UbiEnlight: User-Perception-Aware Real-Time Low-Light Video Enhancement on Mobile Devices and Smartglasses

MINFAN WANG, Northwestern Polytechnical University, China
SICONG LIU, Northwestern Polytechnical University, China
TENG LI, Northwestern Polytechnical University, China
ZIMU ZHOU, City University of Hong Kong, China
SIXUN HE, Northwestern Polytechnical University, China
BIN GUO, Northwestern Polytechnical University, China
JUNZHAO DU, Xidian University, China
ZHIWEN YU, Harbin Engineering University, China

Low-light video capture is now common on smartphones and wearables, yet dynamic illumination coupled with mobile/werable user motion causes noise, blur, and flicker that make videos *hard to perceive and use*, especially for users with night blindness and near-eye displays. Prior work enhances frames with post-hoc temporal constraints or optimizes videos directly; however, low-light and motion degradations are *spatiotemporally coupled*, so post-hoc consistency after per-frame restoration often leaves flicker and motion artifacts. We present UbiEnlight, the first diffusion-transformer-based *on-device* system for *real-time, temporally stable* low-light *video* enhancement. UbiEnlight builds on two insights: (1) illumination and structure separate more robustly in the *frequency* domain, enabling a spectrum-guided diffusion transformer that injects Fourier amplitude/phase priors to correct illumination while anchoring structure without optical flow; and (2) user perceptual tolerance varies by context, motivating a perception-aware controller that adapts sampling depth, attention reuse, and frame caching at runtime. We implement UbiEnlight on smartphones and smartglasses, and evaluate it against state-of-the-art baselines. UbiEnlight improves temporal stability by up to 54.76% under dynamic/extreme conditions while sustaining real-time on-device performance.

CCS Concepts: • **Human-centered computing** → **Ubiquitous and mobile computing**; • **Computing methodologies** → **Machine learning**.

ACM Reference Format:
Minfan wang, Sicong Liu, Teng Li, Zimu Zhou, Sixun He, Bin Guo, Junzhao Du, and Zhiwen Yu. 2026. UbiEnlight: User-Perception-Aware Real-Time Low-Light Video Enhancement on Mobile Devices and Smartglasses. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 10, 3, Article 161 (September 2026), 31 pages. https://doi.org/10.1145/3832024

1 INTRODUCTION

Ubiquitous visual sensing on smartphones and wearables has become a primary interface for *non-expert users* to perceive, record, and interact with the physical world at scale, and is especially important for users with *night*

Corresponding author: scliu@nwpu.edu.cn.
Authors' addresses: Minfan wang, Northwestern Polytechnical University, School of Computer Science, Xi'an, China; Sicong Liu, Northwestern Polytechnical University, School of Computer Science, Xi'an, China ; Teng Li, Northwestern Polytechnical University, School of Computer Science, Xi'an, China; Zimu Zhou, City University of Hong Kong, School of Data Science, Hong Kong, China; Sixun He, Northwestern Polytechnical University, School of Computer Science, Xi'an, China; Bin Guo, Northwestern Polytechnical University, School of Computer Science, Xi'an, China; Junzhao Du, Xidian University, School of Computer Science and Technology, Xi'an, China; Zhiwen Yu, Harbin Engineering University, Harbin, China.

2474-9567/2026/9-ART161
https://doi.org/10.1145/3832024

blindness (affecting roughly 1 in 4,000 worldwide), for whom low-light enhancement on phones and smart glasses can serve as a practical assistive layer beyond general-purpose capture. Video now underpins mass-market uses spanning communication, safety monitoring, assistive perception, and immersive AR/VR (*e.g.*, Google Meet [1], Microsoft Teams [24], Google Clips [2], Tesla Sentry [56], Apple Magnifier [43], Microsoft Seeing AI [15], Google ARCore [33], Meta [16]), across a global smartphone base exceeding 4.3B owners.

Yet, video is routinely captured under uncontrolled lighting, *e.g.*, night commuting, dim indoors, or restricted-illumination venues, making *low light* a pervasive challenge on mobile and wearable devices. This challenge goes beyond aesthetics: Beyond visual quality, degraded visibility and temporal instability reduce usability in safety-critical scenarios such as night driving, elderly care, and home monitoring. Our user study with 53 participants (aged 21∼68, Sec. 2.1) further confirms that users perceive low-light degradation as a threat to safety, reliability, and usability, motivating a responsive and temporally stable on-device enhancement solution.

The key challenge is that real-world videos captured by non-expert users exhibit *non-stationary, mixed degradations*. To increase brightness, mobile cameras often extend exposure time through auto-exposure and ISP adaptation, which inevitably introduces motion blur from camera shake and object motion while leaving residual noise, low contrast, and color distortion. Existing enhancement methods [34] mainly improve brightness and denoising but overlook motion-induced structural degradation; deblurring methods [22] assume well-lit inputs; and existing joint or cascaded approaches [7, 64, 78] remain brittle and largely rely on synthetic paired data.

Existing low-light video enhancement includes *frame-centric pipelines* that add temporal terms as an post-hoc auxiliary constraint, and *video-native* approaches that optimize spatiotemporal restoration end-to-end. Our key insight is that low-light degradation and motion/illumination changes are *spatiotemporally coupled*, so enforcing consistency *after* per-frame restoration often leaves residual flicker and motion artifacts. Accordingly, video-native methods have evolved from i) paired dynamic-video supervision [29, 60], to ii) stronger spatiotemporal guidance for harder motion/lighting [13], and further to iii) unpaired/generative learning [81]. In parallel, diffusion is a promising generative branch. Yet they have a core mismatch: image-oriented enhancers [28, 46, 62, 71] process frames independently and flicker, while video-generation diffusions [21, 47, 79] favor plausibility and can drift or hallucinate structure when repurposed. Specifically, two gaps remain:

• **Challenge #1: Long-stream stability under mixed illumination-motion shifts.** In mobile/wearable capture, illumination can change abruptly (often into extreme low light) while user motion induces fast blur and vibration (especially under increased exposure time). Thus, we must suppress flicker/exposure oscillation *and* prevent structural/semantic drift. Yet most frame-centric pipelines, even with flow warping or recurrent consistency [6, 29, 49, 72], still depend on correspondence estimation, which degrades under low SNR and blur, yielding artifacts or over-smoothing. Moreover, temporal stabilization typically adds extra passes (flow/warping/refinement), raising latency and compute, a major cost for on-device, long-stream deployment. The bottleneck is *preserving coherent illumination and structure when alignment is unreliable*, under tight budgets.

• **Challenge #2: Perception-constrained, resource-aware execution.** Mobile and wearable users are highly sensitive to *delay* and *temporal artifacts*, yet what they notice and value is *context-dependent*. For example, during driving/riding/running, *temporal continuity* and low latency often matter more than marginal per-frame fidelity; on near-eye smartglasses, even subtle jitter or exposure drift becomes salient; in static inspection (*e.g.*, reading signs), fine detail and color fidelity dominate. Ubiquitous enhancement therefore demands *runtime* trade-offs that align *compute* (latency/power/memory) with *human-perceptible* gains, rather than a fixed per-frame policy. However, most backbones run with near-constant per-frame cost [61, 63, 71], and existing adaptivity often tunes a knob without explicitly using perceptual thresholds or real-time resource feedback [9, 40], leading to wasted compute or unstable outputs under tight budgets. The bottleneck is *closing the loop* between perceptual constraints and runtime budgets.

To address these gaps, we present UbiEnlight, a user-centric low-light video enhancement system with two integrated modules, enabling real-time and temporally stable enhancement on mobile and wearables.

