PHOTOMETRIC FLOW FOR TEMPORAL PHOTOMETRIC CONSISTENCY

Anonymous authors
Paper under double-blind review

ABSTRACT

Flow-guided video synthesis leverages optical flow to establish temporal correspondences and model motion dynamics, while typically assuming photometric constancy—that is, corresponding pixels retain a consistent appearance as they move across frames. In practice, illumination changes, cast shadows, automatic exposure adjustments, and white-balance shifts violate this assumption, leading to color drift and visible seams after warping. Existing approaches either impose strict consistency constraints—potentially introducing artifacts and hallucinations—or learn generic error-compensation mechanisms that conflate geometric and photometric errors without explicitly modeling temporal photometric variations. In contrast, we treat photometric variation as a cue complementary to geometric information and explicitly model the evolution of pixel appearance along motion trajectories. Specifically, we introduce photometric flow, a dense, motion-aligned representation of temporal color variation, together with a lightweight Photometric Flow Network (PFN) for its estimation. PFN first predicts a stable additive base field and subsequently refines it through appearance-conditioned residual calibration. This design overcomes the limitations of purely additive correction while preserving a structured representation amenable to temporal composition. The resulting photometric flow provides explicit, temporally consistent appearance guidance that complements the motion information encoded by optical flow. Integrated into an iterative, flow-guided pipeline, our approach consistently improves color fidelity and reduces visible seams, demonstrating broad applicability across diverse tasks including video inpainting, outpainting, and frame interpolation. To the best of our knowledge, this work is the first to formulate and learn photometric flow, particularly tailored for downstream flow-guided video synthesis, providing a principled bridge between motion and appearance variation.

1 INTRODUCTION

Optical flow estimates apparent pixel motion between consecutive frames as a dense vector field, linking per-frame images into a coherent description of scene dynamics. Evolving from classical variational methods (Horn & Schunck, 1981; Zach et al., 2007) to modern deep models (Ilg et al., 2017; Sun et al., 2018; Hui et al., 2018; Yang & Ramanan, 2019; Hofinger et al., 2020; Teed & Deng, 2020), it has become a standard cue for establishing correspondences and enabling temporal propagation across video synthesis (Zhang et al., 2022; Xu et al., 2019; Koroglu et al., 2025; Liang et al., 2024; Jin et al., 2025; Zhou et al., 2023) and editing (Cong et al., 2023; Yang et al., 2023; Liu et al., 2024) pipelines.

Despite its foundational role, optical-flow-based pixel propagation implicitly assumes photometric constancy—that is, corresponding pixels maintain a consistent appearance across frames. However, real-world videos frequently violate this assumption due to illumination changes, shadows, and in-camera auto-exposure or white-balance adjustments. Consequently, even geometrically accurate warping can leave a photometric gap between the propagated content and the target frame. Meanwhile, inaccuracies in optical flow introduce spatial misalignment. These geometric and photometric errors compound over long-range propagation, resulting in color inconsistencies and visible seams after warping (Fig. 1). Therefore, motion-based correction alone is insufficient; explicitly modeling temporal photometric variation is essential yet often overlooked in existing propagation pipelines.

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.

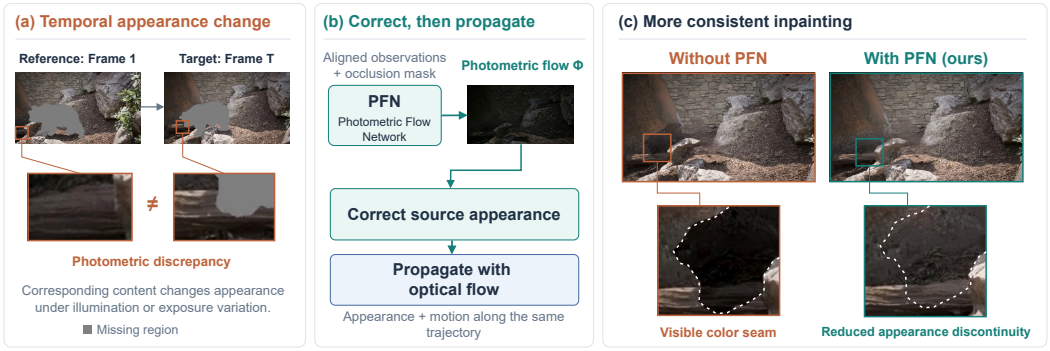


Figure 1: **Photometric flow for propagation-based inpainting.** We illustrate the challenge of photometric inconsistency using video inpainting as an example. When propagating content from a reference (Frame 1) to a target (Frame T), changes in illumination or exposure create a significant **photometric discrepancy** (orange solid box in (a)). Direct propagation ignores these changes, resulting in color seams and artifacts (indicated by the white dotted line in "without PFN" in (c)). Our proposed **Photometric Flow Network (PFN)** predicts a dense **Photometric Flow** field that explicitly represents appearance changes along motion correspondences (b). By compensating for the photometric gap, our method achieves seamless and coherent integration ("with PFN").

Existing remedies largely follow two predominant strategies. Gating-based methods (Xu et al., 2019; Zhou et al., 2023) apply strict consistency checks to discard unreliable warps, preventing severe artifacts but sacrificing temporal cues and increasing hallucination burden downstream. Compensation-based methods, exemplified by an Error Compensation Network (Kang et al., 2022), learn a generic residual to correct propagation errors. However, by treating photometric shifts and geometric misalignments as a single monolithic error—and supervising from local guidance—they can be steered toward erroneous compensation when the guidance region itself exhibits complex effects (e.g., moving shadows or exposure changes). In short, existing approaches either avoid unreliable propagation through rejection or correct artifacts implicitly with generic residuals, but temporal photometric variation itself remains largely unmodeled as an explicit variable in downstream flow-guided synthesis.

To address this limitation, we treat photometric variation as a cue complementary to geometric motion. Motivated by the distinct roles of geometry and appearance in image formation, we introduce *photometric flow*, a dense, motion-aligned representation that characterizes how pixel appearance evolves along motion trajectories. While optical flow describes motion dynamics, photometric flow explicitly models temporal appearance variation. Rather than treating photometric inconsistencies as undifferentiated propagation errors, photometric flow addresses a fundamental question: *as a pixel moves from its source location to its target location, how should its appearance change?*

We implement this idea with a lightweight Photometric Flow Network (PFN) that estimates a dense, per-pixel photometric flow. PFN first predicts an additive base field, which provides a simple and stable representation for temporal composition, and then refines it through appearance-conditioned residual calibration. We adopt an additive representation because it admits stable linear composition over long motion trajectories; in preliminary experiments, multiplicative alternatives suffered from compounding errors during long-range propagation. The calibration conditions on the local RGB appearance and the base photometric flow, relaxing the limitations of a purely additive correction while preserving the explicit and structured nature of the photometric representation. The resulting photometric flow is used directly at synthesis time to correct propagated content, complementing optical flow with appearance guidance along the same motion correspondences.

PFN can be integrated into an iterative flow-guided synthesis pipeline or used as a lightweight enhancement module for existing methods. We evaluate it on video inpainting, outpainting, and frame interpolation, and further validate the learned representation through direct photometric-alignment and optical-flow diagnostics. Across tasks and architectures, photometric flow consistently improves color fidelity, temporal coherence, and propagation quality. To our knowledge, this is the first work to explicitly formulate and learn photometric flow as a downstream representation for flow-guided video synthesis. Our main contributions are:

- We introduce *photometric flow*, a dense, motion-aligned representation of temporal appearance variation that complements optical flow by bridging geometric motion and appearance dynamics. Alongside this formulation, we develop a lightweight Photometric Flow Network (PFN) for its estimation.
- We demonstrate the broad applicability of the proposed photometric flow by successfully integrating the PFN into a unified pipeline for video inpainting, outpainting, and interpolation, achieving consistent gains in color fidelity and seam suppression across tasks.
- We design a plug-and-play Spatio-Temporal Refinement Module that leverages photometric flow as explicit guidance to improve a diverse range of state-of-the-art video synthesis methods, demonstrating its versatility as a general-purpose enhancement cue.

2 Related Work

Photometric Modeling in Optical Flow. Photometric variation remains a challenge in optical flow estimation. Modern deep estimators such as RAFT (Teed & Deng, 2020) learn features more robust to appearance changes than classical photometric matching. Asymmetric photometric augmentation further improves robustness by applying different color transformations to paired frames, but handles appearance changes implicitly through feature invariance rather than explicit modeling.

