenisation bias. In *ACL*, pp. 28325–28340, 2025.

Haipeng Li, Kunming Luo, and Shuaicheng Liu. Gyroflow: Gyroscope-guided unsupervised optical flow learning. In *ICCV*, pp. 12869–12878, 2021.

Shuaicheng Liu, Kunming Luo, Nianjin Ye, Chuan Wang, Jue Wang, and Bing Zeng. Oiflow: Occlusion-inpainting optical flow estimation by unsupervised learning. *IEEE TIP*, 30:6420–6433, 2021.

Ao Luo, Fan Yang, Xin Li, and Shuaicheng Liu. Learning optical flow with kernel patch attention. In *CVPR*, pp. 8906–8915, 2022.

Ao Luo, Fan Yang, Xin Li, Lang Nie, Chunyu Lin, Haoqiang Fan, and Shuaicheng Liu. Gaflow: Incorporating gaussian attention into optical flow. In *CVPR*, pp. 9642–9651, 2023.

Ao Luo, Xin Li, Fan Yang, Jiangyu Liu, Haoqiang Fan, and Shuaicheng Liu. Flowdiffuser: Advancing optical flow estimation with diffusion models. In *CVPR*, pp. 19167–19176, 2024.

Kunming Luo, Chuan Wang, Shuaicheng Liu, Haoqiang Fan, Jue Wang, and Jian Sun. Upflow: Upsampling pyramid for unsupervised optical flow learning. In *CVPR*, pp. 1045–1054, 2021.

Tommaso Mencattini, Riccardo Cadei, and Francesco Locatello. Exploratory causal inference in saence. In *ICLR*, volume 2026, pp. 136910–136957, 2026.

Henrique Morimitsu, Xiaobin Zhu, Roberto M Cesar, Xiangyang Ji, and Xu-Cheng Yin. Dpflow: Adaptive optical flow estimation with a dual-pyramid framework. In *CVPR*, pp. 17810–17820, 2025.

Judea Pearl et al. Causality: models, reasoning, and inference. *Econometric Theory*, 19(675-685): 46, 2003.

Xiaoyu Shi, Zhaoyang Huang, Dasong Li, Manyuan Zhang, Ka Chun Cheung, Simon See, Hongwei Qin, Jifeng Dai, and Hongsheng Li. Flowformer++: Masked cost volume autoencoding for pretraining optical flow estimation. In *CVPR*, pp. 1599–1610, 2023.

Yuheng Shi, Mingjia Li, Minjing Dong, and Chang Xu. Vssd: Vision mamba with non-causal state space duality. In *ICCV*, pp. 10819–10829. IEEE, 2025.

Hao Sun, Fenggen Yu, Huiyao Xu, Tao Zhang, and Changqing Zou. Ll-gaussian: Low-light scene reconstruction and enhancement via gaussian splatting for novel view synthesis. In *ACM MM*, pp. 4261–4270, 2025.

Jiacheng Tang, Zhiyuan Zhou, Zhuolin He, Jia Zhang, Kai Zhang, and Jian Pu. Causalvad: Deconfounding end-to-end autonomous driving via causal intervention. In *CVPR*, pp. 32124–32133, 2026.

Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *ECCV*, pp. 402–419. Springer, 2020.

Wenjing Wang, Xinhao Wang, Wenhan Yang, and Jiaying Liu. Unsupervised face detection in the dark. *TPAMI*, 45(1):1250–1266, 2022.

Yusong Wang, Zheyuan Gu, Keyu Mao, Minghao Shao, Mingkun Xu, Prayag Tiwari, Jiawei Shao, and Qingsong Zhao. Streaming video crime anticipation with spatio-temporal causal reasoning. In *CVPR*, pp. 16933–16943, 2026.

Haofei Xu, Jing Zhang, Jianfei Cai, Hamid Rezatofighi, and Dacheng Tao. Gmflow: Learning optical flow via global matching. In *CVPR*, pp. 8121–8130, 2022.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*, pp. 2048–2057. PMLR, 2015.

Mingfang Zhang, Yinqiang Zheng, and Feng Lu. Optical flow in the dark. *TPAMI*, 2021.

Yinqiang Zheng, Mingfang Zhang, and Feng Lu. Optical flow in the dark. In *CVPR*, pp. 6749–6757, 2020.

Fengyuan Zuo, Zhaolin Xiao, Haiyan Jin, and Haonan Su. Cedflow: Latent contour enhancement for dark optical flow estimation. In *AAAI*, volume 38, pp. 7909–7916, 2024.

Fengyuan Zuo, Haiyan Jin, Zhaolin Xiao, Haonan Su, and Meng Zhang. Cedflow++: Latent contour enhancement for dark optical flow estimation. *IJCV*, 133(10):7222–7241, 2025.

Fengyuan Zuo, Haiyan Jin, Zhaolin Xiao, and Meng Zhang. Position-surpassing flow estimator: Advanced graph-based motion reconstruction in the dark. *IEEE tcsvt*, 36(6):8859–8873, 2026a.

Fengyuan Zuo, Haiyan Jin, Yuanlin Zhang, Zhaolin Xiao, Bin Wang, and Yuerong Mu. Flowfm: Advancing dark optical flow estimation with flow matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6837–6846, 2026b.