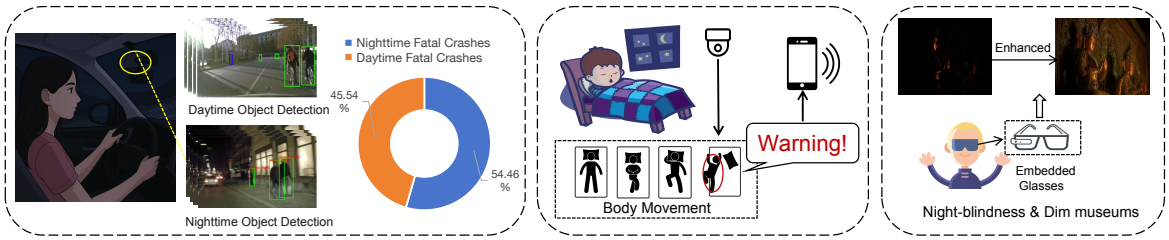


Fig. 1. Representative use cases of low-light video enhancement.

• **First, spectrum-guided diffusion transformer.** Our key idea is to *anchor* diffusion-based low-light *video* enhancement with frequency-domain priors that remain stable under lighting changes, avoiding brittle pixel-domain decomposition and correspondence-based alignment. Motivated by dynamic and extreme low light where Retinex-style decomposition becomes ill-posed, we shift to the *frequency domain*. Using Fourier analysis, we treat amplitude as an illumination *prior* and phase as a structural-semantic *anchor*, and inject both as *conditioning* tokens into a *latent* diffusion transformer throughout denoising. This converts diffusion into *prior-guided* enhancement: amplitude steers exposure correction while phase constrains structure to prevent drift. We further stabilize long streams via spectrum-gated temporal reuse, correlated noise, and phase-aware regularization, achieving temporally coherent enhancement *without* optical-flow alignment.
• **Second, user-perception-aware dynamic controller.** Our key idea is to *redefine* diffusion video enhancement as a *perception-budgeted control problem*: beyond a user's sensitivity thresholds, extra denoising steps or attention computation yield diminishing returns, while instability (flicker/lag) is immediately noticeable and task-dependent. We therefore prioritize *perceptual stability* under strict latency/energy budgets, explicitly coupling *what we compute* to *what users perceive*. Leveraging redundancy across frames, attention states, and denoising steps, we expose four coordinated knobs: keyframe caching, attention reuse, sampling-depth scheduling, and resolution scaling. We then design a closed-loop controller that jointly selects their combination to maximize perceived quality while minimizing energy subject to a real-time latency constraint, supported by a low-overhead profiler that estimates latency/energy/memory online.
Representative Applications. We highlight three scenarios where enhancement is *functionally critical* for users rather than merely aesthetic (Fig. 1).

- **Nighttime Driving and Assistance.** Headlight glare and weak dark-region cues reduce visibility at night. Image-based enhancement can introduce flicker and unstable object boundaries, undermining situational awareness and user trust.
- **In-Home Nighttime Monitoring.** Low-light cameras are widely used for child and elderly monitoring. Brightness oscillation and contour distortion can trigger false alarms and erode long-term reliability.
- **Smartglasses for low-light augmentation (night-blindness and dim museums).** For night-blind users in particular, enhancement must be temporally stable, unobtrusive on near-eye displays, and energy-efficient for prolonged use. In dim museums/heritage sites with restricted lighting, users rely on mobile or wearable devices to perceive fine details.

We implement UbiEnlight as a **joint algorithm-system design** for on-device low-light video enhancement on mobile and wearable platforms. It tightly integrates a spectrum-guided video-native enhancement model with a runtime-adaptive controller, enabling real-time and temporally stable enhancement under dynamic low-light conditions. We evaluate it on *four* heterogeneous platforms (phones and smart glasses) and *four* real-world low-light video datasets covering diverse illumination dynamics, motion patterns, and extreme low-light conditions.

Against *nine* state-of-the-art baselines, UbiEnlight cuts end-to-end latency by up to *99.68%* (*e.g.*, *18.81s* → *60ms*), enabling real-time operation on resource-constrained devices, while delivering *stable* enhancement (improving temporal consistency by up to 54.76%, fidelity by 44.15%) under *mixed* motion and *non-stationary* illumination where prior methods flicker or amplify blur. These gains stem from spectrum-guided conditioning and hierarchical reuse, aligning efficiency with human perception by prioritizing temporal stability and adapting computation online under dynamic and extreme lighting. Our main contributions are summarized as follows.

- To our knowledge, UbiEnlight is the first diffusion-transformer-based real-time low-light video enhancement system designed for non-expert mobile/wearable users under dynamic illumination and mixed user motion.
- We introduce a *spectrum-guided* latent diffusion transformer that injects frequency-domain priors (amplitude for illumination, phase for structure) for temporally coherent enhancement, and a user-perception-aware controller that adaptively allocates computation via hybrid reuse to balance quality, latency, and energy.
- Extensive experiments show that UbiEnlight achieves real-time performance with improved temporal stability and perceptual quality, being robust to dynamic/extreme conditions and aligning with user-perceived usability.

2 BACKGROUND AND OVERVIEW

2.1 Ubiquitous Applications of Low-Light Video Enhancement

Ubiquitous cameras on smartphones and wearables have become a primary interface for users to perceive, record, and interact with the physical world. However, these systems often operate in *uncontrolled lighting environments*, where videos are frequently captured under dim illumination (*e.g.*, nighttime commuting, dark indoor spaces, and illumination-restricted venues such as museums). In such settings, low-light videos suffer from severe noise, motion blur, and suppressed details, directly undermining *user perception*. This motivates *low-light video enhancement* as a key enabler for ubiquitous visual sensing applications.

We further conducted a user study with 53 participants (age 21-68), which shows that low-light degradation is not perceived merely as reduced visual quality. Instead, users strongly associate it with tangible functional consequences, including compromised safety (84.9%), reduced system reliability (77.4%), and impaired day-to-day usability (90.6%). They also report that low-light enhancement benefits both *human-facing* scenarios (*e.g.*, visual communication and media recovery) and *downstream machine perception* tasks (*e.g.*, detection, tracking, and scene understanding).

Design Goals. Informed by these applications and user feedback, we target low-light video enhancement for mobile and wearable platforms with three user-centric objectives:

- **Temporal Continuity**: Users are more sensitive to flicker and drift than to minor per-frame artifacts.
- **Real-Time Responsiveness**: Enhancement latency should remain imperceptible to support interactive and safety-critical usage.
- **Resource-Aware Sustainability**: Continuous enhancement must adapt to battery and resource constraints for long-term operation.

2.2 Limitations of Existing Low-Light Solutions

Despite the demand from ubiquitous applications, low-light video enhancement remains challenging on mobile and wearable platforms. In this work, we focus on **algorithmic enhancement** on commodity RGB cameras, without requiring specialized sensors, auxiliary illumination, or cloud offloading. This scope is motivated by deployability and privacy: many human-facing and safety-critical scenarios require immediate feedback and local processing, and cannot assume extra hardware or reliable connectivity.

Table 1. Comparison of existing low-light visual sensing solutions in terms of efficiency (accuracy, energy cost, real-time performance), characteristics, sensors, and specialized hardware.

Category	Sensors	Efficiency			Suitable applications	Special demands	
		Accuracy	Energy cost	Real-time		Extra illuminator	Custom hardware
Special imaging hardware	CMOS/CCD Low-light sensor	High	Medium	✓	Industrial camera, General imaging	-	Industrial/scientific cameras
	Thermal Infrared / NIR camera	Medium	High	✓	Surveillance/vehicle/UAV	NIR illuminator	IR camera
	Radar	High	High	✓ (low-fps)	Automotive/robotics	Built-in laser emitter	LiDAR (laser, photodetector)
Multi-modal fusion systems	Event camera + RGB	High	Low	×	Robotics/UAV/AR	-	dual cameras
	RGB+NIR fusion camera	High	Medium	×	Industrial camera/inspection	NIR illuminator	dual sensors
	Thermal+visible fusion camera	High	Medium-High	×	Automotive/security	-	dual cameras

2.2.1 Limitations of Hardware and Multi-Modal Approaches. Hardware-based solutions (*e.g.*, low-light sensors [41, 77], IR/NIR modules [19, 44], or active illumination [38, 50, 65]) can improve signal quality under challenging lighting conditions, but they often incur higher energy cost and form-factor complexity, typically requiring specialized sensors or custom hardware (Tab. 1). These requirements hinder scalability and limit practicality for commodity ubiquitous devices such as smartphones and wearables. Multi-modal fusion [30, 51, 75] further improves robustness by combining RGB with depth, thermal, IR, or event sensors. While effective in controlled settings (Tab. 1), such systems introduce substantial calibration overhead, parallel sensing pipelines, and additional hardware dependencies, complicating end-to-end integration. Consequently, they are often incompatible with always-on, and energy-constrained mobile operation, limiting their suitability for pervasive deployment in ubiquitous computing environments.