BrightFlow (Marsal et al., 2023) explicitly models photometric variation in unsupervised optical flow learning by jointly estimating flow and a dense brightness correction map. This map compensates for photometric differences in the reconstruction objective, providing a cleaner geometric learning signal. Controlled gradient coupling prevents the correction branch from trivially steering flow prediction.

Our work shares this explicit treatment but differs in objective and usage. BrightFlow uses brightness correction primarily as an auxiliary training mechanism rather than a downstream synthesis representation. In contrast, our photometric flow is a motion-aligned appearance representation that remains available at inference time and directly guides video synthesis.

Flow-Guided Video Synthesis. Flow-guided propagation is widely used in video inpainting (Zhang et al., 2022; Xu et al., 2019; Zhou et al., 2023; Kang et al., 2022; Xie et al., 2025; Li et al., 2022), outpainting (Dehan et al., 2022; Gao et al., 2020), and frame interpolation (Liu et al., 2020; Li & Cosker, 2017). Typical pipelines warp content or features across frames, handle occluded or unreliable regions using masks or confidence estimates, and refine the propagated results. Multi-step propagation extends temporal coverage but accumulates geometric misalignment and photometric discrepancies from exposure or illumination changes. Both can produce seams and temporal inconsistencies, particularly over long propagation ranges.

Existing methods reject unreliable propagation through consistency checks or gating (Xu et al., 2019; Zhou et al., 2023), or correct artifacts using learned refinement and error compensation (Kang et al., 2022). These mechanisms typically handle photometric variation as part of a broader reconstruction error. Our formulation complements them with a dedicated appearance cue aligned with optical flow correspondences, explicitly representing photometric change during synthesis.

Error Compensation vs. Photometric Flow. Error-compensation modules such as ECN (Kang et al., 2022) predict image-space residuals to improve propagated or synthesized content. Their flexibility allows joint correction of blending errors, geometric misalignment, missing details, and appearance inconsistencies. These task-oriented residuals, however, are generally neither defined along motion trajectories nor designed for temporal composition.

Photometric flow instead explicitly represents appearance changes along motion correspondences. Our PFN first estimates an additive base field for stable temporal accumulation, then applies appearance-conditioned residual calibration to increase the expressiveness of the final correction. This design preserves an explicit photometric representation while relaxing the limitations of a purely additive model. Geometry and photometry thus share correspondences but follow distinct modeling pathways, enabling photometric flow to serve as reusable guidance for downstream synthesis.

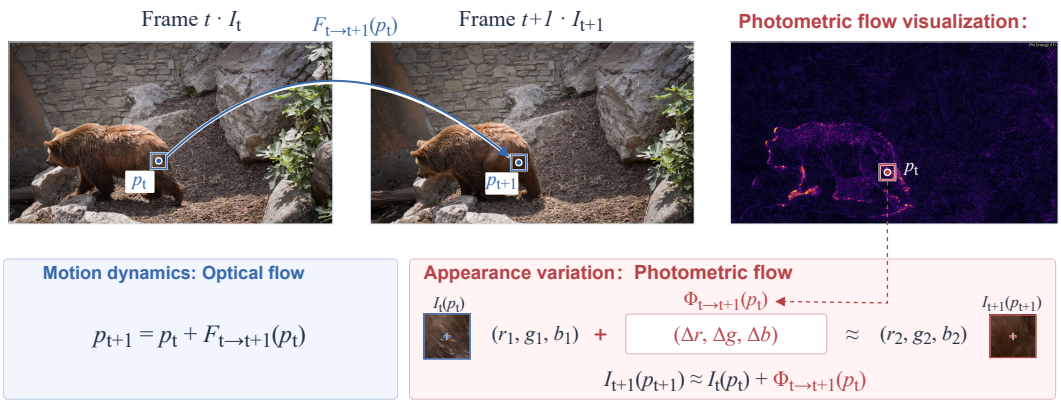


Figure 2: **Definition of Photometric Flow.** Optical flow $F_{t\to t+1}(p_t)$ establishes the geometric correspondence from p_t to p_{t+1}, while photometric flow $\Phi_{t\to t+1}(p_t)$ represents the corresponding RGB appearance change along the same motion trajectory.

3 METHOD

3.1 DEFINITION OF PHOTOMETRIC FLOW

In flow-guided video synthesis, optical flow establishes geometric correspondence across frames. For two adjacent frames I_i and I_{i+1}, the optical flow field $F_{i\to i+1}$ maps a pixel p_i to

$$p_{i+1} = p_i + F_{i\to i+1}(p_i). \tag{1}$$

Conventional flow-guided propagation implicitly assumes photometric constancy between corresponding observations. In real videos, however, illumination, shadows, exposure, and other appearance variations frequently violate this assumption.

We therefore introduce *photometric flow* $\Phi_{i\to i+1}$, a dense three-channel field that represents the appearance change associated with each motion correspondence:

$$I_{i+1}(p_{i+1}) \approx I_i(p_i) + \Phi_{i\to i+1}(p_i). \tag{2}$$

Temporal Composability. Photometric increments can be accumulated along the same motion trajectory used by optical flow. Let $p_k = p_s + F_{s\to k}(p_s)$ denote the location of a source pixel at frame k. The long-range photometric flow is

$$\Phi_{s\to t}(p_s) = \sum_{k=s}^{t-1} \Phi_{k\to k+1}(p_k). \tag{3}$$

This additive formulation enables simple and stable long-range composition. We also explored alternative multiplicative parameterizations, but repeated multiplication compounds local estimation errors and resulted in less stable long-range propagation in our preliminary experiments; additional controlled results are provided in the supplementary material. We therefore retain additive composition while using appearance-conditioned nonlinear calibration to increase the flexibility of estimating the photometric field.

3.2 MODELING PHOTOMETRIC FLOW

Base Estimator. For adjacent frames I_t and I_{t+1}, PFN takes the reference image I_t, a geometrically aligned observation I_{warped}, and an occlusion mask O_t as input, and predicts a three-channel base photometric flow $\Phi^0_{t\to t+1}$. We use a lightweight multi-scale encoder–decoder adapted from the BrightFlow corrector (Marsal et al., 2023), with learned convex upsampling following RAFT (Teed & Deng, 2020). Detailed architectural specifications are provided in the supplementary material.

Appearance-Conditioned Calibration. To adapt the base prediction to the corresponding RGB appearance, we apply a lightweight point-wise calibration function:

$$\Phi(p) = \Phi^0(p) + c_\psi\left([\Phi^0(p), I_t(p), I_{\text{warped}}(p)]\right). \tag{4}$$

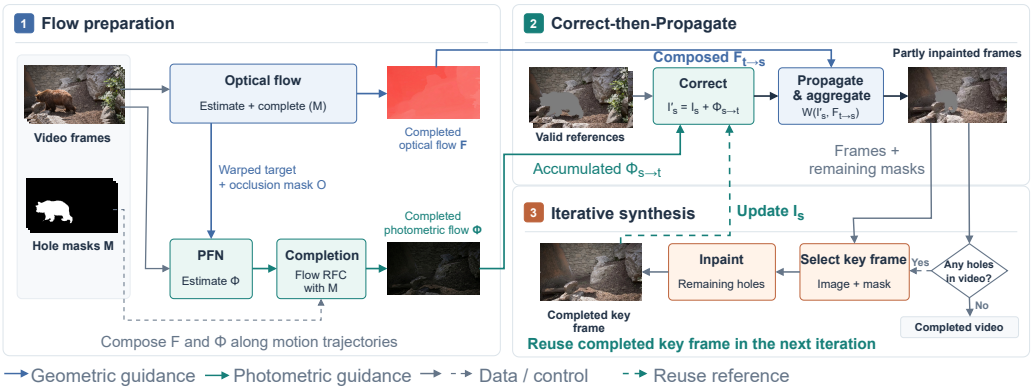


Figure 3: **Correct-then-Propagate instantiated for video inpainting.** Optical and photometric flows are prepared over missing regions. Source appearance is corrected by photometric flow before optical-flow propagation, while remaining holes are synthesized iteratively.

This preserves the additive and temporally composable representation while allowing appearance-dependent nonlinear calibration. Implementation details are provided in the supplementary material.

Synergistic Fine-tuning. We train PFN self-supervisedly using geometric correspondences provided by RAFT. Following BrightFlow (Marsal et al., 2023), we adopt a forward-sharing, backward-decoupled strategy. Given adjacent frames,

$$\begin{aligned} F_{t\to t+1} &= g_\phi(I_t, I_{t+1}), & F_{t+1\to t} &= g_\phi(I_{t+1}, I_t), \\ \bar{I}_{\text{warped}} &= \text{sg}\left(\mathcal{W}(I_{t+1}, F_{t\to t+1})\right), & \bar{O}_t &= \text{sg}(O_t), \end{aligned} \tag{5}$$

where $\text{sg}(\cdot)$ denotes stop-gradient. PFN therefore receives geometrically aligned guidance in the forward pass without propagating gradients through this guidance path to the optical flow estimator.

The target frame is reconstructed using the backward flow:

$$\hat{I}_{t+1} = \mathcal{W}\left(I_t + \Phi_{t\to t+1}, F_{t+1\to t}\right), \qquad \mathcal{L}_{\text{rec}} = \left\| I_{t+1} - \hat{I}_{t+1} \right\|_1. \tag{6}$$

Thus, the optical flow estimator receives gradients through the direct reconstruction-warping path, but not through the PFN guidance branch.

Training follows a staged strategy. We first optimize the base PFN under fixed optical-flow guidance and then jointly fine-tune the base PFN and RAFT. After attaching the appearance-conditioned calibration branch, we first optimize the new branch with the pretrained modules frozen and finally jointly fine-tune the complete PFN and RAFT under the same gradient-routing strategy. Detailed schedules are provided in the supplementary material.

3.3 Integration into Flow-Guided Video Synthesis

We instantiate Photometric Flow in video inpainting through **Correct-then-Propagate** strategy shown in Fig. 3. Optical flow determines where content should be propagated, while photometric flow describes how its appearance should change along the trajectory. Missing regions in both fields are completed before propagation; refer to the supplementary material for more details.

Correct-then-Propagate. Given a source frame I_s and target frame I_t, the accumulated photometric flow first corrects the source appearance before geometric propagation:

$$\hat{I}_{s\to t} = \mathcal{W}\left(I_s + \Phi_{s\to t}, F_{t\to s}\right). \tag{7}$$

Content propagated from multiple valid references is aggregated in the target frame. Remaining holes are synthesized in a selected key frame, which is then reused as a reference in subsequent iterations. Flow completion, aggregation, key-frame selection, and synthesis details are provided in the supplementary material.

Photometric-Guided Refinement. Photometric flow can also serve as a plug-in appearance cue for existing synthesis models. A lightweight spatio-temporal refinement network predicts residual corrections from the coarse video, validity mask, and bidirectional photometric flows while PFN remains frozen; implementation details are provided in the supplementary material.

3.4 Further Task Instantiations

The same photometric-flow representation generalizes to other flow-guided video synthesis tasks without task-specific PFN fine-tuning. For video outpainting, regions outside the observed canvas are treated as missing and the same Correct-then-Propagate strategy is applied. For frame interpolation, photometric flow corrects the two motion-aligned observations from neighboring frames before bidirectional fusion. Detailed implementations are provided in the supplementary material.

4 Experiments

4.1 Experimental Settings

Datasets: Our experiments utilize two standard, large-scale video datasets: YouTube-VOS 2018 (Xu et al., 2018) and DAVIS 2017 (Perazzi et al., 2016). For model training, we use the YouTube-VOS training set, which contains 3,471 video sequences. Following the evaluation protocol of E²FGVI and ProPainter, we evaluate on the 508-video YouTube-VOS test set and the same 50-video subset selected from the 90-video DAVIS 2017 TrainVal set.

Model Settings: We use RAFT (Teed & Deng, 2020) as the optical flow estimator. For the synthesis of missing regions that cannot be filled by propagation, we employ the Stable Diffusion v1.5 inpainting model (Rombach et al., 2022). For optical flow and photometric flow completion, we adopt the efficient Recurrent Flow Completion (RFC) network architecture from ProPainter (Zhou et al., 2023). Specifically, for optical flow completion, we directly use the available pre-trained weights provided by ProPainter. For photometric flow completion, we train our own RFC network from scratch on the YouTube-VOS dataset, following the training configuration of ProPainter.

Baselines: To isolate the contribution of photometric flow, we construct task-specific **Propagate-then-Correct** baselines, where content is first propagated using optical flow and subsequently refined by an ECN-style residual correction module. These baselines are adapted from previous works (Kang et al., 2022; Dehan et al., 2022; Liu et al., 2020) for video inpainting, outpainting, and frame interpolation, respectively. In contrast, our **Correct-then-Propagate** formulation explicitly corrects the source appearance using photometric flow before optical-flow-based propagation. We additionally use photometric flow as refinement guidance for existing synthesis models to evaluate its applicability across architectures. Unless otherwise specified, **PFN** denotes our complete Photometric Flow Network including the appearance-conditioned calibration branch. We use **PFN w/o Calibration** to denote the original base formulation without this calibration mechanism.

4.2 Quantitative Results

Video Inpainting: We evaluate PFN under two usage modes: direct integration into our propagation framework and plug-and-play refinement of existing synthesis models.

As shown in Table 1, PFN consistently improves the corresponding baselines under both propagation and refinement settings. On DAVIS, integrating PFN into the propagation framework improves ECFVI* from 31.81 to 32.15 dB in PSNR while reducing LPIPS from 0.1162 to 0.0969. As a refinement cue, PFN improves E2FGVI, ProPainter, and MiniMax Remover across all reported metrics, with ProPainter reaching 34.72 dB PSNR. On YouTube-VOS, we focus on the cross-architecture refinement setting, where PFN consistently improves E2FGVI, ProPainter, and MiniMax Remover across all reported metrics. The gains on MiniMax Remover are relatively more modest. One possible explanation is that its diffusion-based synthesis can introduce larger appearance deviations from the original observations, making lightweight residual refinement based on photometric guidance more challenging. These results demonstrate that explicit photometric guidance is effective both within flow-guided propagation and as a complementary refinement signal for diverse synthesis architectures.

Table 1: Quantitative comparison for video inpainting on DAVIS and YouTube-VOS. ECFVI* denotes our reproduced baseline. "Application" indicates whether PFN is integrated into the *propagation* stage or used as a plug-and-play *refinement* module. PFN denotes the complete model including appearance-conditioned calibration. "–" indicates settings that were not evaluated. The controlled propagation comparison with ECFVI* is reported on DAVIS, while YouTube-VOS is used for cross-architecture refinement evaluation.

Method	Application	DAVIS				YouTube-VOS			
		PSNR ↑	SSIM ↑	LPIPS ↓	VFID ↓	PSNR ↑	SSIM ↑	LPIPS ↓	VFID ↓
ECFVI (Kang et al., 2022)	-	33.52	0.9621	0.0837	0.1050	33.77	0.9703	0.0771	0.053
RGVI (Cho et al., 2025)	-	31.88	0.9423	0.0585	0.1338	31.98	0.9411	0.0681	0.076
ECFVI* (Kang et al., 2022)	-	31.81	0.9260	0.1162	0.1423	–	–	–	–
ECFVI* + **PFN**	propagation	**32.15**	**0.9558**	**0.0969**	**0.1209**	–	–	–	–
E2FGVI (Li et al., 2022)	-	33.01	0.9680	0.0460	0.1160	33.71	0.9700	0.0426	0.046
E2FGVI + **PFN**	refinement	**33.88**	**0.9756**	**0.0414**	**0.0998**	**34.31**	**0.9714**	**0.0382**	**0.039**
ProPainter (Zhou et al., 2023)	-	33.97	0.9664	0.0265	0.1010	34.23	0.9764	0.0356	0.054
ProPainter + **PFN**	refinement	**34.72**	**0.9760**	**0.0229**	**0.0963**	**34.55**	**0.9864**	**0.0318**	**0.046**
MiniMax Remover (Zi et al., 2025)	-	33.30	0.9545	0.0324	0.1048	33.26	0.9576	0.0377	0.063
MiniMax Remover + **PFN**	refinement	**34.02**	**0.9622**	**0.0315**	**0.1018**	**33.56**	**0.9653**	**0.0356**	**0.054**

Table 2: Quantitative comparison for video outpainting on DAVIS. "App." denotes the application mode: *Pro.* for propagation and *Ref.* for refinement.

Method	App.	PSNR (↑)	SSIM (↑)	LPIPS (↓)	VFID (↓)
ECFVI* (Kang et al., 2022)	–	18.25	0.7377	0.1676	0.2541
ECFVI* + **PFN**	Pro.	**20.26**	**0.7647**	**0.1438**	**0.2158**
M3DDM+ (Murakawa et al., 2026)	–	23.65	0.9610	0.1341	0.3060
M3DDM+ + **PFN**	Ref.	**24.97**	**0.9746**	**0.1199**	**0.2833**

Table 3: Quantitative comparison for video frame interpolation on DAVIS. "App." denotes the mode: *Pro.* for propagation and *Ref.* for refinement.