2.2.2 Limitations of Existing Algorithmic Enhancement. Algorithmic methods avoid extra hardware, but current techniques still fail to meet the goals in Sec. 2.1, mainly due to two coupled limitations.

- **Temporal instability under real-world dynamics.** Many low-light enhancers are designed for images [42, 64, 76, 80] and applied to videos frame by frame, producing exposure jitter and flicker that are highly salient to users and disruptive to downstream perception. Video-specific methods rely on optical flow [52] or recurrent propagation [72], but these mechanisms become fragile under extremely low signal-to-noise ratios and severe motion blur. As a result, existing methods struggle to deliver the **temporal continuity** required by ubiquitous applications.
- **Prohibitive and inflexible computation on ubiquitous devices.** High-quality enhancement increasingly relies on heavy neural models and iterative generative pipelines. Diffusion-based enhancement [27, 46, 71] is promising for perceptual fidelity, yet iterative sampling and repeated attention computation incur substantial latency and energy overhead. Moreover, ubiquitous deployment is inherently dynamic: device resources fluctuate, and video content exhibits time-varying user/target *motion* and *illumination* changes. A fixed configuration cannot sustain a stable quality-latency-energy trade-off across scenes and devices, conflicting with the need for **real-time responsiveness** and **resource-aware sustainability**.

Takeaway. These limitations call for a new design that *(i)* enforces temporally stable enhancement under low-light dynamics and *(ii)* adapts computation to runtime resource and scene variations, enabling real-time and sustainable operation on ubiquitous devices.

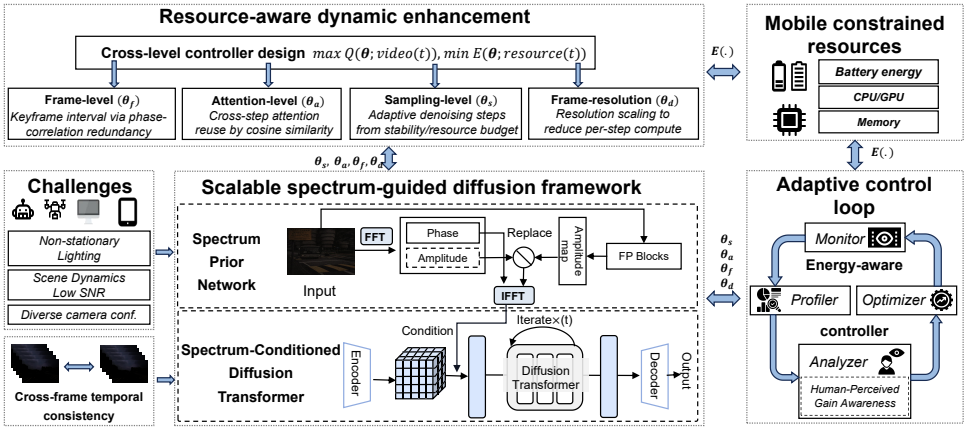


Fig. 2. Workflow of UbiEnlight

2.3 Solution Overview

To meet these requirements, we present UbiEnlight, a low-light video enhancement system that delivers **temporally stable** enhancement with **real-time** and **resource-aware** execution on ubiquitous devices. UbiEnlight is guided by two insights that translate the above goals into system designs.

- **Spectrum priors provide a stable anchor for illumination correction and structure preservation.** Pixel-space appearances are highly unstable in low light. This makes it difficult for image-based models to maintain consistent structure when applied to videos, and it also makes flow-based alignment unreliable. Instead, UbiEnlight leverages a more robust representation in the frequency domain. Fourier *amplitude* reflects illumination and energy distribution, while *phase* encodes geometric layout and structural contours. Importantly, phase tends to remain stable under illumination changes, making it a natural **structural anchor** for temporally coherent enhancement. This motivates a **spectrum-guided diffusion model** (Sec. 3), which injects spectrum priors into diffusion denoising to brighten low-light content while constraining structure evolution.
- **Human-perceived gains saturate, enabling adaptive computation through reuse and skipping.** While diffusion models can significantly improve perceptual quality, running a fixed number of denoising steps with full attention computation is often prohibitive on mobile and wearable devices. Meanwhile, resource availability and video content vary over time (*e.g.*, background workloads, thermal throttling, static vs. fast-motion segments). Since users prioritize temporal continuity and responsiveness, extra computation beyond perceptual thresholds often yields diminishing returns. This motivates **resource-aware dynamic control** (Sec. 4) that adapts computation at runtime to balance enhancement quality, latency, and energy.

Architecture. Based on these insights, UbiEnlight integrates two tightly coupled components (Fig. 2). First, its **spectrum-guided diffusion model** (Sec. 3) predicts amplitude-based illumination guidance and uses phase as a structural anchor to guide denoising, suppressing flicker and geometry drift without relying on motion estimation. Second, it builds a lightweight **closed-loop control stack** (Sec. 4) that exposes efficiency knobs via hierarchical reuse and selective skipping across *frames*, *denoising steps*, and *attention computation*. A runtime profiler and controller then select the configuration to satisfy real-time constraints under fluctuating resources, while preserving user-perceived quality.

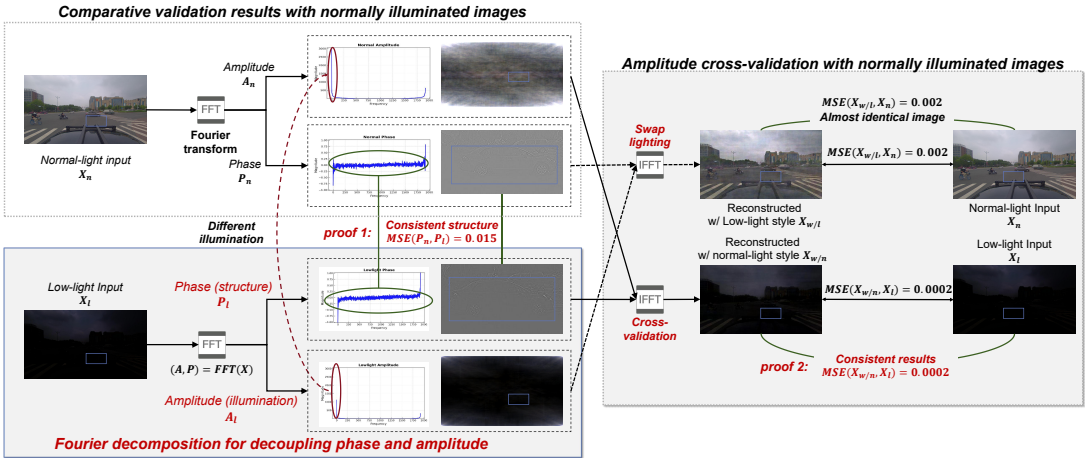


Fig. 3. Illustration of Fourier-based decomposition of illumination and structure.

3 SPECTRUM-GUIDED DIFFUSION MODEL FOR LOW-LIGHT VIDEO ENHANCEMENT

This section presents the spectrum-guided diffusion model of UbiEnlight for low-light video enhancement. Our design is motivated by two requirements: *(i) low-light enhancement* should increase illumination without distorting structure; and *(ii) video enhancement* must be temporally stable. We build our model upon diffusion transformers [48]given their strong performance on image restoration and video generation. However, existing diffusion-based low-light enhancers [28, 46, 62, 71] are developed for *images* and process frames independently when applied to videos, leading to temporal inconsistency. Meanwhile, diffusion models designed for *video generation* [21, 47, 79] emphasize visual plausibility and lack explicit mechanisms to anchor structure, making them prone to structural drift or hallucinated details when repurposed for video enhancement. To meet these requirements, we introduce a *spectrum prior network* that provides amplitude-based illumination priors with phase as a structural anchor (Sec. 3.1), and a *spectrum-conditioned diffusion transformer* that injects these priors to guide denoising and stabilize temporal behavior (Sec. 3.2).