Method	App.	PSNR (↑)	SSIM (↑)	LPIPS (↓)	VFID (↓)
ECFVI* (Kang et al., 2022)	–	29.98	0.9106	0.0846	0.1105
ECFVI* + **PFN**	Pro.	**30.95**	**0.9233**	**0.0785**	**0.0993**
BiM-VFI (Seo et al., 2025)	–	30.63	0.9299	0.0512	0.1080
BiM-VFI + **PFN**	Ref.	**31.64**	**0.9428**	**0.0503**	**0.0936**

Video Outpainting and Frame Interpolation: We further evaluate the same propagation and refinement formulations on video outpainting and frame interpolation.

As shown in Tables 2 and 3, PFN improves both propagation-based and refinement-based baselines across the two tasks. For outpainting under a 19% overall occlusion ratio, PFN raises ECFVI* from 18.25 to 20.26 dB and M3DDM+ from 23.65 to 24.97 dB in PSNR. For interpolation, it improves ECFVI* from 29.98 to 30.95 dB and BiM-VFI from 30.63 to 31.64 dB. The consistent gains across tasks and application modes support the generality of photometric flow beyond video inpainting.

4.3 Analysis of Photometric Flow

The downstream synthesis metrics above jointly reflect geometric alignment, photometric consistency, and the behavior of the synthesis model. We therefore perform additional evaluations to examine the photometric and geometric effects of PFN.

Photometric Alignment. We evaluate photometric alignment on Sintel Final using ground-truth optical flow and measure the pixel-level L2 error between the warped source image and the target image. As shown in Table 5, the error decreases from 5.9094 with warping alone to 5.0403 with the base PFN, and further to 4.9872 with appearance-conditioned calibration. The complete model reduces the error by 15.6% relative to warping alone. All three settings use the same ground-truth flow, so the improvement comes from photometric correction rather than changes in geometric alignment.

Optical Flow Accuracy. Table 6 compares the pretrained RAFT with the model jointly fine-tuned with PFN. On Sintel Final, joint fine-tuning reduces EPE from 3.59 to 3.33 and ER from 10.77% to 9.89%. On KITTI, EPE decreases from 3.00 to 2.75 and ER from 8.48% to 8.24%. Both metrics improve on both datasets after joint fine-tuning. These results suggest that accounting for photometric

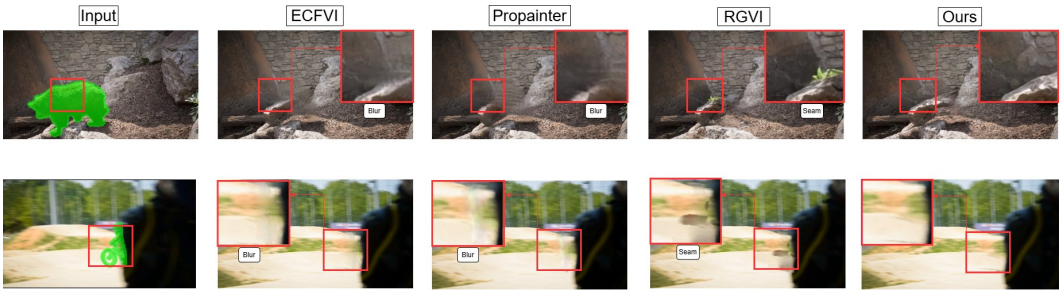


Figure 4: Qualitative comparison with representative video inpainting methods on challenging examples. From left to right: input with mask, ECFVI, ProPainter, RGVI, and our method. Our method reduces visible photometric inconsistencies and produces more coherent integration with the surrounding context.

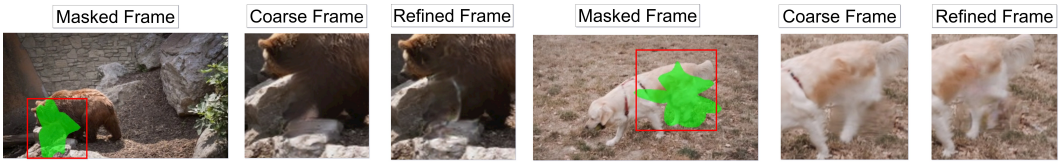


Figure 5: Photometric-flow-guided refinement of ProPainter (Zhou et al., 2023). Left to right: masked input, ProPainter result, and the refined result.

changes during joint training may help the flow estimator learn more accurate motion correspondences.

4.4 Qualitative Results

Figure 4 presents a visual comparison for video inpainting under the Correct-then-Propagate framework against representative methods including RGVI, ProPainter, and ECFVI. While competing methods may suffer from color inconsistencies, visible seams, or blurry artifacts, our method produces content that is more photometrically coherent with the surrounding context.

Figure 5 demonstrates the effectiveness of the refinement formulation. When applied to the output of ProPainter, photometric-flow guidance reduces noticeable color inconsistencies and blurring artifacts, showing that the learned representation can also be used as an external enhancement cue.

Figures 6 and 7 further illustrate the extension to other tasks. For outpainting, photometric correction suppresses visible color discontinuities around newly synthesized regions. For interpolation, the reconstruction difference maps show reduced residual appearance discrepancy after incorporating photometric flow.

4.5 Ablation Study of PFN

We conduct an ablation study to evaluate the contributions of photometric flow modeling, synergistic fine-tuning, and the newly introduced appearance-conditioned calibration. Table 4 reports the results on the DAVIS video inpainting setting.

Table 4: Ablation study of photometric flow modeling, synergistic fine-tuning, and appearance-conditioned calibration.

Setting	PSNR (↑)	SSIM (↑)	LPIPS (↓)	VFID (↓)
Baseline	31.81	0.9260	0.1162	0.1423
+ Photometric Flow	31.96	0.9416	0.1011	0.1314
+ Synergistic Training	31.83	0.9499	0.0982	0.1228
+ Appearance-Conditioned Calibration	**32.15**	**0.9558**	**0.0969**	**0.1209**

Compared with the baseline, introducing photometric flow modeling substantially improves structural and perceptual quality, increasing SSIM from 0.9260 to 0.9416 and reducing LPIPS from 0.1162 to 0.1011. This demonstrates the benefit of ex-



Figure 6: Qualitative comparison for video outpainting. The baseline without photometric flow produces color seams, while photometric correction yields more coherent transitions.

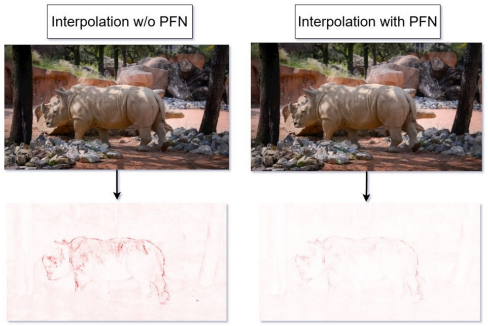


Figure 7: Qualitative comparison for frame interpolation. The top row compares interpolation results, while the bottom row visualizes L1 error, where darker red indicates larger error.

Table 5: Direct photometric alignment on Sintel Final using ground-truth optical flow.

Setting	L2 Error (↓)	Improv. (%)
Warp only	5.9094	–
PFN w/o Calibration	5.0403	14.7%
PFN	**4.9872**	**15.6%**

Table 6: Optical flow evaluation of RAFT before and after joint fine-tuning with PFN.

Dataset	RAFT		Joint FT	
	EPE (↓)	ER (↓)	EPE (↓)	ER (↓)
Sintel Final	3.59	10.77	**3.33**	**9.89**
KITTI	3.00	8.48	**2.75**	**8.24**

plicitly modeling temporal photometric variation rather than relying solely on generic propagation-error correction.

Further incorporating synergistic training improves SSIM, LPIPS, and VFID, although PSNR decreases slightly from 31.96 to 31.83 in this setting. The reduction in VFID from 0.1314 to 0.1228 is consistent with the improved optical-flow accuracy reported in Table 6, indicating that joint optimization improves the temporal and geometric consistency of the propagated content.

Finally, adding appearance-conditioned calibration yields the best overall performance across all four metrics. Compared with PFN without calibration, PSNR increases from 31.83 to 32.15 and SSIM from 0.9499 to 0.9558, while LPIPS and VFID further decrease to 0.0969 and 0.1209, respectively. These results show that explicitly recalibrating the base photometric flow according to the corresponding RGB observations provides complementary improvement beyond the base photometric representation and synergistic fine-tuning.

5 Conclusion

In this paper, we introduced Photometric Flow, a novel and interpretable representation designed to explicitly model temporal appearance variation in flow-guided video synthesis. We demonstrated that a Photometric Flow Network (PFN) can be trained in a self-supervised manner and further enhanced through lightweight appearance-conditioned calibration. Through synergistic fine-tuning, PFN explicitly accounts for photometric discrepancies while the optical flow estimator focuses on geometric correspondence under a forward-sharing, backward-decoupled optimization strategy. Our experiments show that this approach improves our primary video inpainting framework and generalizes to video outpainting and frame interpolation. Furthermore, we validated the versatility of Photometric Flow as a reusable appearance cue by using it to refine existing video synthesis methods. Direct photometric and optical-flow evaluations further support the effectiveness of the learned representation and joint-training strategy. We believe that explicitly modeling geometry and temporal appearance as complementary cues is an important step toward more robust and high-fidelity video synthesis.

AI USE STATEMENT

Generative AI tools were used to assist with language editing and polishing of the manuscript, including improving clarity, grammar, and readability. They were also used to assist with writing, editing, and debugging software code. The tools were not used to formulate the research hypotheses, design the methodology or experiments, or interpret the experimental results. All AI-assisted text was reviewed and revised by the authors, and all AI-assisted code was inspected and tested for correctness. The authors take full responsibility for the final content of the manuscript, the reported results, and all associated artifacts.

ETHICS STATEMENT

This work studies photometric consistency in flow-guided video synthesis, including video inpainting, outpainting, and frame interpolation. Our experiments are conducted on publicly available benchmark datasets, and we do not collect private user data or conduct experiments involving human participants. As with other image and video synthesis technologies, the proposed method could potentially be used to improve the realism of manipulated visual content. We therefore encourage responsible use of the method and any released artifacts in accordance with applicable laws, licenses, and ethical guidelines.

REPRODUCIBILITY STATEMENT

We provide detailed descriptions of the proposed photometric flow formulation, Photometric Flow Network (PFN), training strategy, and its integration into flow-guided video synthesis in Section 3. Experimental settings, including datasets, model configurations, baselines, and evaluation protocols, are described in Section 4.1. Additional implementation details, network architectures, training schedules, and downstream framework configurations are provided in the supplementary material, together with further ablation studies and qualitative results. These details are intended to facilitate reproduction and verification of the reported results.

REFERENCES

Suhwan Cho, Seoung Wug Oh, Sangyoun Lee, and Joon-Young Lee. Elevating flow-guided video inpainting with reference generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 2527–2535, 2025.

Yuren Cong, Mengmeng Xu, Christian Simon, Shoufa Chen, Jiawei Ren, Yanping Xie, Juan-Manuel Perez-Rua, Bodo Rosenhahn, Tao Xiang, and Sen He. Flatten: optical flow-guided attention for consistent text-to-video editing. *arXiv preprint arXiv:2310.05922*, 2023.

Loïc Dehan, Wiebe Van Ranst, Patrick Vandewalle, and Toon Goedemé. Complete and temporally consistent video outpainting. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 686–694. IEEE, 2022.

Chen Gao, Ayush Saraf, Jia-Bin Huang, and Johannes Kopf. Flow-edge guided video completion. In *European Conference on Computer Vision*, pp. 713–729. Springer, 2020.

Markus Hofinger, Samuel Rota Bulò, Lorenzo Porzi, Arno Knapitsch, Thomas Pock, and Peter Kontschieder. Improving optical flow on a pyramid level. In *European Conference on Computer Vision*, pp. 770–786. Springer, 2020.

Berthold KP Horn and Brian G Schunck. Determining optical flow. *Artificial intelligence*, 17(1-3): 185–203, 1981.

Tak-Wai Hui, Xiaoou Tang, and Chen Change Loy. Liteflownet: A lightweight convolutional neural network for optical flow estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 8981–8989, 2018.

Eddy Ilg, Nikolaus Mayer, Tonmoy Saikia, Margret Keuper, Alexey Dosovitskiy, and Thomas Brox. Flownet 2.0: Evolution of optical flow estimation with deep networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2462–2470, 2017.

Wonjoon Jin, Qi Dai, Chong Luo, Seung-Hwan Baek, and Sunghyun Cho. Flovd: Optical flow meets video diffusion model for enhanced camera-controlled video synthesis. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 2040–2049, 2025.

Jaeyeon Kang, Seoung Wug Oh, and Seon Joo Kim. Error compensation framework for flow-guided video inpainting. In *European conference on computer vision*, pp. 375–390. Springer, 2022.

Mathis Koroglu, Hugo Caselles-Dupré, Guillaume Jeanneret, and Matthieu Cord. Onlyflow: Optical flow based motion conditioning for video diffusion models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 6216–6226. IEEE, 2025.

Wenbin Li and Darren Cosker. Video interpolation using optical flow and laplacian smoothness. *Neurocomputing*, 220:236–243, 2017.

Zhen Li, Cheng-Ze Lu, Jianhua Qin, Chun-Le Guo, and Ming-Ming Cheng. Towards an end-to-end framework for flow-guided video inpainting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 17562–17571, 2022.

Feng Liang, Bichen Wu, Jialiang Wang, Licheng Yu, Kunpeng Li, Yinan Zhao, Ishan Misra, Jia-Bin Huang, Peizhao Zhang, Peter Vajda, et al. Flowvid: Taming imperfect optical flows for consistent video-to-video synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8207–8216, 2024.

Chang Liu, Rui Li, Kaidong Zhang, Yunwei Lan, and Dong Liu. Stablev2v: Stablizing shape consistency in video-to-video editing. *arXiv preprint arXiv:2411.11045*, 2024.

Xiaozhang Liu, Hui Liu, and Yuxiu Lin. Video frame interpolation via optical flow estimation with image inpainting. *International Journal of Intelligent Systems*, 35(12):2087–2102, 2020.

Rémi Marsal, Florian Chabot, Angélique Loesch, and Hichem Sahbi. Brightflow: Brightness-change-aware unsupervised learning of optical flow. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 2061–2070, 2023.

Takuya Murakawa, Takumi Fukuzawa, Ning Ding, and Toru Tamaki. M3ddm+: An improved video outpainting by a modified masking strategy. *arXiv preprint arXiv:2601.11048*, 2026.

Federico Perazzi, Jordi Pont-Tuset, Brian McWilliams, Luc Van Gool, Markus Gross, and Alexander Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 724–732, 2016.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Wonyong Seo, Jihyong Oh, and Munchurl Kim. Bim-vfi: Bidirectional motion field-guided frame interpolation for video with non-uniform motions. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 7244–7253, 2025.

Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz. Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 8934–8943, 2018.

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

Chaohao Xie, Kai Han, and Kwan-Yee K Wong. Vipdiff: Towards coherent and diverse video inpainting via training-free denoising diffusion models. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 2411–2420. IEEE, 2025.

Ning Xu, Linjie Yang, Yuchen Fan, Jianchao Yang, Dingcheng Yue, Yuchen Liang, Brian Price, Scott Cohen, and Thomas Huang. Youtube-vos: Sequence-to-sequence video object segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 585–601, 2018.

Rui Xu, Xiaoxiao Li, Bolei Zhou, and Chen Change Loy. Deep flow-guided video inpainting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3723–3732, 2019.

Gengshan Yang and Deva Ramanan. Volumetric correspondence networks for optical flow. *Advances in neural information processing systems*, 32, 2019.

Shuai Yang, Yifan Zhou, Ziwei Liu, and Chen Change Loy. Rerender a video: Zero-shot text-guided video-to-video translation. In *SIGGRAPH Asia 2023 Conference Papers*, pp. 1–11, 2023.

Christopher Zach, Thomas Pock, and Horst Bischof. A duality based approach for realtime tv-l 1 optical flow. In *Joint pattern recognition symposium*, pp. 214–223. Springer, 2007.

Kaidong Zhang, Jingjing Fu, and Dong Liu. Flow-guided transformer for video inpainting. In *European conference on computer vision*, pp. 74–90. Springer, 2022.

Shangchen Zhou, Chongyi Li, Kelvin CK Chan, and Chen Change Loy. Propainter: Improving propagation and transformer for video inpainting. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10477–10486, 2023.

Bojia Zi, Weixuan Peng, Xianbiao Qi, Jianan Wang, Shihao Zhao, Rong Xiao, and Kam-Fai Wong. Minimax-remover: Taming bad noise helps video object removal. *arXiv preprint arXiv:2505.24873*, 2025.