3.1 Spectrum Prior Network

This subsection introduces the *Spectrum Prior Network (SP)*, a lightweight module that extracts spectrum-based priors for diffusion denoising. SP is motivated by a simple observation: in low-light videos, illumination degradation mainly reduces signal energy, while the underlying scene layout is comparatively stable (Fig. 3). Accordingly, SP predicts an *illumination modulation prior* in the *amplitude* spectrum and retains the *phase* spectrum as a *structural anchor*. These priors are projected into conditioning tokens and injected into the diffusion transformer (Sec. 3.2) to guide both enhancement quality and temporal stability.

Spectrum Decomposition. Given an input frame x, we compute a luminance proxy $Y = \psi(x)$ (*e.g.*, the Y channel) and apply a Fourier transform:

$$\mathbf{S} = \mathcal{F}(Y), \qquad \mathbf{A}_{in} = |\mathbf{S}|, \qquad \mathbf{P}_{in} = \angle \mathbf{S}, \tag{1}$$

where $\mathcal{F}(\cdot)$ denotes the 2D Fourier transform, and $\mathbf{A}_{in}$ and $\mathbf{P}_{in}$ are the amplitude and phase spectra.

Network Design. We adopt the amplitude-transform formulation (i.e., predicting a transform map rather than directly regressing Fourier amplitudes), which has been shown to better preserve amplitude structure and avoid detail/color degradation under low-light conditions [58]. Specifically, SP predicts a spatially varying *amplitude*

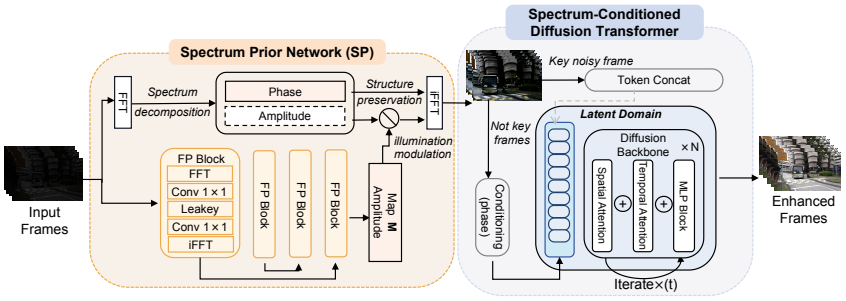


Fig. 4. Architecture of the Spectrum-guided Diffusion Transformer model for lowlight enhancement.

transform map $\mathbf{M}$ that regulates how much illumination should be amplified. We implement SP as a lightweight network with four Fourier Processing (FP) blocks and skip connections [58]. Each FP block follows an FFT-process-IFFT routine: intermediate features are modulated in the frequency domain using lightweight 1×1 convolutions on the amplitude and phase branches, and then mapped back to the spatial domain via a 3×3 convolution with residual connections. A sigmoid layer constrains $\mathbf{M}$ to $(0, 1)$ to avoid excessive amplification and improve stability.

Amplitude Modulation and Phase Preservation. SP performs illumination correction by modulating the amplitude spectrum while keeping the phase fixed:

$$\mathbf{A}_{out} = \mathbf{A}_{in} \odot \frac{1}{\mathbf{M} + \varepsilon}, \qquad Y^{(s1)} = \mathcal{F}^{-1}\Big(\mathbf{A}_{out} \odot e^{j\mathbf{P}_{in}}\Big), \tag{2}$$

where $\odot$ denotes element-wise multiplication, ε is a small constant for numerical stability, and $\mathcal{F}^{-1}(\cdot)$ is the inverse Fourier transform. This yields a structure-faithful luminance estimate $Y^{(s1)}$, which serves as an effective prior for subsequent diffusion denoising.

SP is supervised using an amplitude alignment objective against the normal-light ground truth:

$$\mathcal{L}_{\text{freq}} = \Big\| \big|\mathcal{F}(Y^{(s1)})\big| - \big|\mathcal{F}(Y_{\text{gt}})\big| \Big\|_2^2, \tag{3}$$

where $Y_{\text{gt}} = \psi(x_{\text{gt}})$ is the luminance proxy of the ground-truth frame. Finally, SP outputs are projected into conditioning tokens $c = \Pi(\mathbf{M}, \mathbf{P}_{in})$ using a lightweight projector $\Pi(\cdot)$, and injected into the diffusion transformer for spectrum-conditioned denoising.

3.2 Spectrum-Conditioned Diffusion Transformer

This subsection describes how we use the spectrum priors from SP to guide diffusion denoising (Fig. 4) for *temporally stable* low-light video enhancement. In particular, the priors are used in two complementary ways: *(i)* they condition the denoising backbone (Sec. 3.2.1) to improve illumination correction and structure preservation; and *(ii)* they provide signals to stabilize temporal behavior (Sec. 3.2.2) by regulating reuse and suppressing diffusion-induced randomness.

3.2.1 Diffusion Backbone. We adopt a DiT-style diffusion backbone [8, 48] for reverse denoising in the *latent domain*, where input video frames are first encoded into compact spatial tokens using a lightweight encoder, following the standard latent diffusion paradigm. This design reduces spatial resolution by 4-8×, substantially lowering computational and memory overhead on ubiquitous devices. The backbone follows a spatio-temporal Transformer design, and each denoising block is composed of three functional components: each denoising block contains three components: *(i) spatial self-attention* to model intra-frame structures and textures; *(ii)*

temporal self-attention to propagate reliable information across neighboring frames and suppress frame-wise stochastic variations; and *(iii) Spectrum-guided conditional diffusion* that injects SP-derived conditioning tokens c into denoising process by encoding spectrum-derived representations as latent conditioning signals, which are incorporated into the diffusion model to steer the denoising trajectory toward frequency-consistent and temporally stable video enhancement results. Unlike purely frame-wise diffusion, temporal attention enables the model to reuse latent representations from adjacent frames as an implicit temporal prior, which improves stability under motion and reduces diffusion-induced flicker. Meanwhile, spectrum-conditioned cross-attention decouples illumination control from structural generation by explicitly providing amplitude and phase cues at each denoising step. By applying spectral conditioning consistently throughout the reverse diffusion trajectory, the model enhances visibility while preserving scene layout and temporal coherence, which is critical for low-light video enhancement in mobile and wearable scenarios.

3.2.2 Temporal Stabilization Mechanisms. Diffusion-based video enhancement can still exhibit temporal artifacts due to *unstable propagation*, *stochastic sampling*, and *long-horizon drift*. Building on the spectrum-conditioned backbone, we introduce three lightweight stabilization mechanisms that address these issues.

- **Stabilizing Propagation: Spectrum-Gated Temporal Reuse.** Temporal attention enables frame-to-frame propagation, but naively reusing past information can be harmful when illumination changes abruptly or motion is fast. We therefore *adapt the reuse strength* using spectrum cues. Specifically, we observe that the *phase spectrum* is closely related to scene structure and is often more stable across frames, whereas the *amplitude spectrum* varies more with illumination. Based on amplitude/phase consistency, we compute a compact similarity score s_i and use it to modulate a temporal reuse gate:

$$\alpha_i = \text{clip}\left(\gamma\, s_i,\ 0,\ 1\right), \tag{4}$$

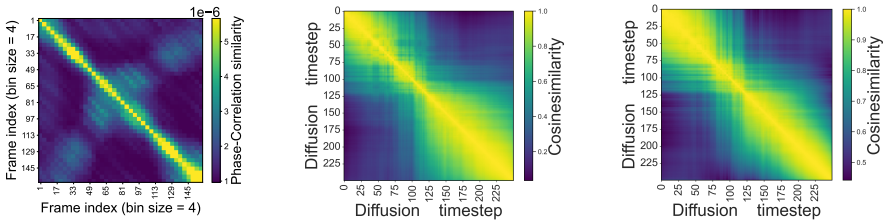
  where γ is a scaling factor and $\alpha_i \in [0, 1]$ controls how strongly temporal information is reused. When adjacent frames have similar spectra, α_i increases, encouraging stronger reuse.
- **Stabilizing Sampling: Correlated Noise across Frames.** Even with controlled propagation, diffusion sampling itself introduces randomness. If each frame samples noise independently, small sampling differences can accumulate into perceptible flicker. To suppress such frame-wise jitter without optical-flow alignment, we adopt *correlated* noise sampling. At denoising step t, the noise for frame i is generated as:

$$\epsilon_{i,t} = \rho\, \epsilon_{i-1,t} + \sqrt{1-\rho^2}\, \eta_{i,t}, \tag{5}$$

  where $\eta_{i,t} \sim \mathcal{N}(0, I)$ is newly sampled Gaussian noise and $\rho \in [0, 1)$ controls the correlation strength. A larger ρ increases stochastic consistency across adjacent frames and reduces flicker, while the injected noise maintains diversity and prevents over-smoothing.
- **Stabilizing Structure: Phase-Aware Temporal Stabilization.** Over long sequences, stochastic denoising can still accumulate small inconsistencies and introduce *geometry drift*. To further stabilize structure, we introduce phase-aware temporal stabilization, which explicitly enforces *continuity of structural evolution* across frames.
  - For consecutive frames, we compute luminance proxies (Y_{i-1}, Y_i) and their phase spectra (P_{i-1}, P_i). We characterize inter-frame structural evolution using a phase-change descriptor:

$$\Delta P_i = P_i - P_{i-1}, \qquad w_k = A_{i-1}(k)^{\beta}, \tag{6}$$

    where k indexes frequency bins and w_k suppresses noise-dominated components via amplitude-aware weighting. The exponent $\beta > 0$ controls the sensitivity to reliable frequency components; ΔP_i captures motion-consistent structural changes while filtering enhancement-induced artifacts.
  - Rather than being injected through an explicit projection or concatenation, the phase-change cue influences temporal attention by *constraining the set of plausible contextual representations* reused across



(a) frame-wise redundancy (b) attention redundancy (c) mlp redundancy

Fig. 5. Temporal redundancy and efficiency characteristics in diffusion-based video enhancement.

frames. Specifically, correlated noise sampling aligns the diffusion trajectories of adjacent frames, while cached latent tokens from the previous frame already encode spectrum-guided structural priors from Stage 1. Under this setting, temporal attention operates over a restricted and structurally consistent context space. The phase-change descriptor ΔP_i thus serves as a criterion for distinguishing motion-consistent structural evolution from stochastic sampling variations, encouraging the model to reuse phase-consistent latent structures while suppressing diffusion-induced jitter. From a modeling perspective, this effect can be abstractly described as current-frame latent tokens attending to prior-frame context under phase-consistency constraints:

$$\mathbf{z}_i' \leftarrow \mathrm{Attn}_{\Delta P_i}\big(Q(\mathbf{z}_i), K(\mathbf{z}_{i-1}), V(\mathbf{z}_{i-1})\big), \tag{7}$$

where ΔP_i does not act as an explicit conditioning input, but modulates the attention behavior by favoring reuse of phase-consistent structures while suppressing stochastic variations introduced by diffusion sampling. This encourages the model to reuse reliable *structural* evidence, distinguish *motion-induced* phase changes from diffusion randomness, and stabilize geometry without optical-flow alignment.

– We further enforce inter-frame consistency with a regularization loss:

$$\mathcal{L}_{phase} = \sum_i \Big(1 - \mathrm{Coh}(P_{\hat{Y}_i}, P_{\hat{Y}_{i-1}})\Big), \qquad \mathrm{Coh}(P_a, P_b) = \frac{\left|\sum_k w_k e^{j(P_a(k)-P_b(k))}\right|}{\sum_k w_k}, \tag{8}$$

which penalizes abnormal phase deviations while preserving legitimate motion. As a result, the enhanced video exhibits smooth and physically plausible structural evolution.

All components of the diffusion transformer, including the backbone and the temporal stabilization mechanisms, operate in the *latent domain* for efficiency. Operating in the latent space reduces spatial resolution by 8×, lowering latency and memory usage (Sec. 3.2.1). Finally, beyond these architectural choices, we further design explicit compute reuse mechanisms to boost real-time performance on resource-limited devices, as explained next.

4 USER-PERCEPTION-AWARE DYNAMIC CONTROL FOR REAL-TIME VIDEO ENHANCEMENT

This section presents the user-perception-aware dynamic control stack of UbiEnlight for real-time low-light video enhancement. While our spectrum-guided diffusion model (Sec. 3) improves perceptual quality, iterative denoising and repeated attention make it hard to meet mobile/wearable latency and energy budgets. To address this, UbiEnlight adopts a lightweight closed-loop design with three components: a **hierarchical reuse and skipping strategy** (Sec. 4.1) that exploits redundancy across frames, attention, and denoising steps; a **resource-aware controller** (Sec. 4.2) that meets a real-time latency budget while balancing quality and energy; and a **runtime cost profiler** (Sec. 4.3) that estimates resource usage with minimal overhead.

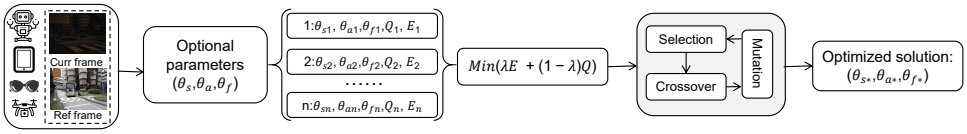
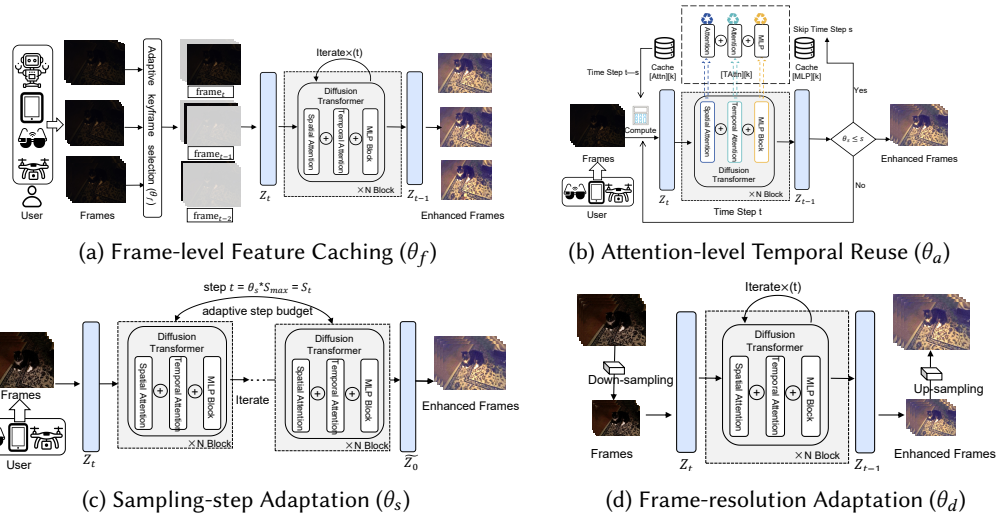


Fig. 6. Workflow of the optimizer in the resource-aware dynamic controller.



(a) Frame-level Feature Caching (θ_f)

(b) Attention-level Temporal Reuse (θ_a)

(c) Sampling-step Adaptation (θ_s)

(d) Frame-resolution Adaptation (θ_d)

Fig. 7. Illustration of hierarchical reuse and skipping mechanisms in the proposed runtime scaling strategy.