In the supplementary material, we provide additional implementation and training details, together with extended experimental analyses that complement the main manuscript. Specifically, Sec. A elaborates on the network architectures, training protocols, and downstream framework implementations, including the Photometric Flow Network (PFN), its appearance-conditioned calibration branch, the Recurrent Flow Completion (RFC) network, and the Spatio-Temporal Refinement Module (STRM). We also provide further details on the forward-sharing, backward-decoupled training strategy and the baseline and photometric-guided refinement frameworks. In Sec. B, we present additional ablation studies and photometric-flow visualizations to further analyze the behavior and effectiveness of our method. Additional qualitative comparisons are provided in the accompanying video.

A Implementation and Training Details

We provide implementation details for the network architectures, training protocols, and downstream framework components used in our method.

A.1 Network Architectures

We describe the three learnable components used in our system: the Photometric Flow Network (PFN), the Recurrent Flow Completion (RFC) network, and the Spatio-Temporal Refinement Module (STRM).

1. Photometric Flow Network (PFN)

The PFN (f_θ) is adapted from the Corrector architecture in BrightFlow (Marsal et al., 2023). PFN consists of a lightweight encoder–decoder for estimating a base photometric flow, followed by an appearance-conditioned calibration branch that produces the final prediction.

Input/Output: PFN takes a 7-channel input tensor formed by concatenating the reference/source image I_t (3 channels), the target observation warped back to the coordinate system of I_t, I_{warped} (3 channels), and the occlusion mask O_t (1 channel).The encoder–decoder first predicts a 3-channel base photometric flow $\Phi^0_{t\to t+1}$. An appearance-conditioned calibration branch subsequently refines this base prediction to obtain the final 3-channel photometric flow $\Phi_{t\to t+1}$.

Encoder: The encoder extracts multi-scale features through three stages of residual downsampling:

- *Input Processing:* A 7×7 convolution with stride 2 and padding 3 maps the 7-channel input to 64 channels.
- *Core Structure:* Three stages are used, each containing two residual blocks. Stage 1 operates at stride 1 with 64 channels at $H/2$ resolution; Stage 2 operates at stride 2 with 96 channels at $H/4$ resolution; and Stage 3 operates at stride 2 with 128 channels at $H/8$ resolution.
- *Output:* The encoder produces a feature map $F \in \mathbb{R}^{128\times H/8\times W/8}$.

Base Photometric Flow Decoder: The decoder predicts the base photometric flow using a learned convex upsampling mechanism following RAFT (Teed & Deng, 2020). It contains two heads operating on the $1/8$-resolution feature F:

- *Offset Head:* Predicts a coarse 3-channel photometric field through Conv $\to$ ReLU $\to$ Conv $\to$ Tanh.
- *Mask Head:* Predicts per-pixel convex upsampling weights ($9\times 8\times 8 = 576$ channels) through Conv $\to$ ReLU $\to$ Conv.
- *Upsampling:* The coarse field is reconstructed at full resolution through a convex combination of neighboring coarse predictions, yielding the base photometric flow $\Phi^0 \in \mathbb{R}^{3\times H\times W}$.

Appearance-Conditioned Calibration: To explicitly recalibrate the base prediction according to the corresponding RGB observations, we concatenate the base photometric flow with the target and warped images:

$$X_{\text{cal}} = [\Phi^0, I_t, I_{\text{warped}}] \in \mathbb{R}^{9\times H\times W}. \tag{8}$$

A shared point-wise MLP, implemented using two 1×1 convolutions with channel dimensions $9 \to 32 \to 3$, predicts a residual photometric correction:

$$\Delta\Phi = W_2\, \sigma(W_1 X_{\text{cal}} + b_1) + b_2, \tag{9}$$

where σ denotes the nonlinear activation function. The final photometric flow is obtained through residual calibration:

$$\Phi = \Phi^0 + \Delta\Phi. \tag{10}$$

Since the calibration is implemented entirely with 1×1 convolutions, the same nonlinear mapping is applied independently at each spatial location while jointly mixing the RGB and photometric-flow channels. The spatial context is provided by the preceding encoder–decoder through Φ^0.

2. RECURRENT FLOW COMPLETION NETWORK (RFC)

For completing both optical flow and photometric flow, we adopt the Recurrent Flow Completion (RFC) architecture from ProPainter (Zhou et al., 2023).

Input/Output: The RFC takes a temporal tensor formed by concatenating the raw field (Φ_{raw} or F_{raw}) with its corresponding binary mask M. The network outputs a dense completed optical-flow or photometric-flow field over the video sequence.

Architecture: The network follows the spatio-temporal encoder–propagation–decoder design of ProPainter and leverages Pseudo-3D (P3D) blocks and deformable alignment:

- *Encoder:* Downsampling stages based on P3D blocks extract spatio-temporal features and reduce the spatial resolution to $1/8$.
- *Recurrent Propagation:* A bidirectional propagation module with second-order deformable alignment aggregates temporally neighboring features and propagates information into missing regions.
- *Decoder:* The fused features are progressively upsampled using bilinear interpolation followed by convolutions, with encoder skip connections used to preserve spatial details.

3. SPATIO-TEMPORAL REFINEMENT MODULE (STRM)

The STRM is a lightweight pluggable refinement network adapted from the ECN architecture.

Input: For a temporal window of length $L = 10$, the input tensor has shape $L \times 10 \times H \times W$. For each frame I_t, the 10-channel input consists of the coarse frame $I_{\text{coarse},t}$ (3 channels), the bidirectional photometric flows $\Phi_{t\to t+1}$ and $\Phi_{t\to t-1}$ (6 channels), and the validity mask M_t (1 channel).

Architecture: STRM adopts a U-Net-like encoder–decoder with a Transformer bottleneck:

- *Encoder/Decoder:* The encoder downsamples the input to $1/4$ resolution with channel dimensions $10 \to 64 \to 128 \to 320$. The decoder restores the original resolution using bilinear upsampling followed by gated convolutions.
- *Transformer Bottleneck:* The bottleneck contains four Transformer blocks with multi-head attention and five multi-scale patch sizes for modeling longer-range spatio-temporal context.

Output: The network predicts a residual correction clip V_{residual}, which is added to the coarse input clip:

$$V_{\text{refined}} = V_{\text{coarse}} + V_{\text{residual}}. \tag{11}$$

A.2 TRAINING PROTOCOL

All models are trained on the **YouTube-VOS** training set using the **AdamW** optimizer on four NVIDIA RTX 3090 GPUs. Our training follows a staged curriculum:

Stage 1: PFN Training. We first train the base photometric-flow estimator, i.e., PFN without the appearance-conditioned calibration branch. For the initial **80k steps**, we use a **frozen** pre-trained RAFT to provide geometric guidance and optimize only the base PFN with a learning rate of 1×10^{-4} and a batch size of 32.

We then **unfreeze** RAFT and jointly fine-tune the base PFN and the optical flow estimator for an additional **40k steps** with a reduced learning rate of 1×10^{-5}. During this stage, we adopt the forward-sharing, backward-decoupled optimization strategy described in the main paper. Specifically, the warped image and occlusion mask derived from the predicted optical flow are detached before being fed into PFN:

$$\bar{I}_{\text{warped}} = \text{sg}(I_{\text{warped}}), \qquad \bar{O}_t = \text{sg}(O_t), \tag{12}$$

where $\text{sg}(\cdot)$ denotes stop-gradient. Thus, flow-derived guidance is shared with PFN in the forward pass, while gradients through the PFN guidance branch do not propagate back to RAFT. The optical flow estimator is still updated through the direct reconstruction/warping path.

After obtaining the synergistically fine-tuned base PFN, we attach the appearance-conditioned calibration branch to form the complete PFN. We first freeze both the base PFN and RAFT and optimize only the calibration branch for **5k steps**. We then unfreeze the base PFN and RAFT and jointly fine-tune the complete PFN together with the optical flow estimator for an additional **40k steps**. During this stage, we retain the same forward-sharing, backward-decoupled optimization strategy: flow-derived guidance is detached before entering PFN, while the optical flow estimator receives gradients through the direct reconstruction-warping path.

Stage 2: Flow Completion Network Training. The RFC network for completing photometric flow is trained for **200k iterations**, following the training configuration of ProPainter (Zhou et al., 2023).

- *Self-Supervised Data Generation:* We generate dense photometric flows using the trained PFN on unmasked video sequences to serve as pseudo-ground truth. We then simulate corruption by applying random stationary and moving masks.
- *Optimization:* The network is optimized using an L_1 reconstruction loss on both valid and masked regions, together with a second-order smoothness loss. We use a batch size of 4 and an initial learning rate of 1×10^{-4}.

Stage 3: Refinement Module (STRM) Training. The STRM is trained for **200k iterations**. Following the strategy of ECFVI, the pre-trained PFN and the upstream optical- and photometric-flow estimation/completion modules are kept **frozen** during this stage. This ensures that the refinement module focuses solely on correcting residual errors from the coarse video and the fixed photometric-flow guidance. We use an initial learning rate of 1×10^{-4}, an L_1 reconstruction loss, and a batch size of 4.

A.3 Detailed Frameworks

In this section, we provide additional implementation details for the baseline framework and the Photometric-Guided Refinement Framework summarized in the main paper.

1. Baseline Framework

To isolate the effect of explicitly modeling photometric flow, we construct a baseline framework that removes PFN while retaining a generic error correction mechanism. This baseline follows a conventional "Propagate-then-Correct" strategy, in contrast to our "Correct-then-Propagate" formulation.

Pipeline: As illustrated in Fig. 8, the baseline operates in the following sequence:

1. **Flow Completion:** As in our main framework, optical flow fields are first estimated and completed within masked regions using the RFC network.
2. **Direct Content Propagation:** Content from reference frames is directly warped to the target frame using the completed optical flow, without photometric correction before propagation. Temporal appearance differences may therefore remain in the propagated content.

3. **Error Compensation:** A generic Error Compensation Network (ECN) is applied to the propagated result and predicts a residual correction map. Unlike our method, this correction is performed *after* geometric propagation.
4. **Synthesis:** The corrected propagated result is subsequently passed to the synthesis module (Stable Diffusion) to fill regions that remain unavailable from propagation.

Key Difference: The main distinction lies in when and how the correction is introduced. The baseline performs generic residual compensation after propagation, whereas our method estimates an explicit motion-aligned photometric field and corrects the source appearance before geometric propagation. This comparison allows us to evaluate the benefit of explicitly modeling temporal photometric variation rather than correcting propagation errors only after they occur.

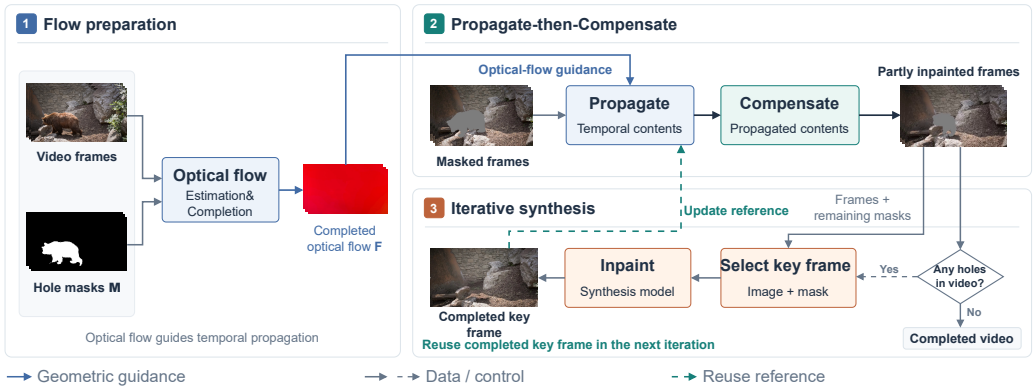


Figure 8: **Baseline framework.** The baseline follows a Propagate-then-Correct pipeline, where content is first propagated using optical flow and a generic residual correction is subsequently applied.

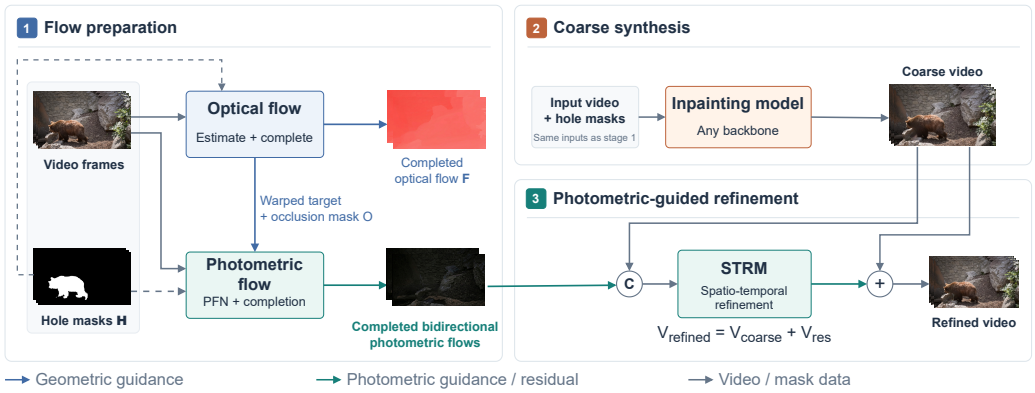


Figure 9: **Photometric-Guided Refinement Framework.** Bidirectional photometric flows extracted from the input video provide appearance guidance to STRM, which predicts a residual correction for the coarse video produced by an external synthesis model.

2. Photometric-Guided Refinement Framework

As shown in Fig. 9, we construct a lightweight refinement framework that uses photometric flow as an explicit guidance signal for the output of an existing video synthesis method. The framework operates in a "Guidance-then-Correction" manner and does not require modification or re-training of the upstream synthesis backbone.

Pipeline Stages: The framework consists of three sequential stages:

1. **Coarse Input Acquisition:** The framework receives a coarse video sequence $\mathbf{V}_{\text{coarse}}$ generated by an upstream synthesis method (e.g., ProPainter or E2FGVI), together with the corresponding masks.
2. **Photometric Guidance Generation:** Photometric guidance is extracted from the original input sequence rather than from the potentially degraded coarse prediction. We first estimate optical flow between adjacent frames and complete the missing flow regions using the pre-trained RFC network. The frozen PFN then estimates bidirectional photometric flows from the corresponding observations and geometric guidance. Missing regions in the photometric-flow fields are completed using the photometric-flow RFC, following the same flow-preparation procedure as in our main propagation framework. The resulting bidirectional photometric flows provide explicit appearance guidance for subsequent refinement.
3. **Spatio-Temporal Residual Correction:** The coarse video, masks, and bidirectional photometric-flow guidance are stacked and fed into STRM. The module processes the sequence in a sliding-window manner and predicts a residual correction $\mathbf{V}_{\text{res}}$. The final refined video is obtained as

$$\mathbf{V}_{\text{refined}} = \mathbf{V}_{\text{coarse}} + \mathbf{V}_{\text{res}}. \tag{13}$$

Model-Agnostic Design: The refinement framework treats the upstream synthesis method as an external backbone and uses only its coarse output together with photometric-flow guidance. Therefore, the refinement module can be applied to different synthesis architectures without re-training or modifying the upstream backbone.

B Additional Results

In this section, we provide additional experimental analyses and visual demonstrations to complement the main manuscript. We first examine the representation choice underlying Photometric Flow by comparing additive and multiplicative parameterizations under a controlled reduced-training setting, providing empirical support for the additive formulation adopted in the main method. We then investigate the impact of different color spaces on PFN training. Finally, we visualize the learned Photometric Flow to further analyze its spatial behavior.

B.1 Ablation Study: Additive vs. Multiplicative Parameterization

To further examine the choice of photometric-flow representation, we compare our additive formulation with a multiplicative alternative under a controlled, shortened training setting. The additive formulation models temporal appearance variation as

$$I_{i+1}(p_{i+1}) \approx I_i(p_i) + \Phi_{i\to i+1}(p_i), \tag{14}$$

whereas the multiplicative variant predicts a per-channel scaling field $A_{i\to i+1}$:

$$I_{i+1}(p_{i+1}) \approx I_i(p_i) \odot A_{i\to i+1}(p_i). \tag{15}$$

For long-range propagation, the multiplicative formulation requires repeated products of intermediate scaling fields, while the additive formulation uses simple accumulation. To isolate this representation choice, we train both variants using the same reduced training schedule and evaluate them under the same propagation-based video inpainting setting on DAVIS. All other components and training configurations are kept unchanged.

As shown in Table 7, the multiplicative parameterization performs substantially worse than the additive formulation under the same propagation-based evaluation setting. In particular, its PSNR drops to 23.67 dB, compared with 31.98 dB for the additive formulation, while SSIM decreases from 0.9379 to 0.7667 and both LPIPS and VFID deteriorate substantially. Notably, the multiplicative variant also performs considerably worse than the propagation baseline (31.81 dB), whereas the additive formulation remains stable and slightly improves upon it.

This behavior is consistent with the different temporal composition properties of the two representations. Additive photometric flow is accumulated through summation along a motion trajectory,

Table 7: Comparison of additive and multiplicative photometric parameterizations under a reduced training schedule. The two PFN variants use the same training budget and are evaluated under the same propagation-based protocol.