4.1 Hierarchical Reuse and Skipping Strategies

Diffusion-based low-light video enhancement is computationally expensive, as it iteratively denoises each frame and repeatedly computes transformer attention. In mobile/wearable use, users are often more sensitive to *flicker and delay* than marginal per-frame fidelity, especially during driving/riding/running or on near-eye displays. And videos exhibit redundancy at three levels: **frame-wise redundancy** across adjacent frames with similar structure/illumination; **attention redundancy** when attention states change slowly across neighboring denoising steps; and **sampling redundancy** where enhancing every frame is unnecessary under *mild motion*. This motivates UbiEnlight to trade *redundant computation* for *perceptual stability* and *real-time deployment* by exploiting four coordinated adaptation mechanisms.

Frame-Level Redundancy Exploitation via Keyframe Selection (θ_f). Diffusion-based video enhancement applies multi-step denoising to every frame, making latency proportional to the number of processed frames. However, adjacent mobile video frames exhibit strong temporal redundancy. As shown in Fig. 5a, over 70% of adjacent frame pairs have phase-correlation similarity above 0.9, indicating largely consistent amplitude–phase structures under moderate motion and illumination changes. To exploit this redundancy, UbiEnlight enhances only selected *keyframes* and reuses cached latent states (*e.g.*, features and attention context) for intermediate frames, avoiding explicit motion estimation. Illumination-related amplitude components are inherited from the latest keyframe, while fine-grained variations are adaptively refined. The keyframe interval is determined by the phase-correlation similarity $s_t \in [0, 1]$ between the current frame and the latest keyframe:

$$K_t = \text{clip}\left(\left\lfloor \frac{\kappa}{1 - s_t + \epsilon} \right\rfloor, K_{\min}, K_{\max}\right), \tag{9}$$

where κ is a scaling factor, ϵ prevents division by zero, and $K_{\min}/K_{\max}$ bound the interval for responsiveness and stability (Fig. 7a). Higher similarity results in a longer reuse interval, whereas lower similarity triggers earlier keyframe updates, reducing redundant computation while preserving temporal consistency.

Attention-Level Reuse across Denoising Steps (θ_a). Within diffusion transformers, neighboring denoising steps often operate on highly correlated latent states, making attention computation partially redundant as shown in Fig.5b and Fig.5c. Within diffusion transformers, neighboring denoising steps often operate on highly correlated latent states, making attention computation partially redundant . Hence, UbiEnlight reuses cached attention outputs when they remain sufficiently similar across steps. Let O_s^{attn} be the multi-head attention output at denoising step s, and $O_{s'}^{\text{attn}}$ the cached output from the most recent step $s' < s$. We compute cosine similarity $S_s = \frac{\left\langle O_s^{\text{attn}}, O_{s'}^{\text{attn}} \right\rangle}{\|O_s^{\text{attn}}\|_2 \cdot \|O_{s'}^{\text{attn}}\|_2}$. If $S_s > \tau_{\text{attn}}(\theta_a)$, UbiEnlight reuses $O_{s'}^{\text{attn}}$; otherwise, it recomputes attention and updates the cache(seeing Fig. 7b). A larger θ_a promotes reuse for efficiency, while a smaller θ_a enforces more frequent recomputation to preserve temporal stability.

Sampling-Step Adaptation via Stability-Aware Denoising Scheduling (θ_s). Diffusion inference remains the primary latency bottleneck, because the denoising network is executed repeatedly over a step schedule $\{t_1, \ldots, t_S\}$, making runtime approximately proportional to the number of sampling steps S. To provide fine-grained control of per-keyframe computation under dynamic resource constraints, *sysname* introduces a sampling-step knob θ_s that directly determines the step budget used for each enhanced keyframe. Mechanistically, θ_s maps the current resource state (*e.g.*, latency/energy budget) and the estimated input difficulty (*e.g.*, darkness level or structural change) to a step count:

$$S_t = \text{clip}\left(\lfloor \theta_s \cdot S_{\max} \rfloor, S_{\min}, S_{\max}\right) \tag{10}$$

where S_t is the number of denoising steps executed for keyframe t, and $[S_{\min}, S_{\max}]$ bounds the feasible range. When energy is sufficient or latency constraints are relaxed, θ_s increases, allocating more denoising steps to better refine high-frequency details and reduce residual noise. Under tight latency or energy budgets, θ_s decreases, shortening the denoising trajectory and reducing compute(seeing Fig. 7c), while relying on (i) the Stage-1 spectrum-guided prior and (ii) temporal stabilization (e.g., correlated noise) to limit flicker and maintain coherent structure. This sampling-step adaptation complements keyframe selection: keyframe selection reduces the number of frames requiring diffusion, whereas θ_s controls the per-keyframe denoising depth, jointly balancing quality, latency, and energy for streaming deployment.

Resolution Scaling (θ_d). In addition to reducing denoising steps, UbiEnlight adapts the input resolution to directly control per-frame computation. Specifically, $\theta_d \in (0, 1]$ denotes a down-sampling factor applied to each incoming frame before enhancement(seeing Fig. 7); UbiEnlight performs enhancement on the resized frames (aligned to the model stride) and then upsamples the result back to the target resolution.

4.2 Resource-Aware Controller

We design a controller that dynamically coordinates the four efficiency knobs in Sec. 4.1, frame-level caching (θ_f), attention reuse (θ_a), and sampling-step adaptation (θ_s), and frame-resolution adaptation(θ_d) to maintain real-time enhancement under time-varying scene dynamics and device budgets.

Formulation. At each time t, given the incoming stream video(t) and measured resource state resource(t), the controller selects a configuration $\boldsymbol{\theta} = (\theta_s, \theta_a, \theta_f, \theta_d)$ to balance enhancement quality, energy cost, and latency. We model this as a constrained multi-objective control problem:

$$\begin{aligned} \max_{\boldsymbol{\theta}} \quad & Q(\boldsymbol{\theta}; \text{video}(t)), \qquad \min_{\boldsymbol{\theta}} \quad E(\boldsymbol{\theta}; \text{resource}(t)), \\ \text{s.t.} \quad & T(\boldsymbol{\theta}; \text{resource}(t)) \le T_{\max}, \end{aligned} \tag{11}$$

where $Q(\cdot)$ is the resulting video quality, $E(\cdot)$ is the energy cost, and $T(\cdot)$ is the per-frame latency. This formulation captures the trade-off between perceptual quality and resource expenditure while enforcing real-time constraints.

Optimizer. We adopt a two-stage optimization strategy consisting of offline profiling and online adaptation. *i) Offline.* We construct a Pareto-optimal configuration library that maps runtime knobs $\boldsymbol{\theta} = (\theta_s, \theta_a, \theta_f, \theta_d)$ to their measured quality, energy, and latency (Q, E, T). Here, θ_s, θ_a, θ_f, and θ_d control the sampling-step budget, attention reuse, keyframe interval, and input resolution, respectively. The decision space is intentionally limited, with $\theta_s \in \{2, 4...18, 20\}$,$\theta_a \in \{1, 1/2...1/5\}$, $\theta_f \in \{0, 1/5, 1/3, 1/2, 1\}$, and $\theta_d \in \{1, 1/2, 1/3\}$. For each candidate configuration, we measure the corresponding enhancement quality, energy cost, and per-frame latency on representative low-light videos and target hardware profiles. We profile all candidate configurations on representative low-light videos and target devices, discard those violating the latency constraint ($T > 300$ ms), and retain only Pareto-optimal configurations. The resulting library typically contains 35∼45 configurations per platform. *ii) Online.* The controller monitors device status and lightweight scene indicators, and selects a runtime configuration from the Pareto library according to a preference weight $\lambda \in [0, 1]$, which balances enhancement quality and efficiency:

$$\boldsymbol{\theta}^{\star} = \arg\max_{\boldsymbol{\theta} \in \mathcal{L}} U(\boldsymbol{\theta}) \quad \text{s.t.} \quad T(\boldsymbol{\theta}) \le T_{\max}, \qquad U(\boldsymbol{\theta}) = (1-\lambda)\, Q(\boldsymbol{\theta}) - \lambda\, E(\boldsymbol{\theta}), \tag{12}$$

where $\mathcal{L}$ denotes the offline Pareto library. As scene dynamics or resource availability change, the controller updates λ and switches configurations accordingly, maintaining a balanced tradeoff among quality, latency, and energy during real-time enhancement.