Setting	PSNR (↑)	SSIM (↑)	LPIPS (↓)	VFID (↓)
Propagation Baseline	31.81	0.9260	0.1162	0.1423
Multiplicative Photometric Flow	23.67	0.7667	0.1598	0.2341
Additive Photometric Flow (Ours)	**31.98**	**0.9379**	**0.1087**	**0.1388**

whereas multiplicative scaling fields must be repeatedly multiplied. Consequently, local estimation errors in the multiplicative field can be repeatedly compounded during temporal propagation, making the resulting correction increasingly sensitive to prediction errors as the propagation range grows. The pronounced degradation observed in this controlled experiment is consistent with the instability we encountered during our preliminary investigation and motivates the additive representation adopted in the main model.

B.2 Ablation Study: Color Space Selection

Intuitively, color spaces that decouple luminance from chrominance, such as HSL (Hue, Saturation, Lightness), might seem more suitable for modeling photometric changes and photometric inconsistencies. To determine the optimal representation for our Photometric Flow Network (PFN), we conducted an ablation study comparing models trained in the RGB space versus the HSL space. For a fair comparison, we employed a shortened training schedule for this ablation: the PFN was pre-trained for 20k steps, followed by 40k steps of training with the RFC network for photometric flow on the video inpainting task. We evaluate the results on DAVIS test sets like what we do in the main paper.

Table 8: Ablation study on the training color space. Models were trained with a reduced schedule (20k PFN + 40k RFC steps).

Setting	PSNR (↑)	SSIM (↑)	LPIPS (↓)
Baseline (w/o ECN)	22.38	0.7648	0.1896
Ours (HSL Space)	27.65	0.7953	0.1667
Ours (RGB Space)	**30.76**	**0.8832**	**0.1309**

As shown in Table 8, while the HSL variant improves upon the baseline, the model trained in the **RGB space** demonstrates significantly superior performance across all metrics (e.g., +3.11 dB in PSNR over HSL). We attribute this primarily to **Feature Space Compatibility.**

Since the other core components of our pipeline—specifically the pre-trained optical flow estimator (RAFT) and the downstream synthesis modules—are designed and trained to operate on RGB representations, they extract features optimized for this color space. Forcing the PFN to operate in HSL introduces a representation gap, requiring the network to implicitly bridge the discrepancy between RGB-based geometric guidance and HSL-based photometric correction, which likely hinders optimization.

Limitation: We acknowledge that determining whether HSL could theoretically outperform RGB if the entire pipeline (including the optical flow backbone) were retrained in HSL space remains an open question. However, given the computational cost and the dominance of RGB in pre-trained video models, we adopt RGB as the most practical and effective choice for our framework. Consequently, we adopt RGB as the default color space for all subsequent experiments.

B.3 Additional Photometric Flow Visualizations

Photometric changes between adjacent frames are often subtle, making direct visualization of the three-channel photometric-flow field difficult to interpret. We therefore visualize the spatial magnitude of the predicted correction using a **Photometric Flow Magnitude Map**.

Visualization Protocol. For a predicted photometric flow $\Phi \in \mathbb{R}^{3\times H\times W}$, we compute the pixel-wise L_1 magnitude as

$$M(x, y) = \sum_{c\in\{R,G,B\}} |\Phi_c(x, y)|. \tag{16}$$

For visualization, the resulting magnitude map is mapped to a sequential pseudo-color scale, where darker regions correspond to smaller predicted corrections and brighter regions indicate larger corrections.

Analysis. As shown in Figs. 10 and 11, the predicted photometric corrections are spatially non-uniform. Large portions of the frames receive only small adjustments, while stronger responses appear in regions with more pronounced local appearance variation. Strong responses are also frequently observed near motion boundaries and complex contours, where the prediction may be influenced by both local appearance changes and residual uncertainty in geometric correspondence. These visualizations therefore show that photometric flow captures spatially structured correction patterns rather than a uniform global color shift.

C Limitations

Photometric Flow retains an additive representation to enable simple and stable temporal composition. Although appearance-conditioned calibration increases the flexibility of estimating this field, the representation does not model arbitrary nonlinear radiometric transformations, and may remain challenged by strong specularities, saturation or overexposure, and nonlinear camera responses. Moreover, PFN relies on optical-flow correspondences; severe correspondence failures may contaminate photometric estimation and lead to inaccurate color correction or smearing. Extending Photometric Flow to uncertainty-aware or more general radiometric transformations is an interesting direction for future work.



Figure 10: **Visualization of photometric flow.** The magnitude map visualizes the spatial distribution of the predicted photometric correction, with darker regions indicating smaller corrections and brighter regions indicating larger corrections.

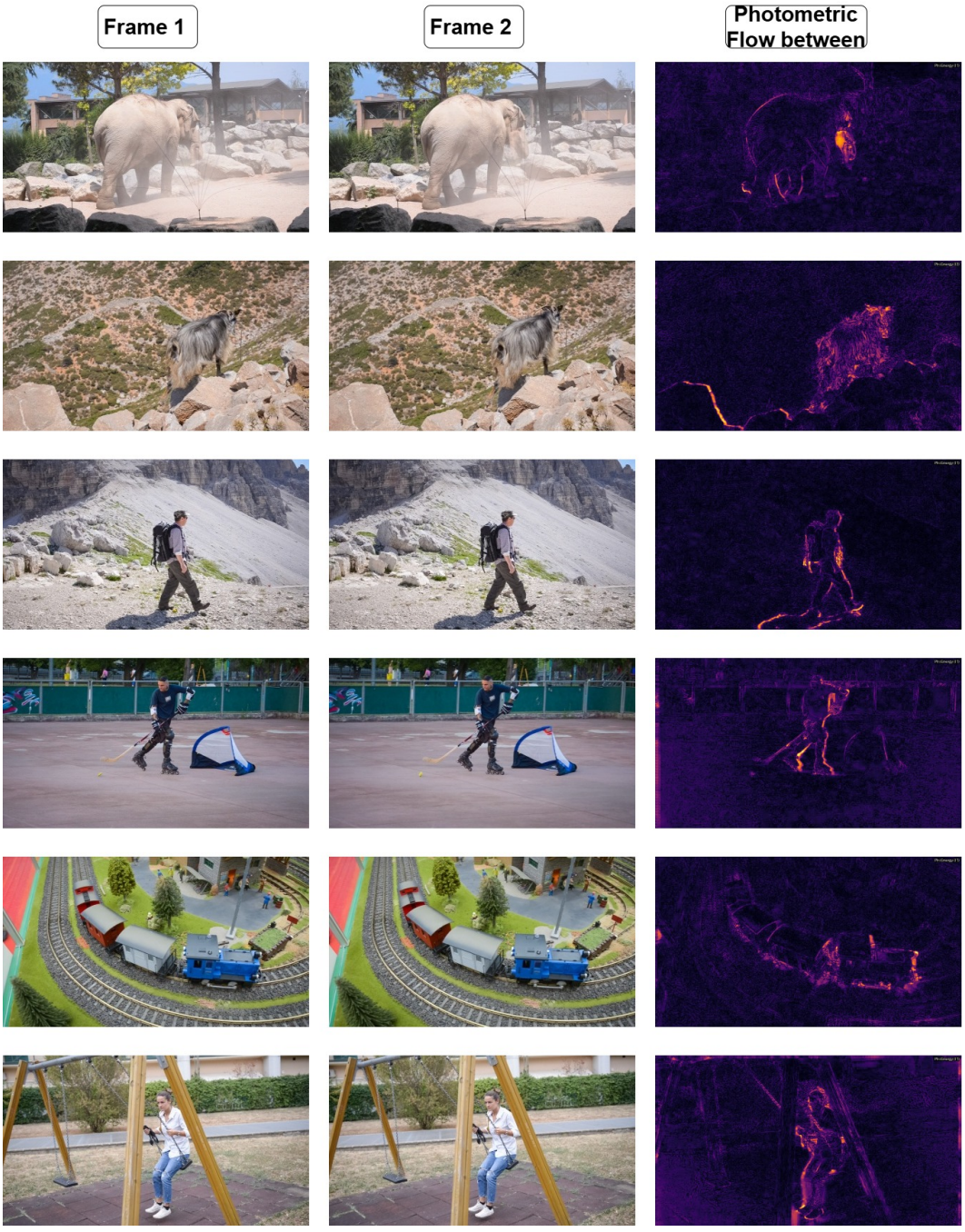


Figure 11: **Additional photometric-flow visualizations.** The predicted corrections exhibit spatially varying magnitudes across different objects, regions, and motion configurations.