4.3 Runtime Cost Profiler

To support online control, UbiEnlight includes a lightweight profiler that estimates the **energy** and **latency** cost of a candidate configuration $\boldsymbol{\theta} = (\theta_s, \theta_a, \theta_f, \theta_d)$ without invasive instrumentation. The profiler combines an analytic layer-level cost model with a small set of runtime signals, and provides cost feedback to the controller in Sec. 4.2.

Energy Cost. Extending AdaEnlight [40] to diffusion transformers for videos, we model layer-level computation and memory behavior to guide perception-aware runtime control. The energy for layer l is decomposed as:

$$E_l = \delta_C \cdot C_l + \epsilon \cdot \delta_{\text{cache}} \cdot M_l + (1-\epsilon) \cdot \delta_{\text{DRAM}} \cdot M_l + M_l \cdot \delta_{\text{SM}} \tag{13}$$

where C_l is the number of MAC operations, M_l is the memory traffic, and $(\delta_C, \delta_{\text{cache}}, \delta_{\text{DRAM}}, \delta_{\text{SM}})$ are device-calibrated energy costs for computation, cache, DRAM, and on-chip memory. ϵ represents the cache-hit rate, obtained offline through micro-benchmarks. Key components include: i) MAC operations C_l: computes offline based on the layer's architecture. For a diffusion transformer with windowed attention, C_l depends on QKV projections, attention multiplication, and FFN operations. ii) Memory accesses M_l: estimates offline based on dataflow in attention and FFN layers. Memory access includes parameter loads, activation reads/writes, and intermediate buffers. iii) Cache-hit ratio ϵ: measures online to reflect real-time cache performance, updated every few frames. The total energy cost for enhancing a video with n_{frame} frames under adaptive hyperparameters $\theta_f, \theta_a, \theta_d, \theta_s$ is:

$$E_{\text{enhancer}} = \sum_{f=1}^{n_{\text{frame}}} \sum_{t=1}^{S(\theta_s)} \sum_{l=1}^{L} \theta_f(f)\, \theta_a(l)\, E_l\big(\theta_d^{f,t}\big), \tag{14}$$

where $\theta_f(f)$ and $\theta_a(l)$ are binary indicators for frame f and layer l, and $S(\theta_s)$ is the number of diffusion steps. This allows the controller to assess the impact of skipping steps and frame down-sampling on energy.

Memory Access Modeling. We model memory access mainly as a runtime constraint, since memory is not the active bottleneck in our deployment. For each layer l, we estimate memory accesses M_l from the attention and FFN dataflow, and aggregate the per-step cost as $M_{\text{step}} = \sum_l \theta_l M_l$. We use this metric to monitor memory pressure during execution. Across tested platforms and configurations, memory usage stays stable (peak < 1.7 GB, average ~ 1.5 GB), making latency and energy the dominant constraints.

Resource Cost. We further define a composite resource cost to balance energy, computation, and memory:

$$R(\theta_s, \theta_a, \theta_f, \theta_d) = \beta_E \, \widehat{E}_{\text{enhancer}} + \beta_C \, \widehat{C}_{\text{total}} + \beta_M \, \widehat{M}_{\text{total}}, \tag{15}$$

where $\widehat{E}$enhancer, $\widehat{C}$total, and $\widehat{M}_{\text{total}}$ are normalized costs for energy, computation, and memory, and β_E, β_C, and β_M are tunable weights reflecting resource priorities for the platform.

4.3.1 Profiler Workflow. Accurate, low-overhead performance profiling is essential for adaptive control on mobile platforms. UbiEnlight uses a two-stage workflow that combines offline calibration with online integration to bridge hardware-level resource cost modeling and runtime adaptation. *Offline Calibration and Modeling.* We derive platform-specific energy coefficients ($\delta_C, \delta_{\text{cache}}, \delta_{\text{DRAM}}, \delta_{\text{SM}}$) and peak throughput baselines through hardware profiling. By correlating power measurements with micro-benchmarks, we normalize memory access costs relative to MAC operations, creating a static energy profile independent of the model architecture. *Online Runtime Monitoring and Integration.* During inference, the profiler monitors lightweight signals (*e.g.*, cache-hit rate) and the current configuration Θ. By combining analytic models with runtime feedback, it estimates energy and latency *without disrupting user interaction.* This feedback closes the loop with the controller, enabling continuous adaptation to *scene dynamics*, *device state*, and *user-perceived responsiveness.*

4.4 Adaptive Control Loop

UbiEnlight employs a *human-centric adaptive control loop* to balance perceived visual quality, energy efficiency, and responsiveness under dynamic usage conditions (Fig. 6). The controller periodically (*e.g.*, every second) monitors device resources (*e.g.*, battery, compute load, memory, and thermal headroom) together with profiler-estimated latency and energy of the current configuration. When resource budgets or scene dynamics change, it re-optimizes the runtime knobs ($\theta_s, \theta_a, \theta_f, \theta_d$), covering sampling depth, attention reuse, frame caching, and frame-resolution adaptation, using a Pareto-based multi-objective formulation that maximizes perceived quality under latency and energy constraints. The selected configuration is applied only after throughput and thermal safety checks, enabling continuous adaptation without disrupting user interaction. When the input frame rate exceeds the achievable enhancement throughput, the controller favors bounded-latency enhancement by adaptively selecting keyframes and reuse configurations, instead of executing full diffusion on every frame.

5 EXPERIMENT

5.1 Settings

Implementation. We implement UbiEnlight in PyTorch and train all models on a workstation with an NVIDIA GeForce RTX 3090 GPU (24GB VRAM, CUDA 12.6). We deploy UbiEnlight on four representative platforms: Google Pixel 7 (Android, 8GB RAM; Dev_1), NVIDIA Jetson AGX Orin (Linux, 64GB RAM; Dev_2), Xiaomi 17 Ultra (Android, 16GB RAM; Dev_3), and XREAL One Pro AR glasses paired with a smartphone for computation (Dev_4). Deployment uses platform-specific runtimes: PyTorch Mobile (.ptl) on Android devices, CUDA-accelerated PyTorch on Jetson, and a tethered pipeline for XREAL One Pro, where captured frames are streamed to the paired smartphone for enhancement and then returned for display. Our diffusion backbone is built on STDiT and follows the standard DDPM reverse denoising process. The adaptive controller adjusts only the denoising-step budget, while the sampling algorithm remains unchanged. For reproducibility, we fix the random seed and

Table 2. Overview of low-light frame and video datasets (including our self-collected UbiVideo).

Datasets	Data type	Sample number	Description
LSRW [18]	Paired frames	5,650 pairs	Real-world paired low/normal-light frames.
LOL-v2 [64]	Paired frames	1,789 pairs	Paired low-light frames (Real + Synthetic).
DID [13]	Paired videos	413 videos / 41,038 frames	Real-world paired low/normal-light videos.
SDSD [60]	Paired videos	150 videos	Paired low/normal-light videos for dynamic scenes.
DRV [6]	Paired RAW videos	202 videos (up to 110 frames)	Static RAW Bayer sequences with long-exposure GT.
LoLi-Phone [35]	Unpaired videos	120 videos / 45,148 frames	Smartphone-captured low-light videos.
UbiVideo	Unpaired videos	120 videos	Self-collected mobile low-light videos across four motion scenarios.

deterministically initialize the diffusion noise under each runtime configuration. Across all platforms, videos are captured at 20 FPS, resized to 270×480, and processed frame by frame. For platforms with multiple compute backends (*e.g.*, CPU/GPU on Jetson), inference is pinned to a single backend (GPU-only) to avoid synchronization overhead and redundant memory transfers.

Datasets. We evaluate UbiEnlight on both public and self-collected low-light datasets, covering paired/unpaired supervision and diverse camera/object motion patterns Tab. 2 lists the details of these datasets.

- **Public datasets.** We use two paired low-light frame datasets, LSRW [18] and LOL-v2 [64], and four low-light video datasets, DID [13], SDSD [60], DRV [6], and LoLi-Phone [35]. LSRW/LOL-v2 provide large-scale paired frames with realistic illumination degradations, and are used to learn single-frame priors for exposure correction, color restoration, and denoising. DID is a paired video dataset which can be used to adapt UbiEnlight to temporal dynamics and motion-aware enhancement. SDSD and DRV are used only for full-reference testing, while LoLi-Phone is used only for real-world streaming evaluation.
- **Training pipeline.** UbiEnlight is trained in two stages. In Stage 1, we train the spectrum-prior branch on paired low/normal-light image datasets as an optional warm-start. Input images are resized/cropped to 256×256. We train with Adam (lr 1×10^{-4}, batch size 8). In Stage 2, we train the video enhancement model on fixed-length clips of $T = 24$ at 256×256. We train with AdamW (backbone lr 1×10^{-5}). During Stage 2, the diffusion objective and the auxiliary temporal/frequency/exposure consistency terms are optimized jointly within a unified objective. It supports both paired and unpaired video training.
- **Self-collected UbiVideo dataset.** We collect *UbiVideo* to evaluate robustness under realistic mobile capture behaviors. UbiVideo contains four scenarios: (A) moving camera + moving objects, (B) moving camera + static objects, (C) static camera + moving objects, and (D) static camera + static objects. Each includes 30 clips recorded on a Google Pixel 7 (120 videos total), each >30 s, captured in real low-light environments with varying exposure and motion. We primarily use UbiVideo to assess temporal stability and perceptual consistency under dynamic camera/object motion during continuous on-device enhancement.

Baselines. We compare UbiEnlight with twelve representative low-light enhancement methods.

Frame-wise (single-image) Baselines. They process videos frame-by-frame, providing strong per-frame quality/efficiency references but often inducing temporal flicker/jitter due to independent optimization.

- **Zero-DCE++** [36]: A **state-of-the-art** curve-based method that estimates **image-adaptive exposure curves** with (little or) no paired supervision, enabling lightweight deployment.
- **LLFlow** [61]: A **learning-based single-image** generative model that enhances low-light images by modeling their conditional distribution, applied **frame-by-frame** on videos.
- **URetinex-Net** [67]: A **learning-based single-image** Retinex network estimating illumination and reflectance via end-to-end architecture, applied frame-by-frame.
- **Diff-Retinex** [71]: An image-oriented diffusion model combining generative refinement with illumination-reflectance decomposition; it delivers **strong perceptual quality** but requires **iterative sampling** when used **frame-by-frame**.

Table 3. Capability checklist under two hardest real-world settings using usability-aware, no-reference constraints. A method is marked ✓ only if it satisfies the threshold for that constraint.

Hard Case	Constraint	Methods				
		Ours	SDSDNet	MBLLEN	StableLLVE	FastLLVE
Dynamic illumination + intense motion	Visibility (VPR@0.15 > 0.9)	✓(0.99)	×(0.58)	×(0.76)	✓(0.99)	✓(0.91)
	Detail (MGE > 0.07)	✓(0.11)	✓(0.07)	×(0.04)	×(0.05)	×(0.03)
	Stability (MF@0.15 < 0.05)	✓(0.02)	×(0.13)	×(0.06)	×(0.07)	×(0.09)
Extreme low-light + intense motion	Visibility (VPR@0.15 > 0.9)	✓(0.99)	×(0.53)	✓(0.99)	✓(0.99)	×(0.48)
	Detail (MGE > 0.07)	✓(0.10)	×(0.06)	×(0.03)	×(0.05)	✓(0.07)
	Stability (MF@0.15 < 0.05)	✓(0.01)	×(0.11)	✓(0.03)	✓(0.04)	×(0.10)

Notes: $\mathrm{VPR}_\tau = \frac{1}{|\Omega|}\sum_p \mathbf{1}(Y > \tau)$; $\mathrm{MGE} = \frac{1}{|\Omega|}\sum_p \|\nabla Y\|$; $\mathrm{MF}_\tau = \mathbb{E}_t[\,|Y_t - Y_{t-1}| \mid Y_t > \tau \vee Y_{t-1} > \tau\,]$. where Ω denotes the pixel domain, $|\Omega|$ its size, p a pixel index, Y_t luminance at time t, ∇Y spatial gradients, τ a fixed visibility threshold, and $\mathbb{E}_t[\cdot]$ temporal expectation.

- **FourierDiff** [46]: A **more recent** frequency-domain diffusion model leveraging Fourier priors; it offers **strong perceptual quality** but remains costly due to iterative sampling when applied **frame-by-frame**.

Frame-centric enhancement + post-hoc temporal modeling Baselines. They follow frame-centric pipelines augmented with explicit temporal constraints (e.g., flow-based warping, recurrent propagation, or short-term temporal modules) to suppress flicker.

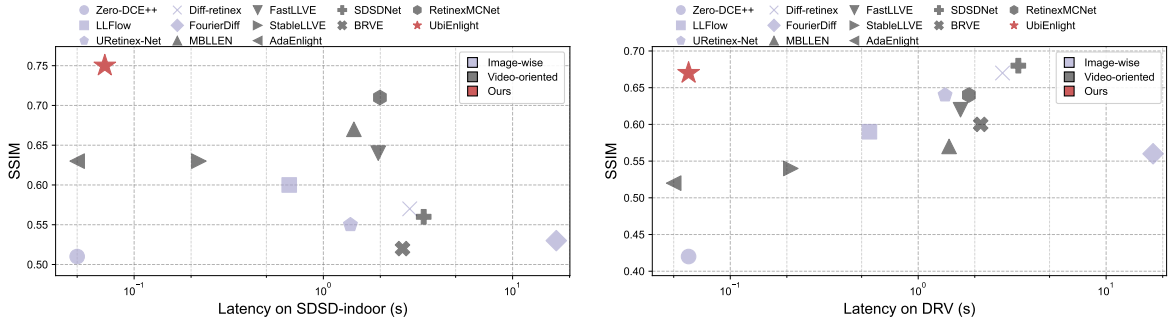
- **MBLLEN** [45]: A **classic** multi-branch network for video enhancement, fusing multi-scale features to recover details while maintaining temporal consistency (a representative **short-term temporal modules** design).
- **StableLLVE** [72]: A **state-of-the-art** model with temporal regularization that suppresses flickering and maintains motion coherence via **inter-frame consistency constraints**; it can be fragile when alignment becomes unreliable.
- **FastLLVE** [39]: A lightweight CNN optimized for real-time inference on commodity devices, balancing enhancement quality and computational efficiency with **short-term temporal modules**.
- **AdaEnlight** [40]: An adaptive high-fidelity framework designed for real-time mobile video enhancement, leveraging lightweight architectures and content-aware adaptation under resource constraints (with **short-term temporal modules** for stabilization).

Video-native Baselines. They optimize videos directly with video-centric data, models, and objectives, typically relying on paired dynamic-video supervision.

- **SDSDNet** [60]: A **more recent** static-to-dynamic supervised framework trained with **paired dynamic-video supervision** on paired low/normal-light video data, using spatiotemporal feature fusion for dynamic scene enhancement.
- **RetinexMCNet** [59]: A **recent** Retinex-based supervised LLVE method that combines illumination-reflectance enhancement with a memory controller for temporal consistency.
- **BRVE** [74]: A **recent** binarized raw-domain LLVE method that uses distribution-aware binary convolutions and spatial-temporal shift operations for efficient temporal enhancement.
- **UbiEnlight W/O_Align**: A variant of UbiEnlight without the second-stage Diffusion Transformer, reported to isolate the contribution of our video-native design.
- **UbiEnlight W/O_Spectrum Prior**: A variant of UbiEnlight that removes the amplitude and phase spectrum priors, reported to isolate the contribution of our spectrum-guided conditioning design.

5.2 Performance Comparison

We evaluate UbiEnlight on the SDSD-indoor and DRV benchmarks against frame-wise, temporal-aware, and recent video-native methods (e.g., BRVE and RetinexMCNet) under a unified deployment-oriented protocol. Besides quality and efficiency, we compare whether methods require paired supervision or explicit motion



(a) Quality-latency trade-off on SDSD

(b) Quality-latency usage trade-off on DRV

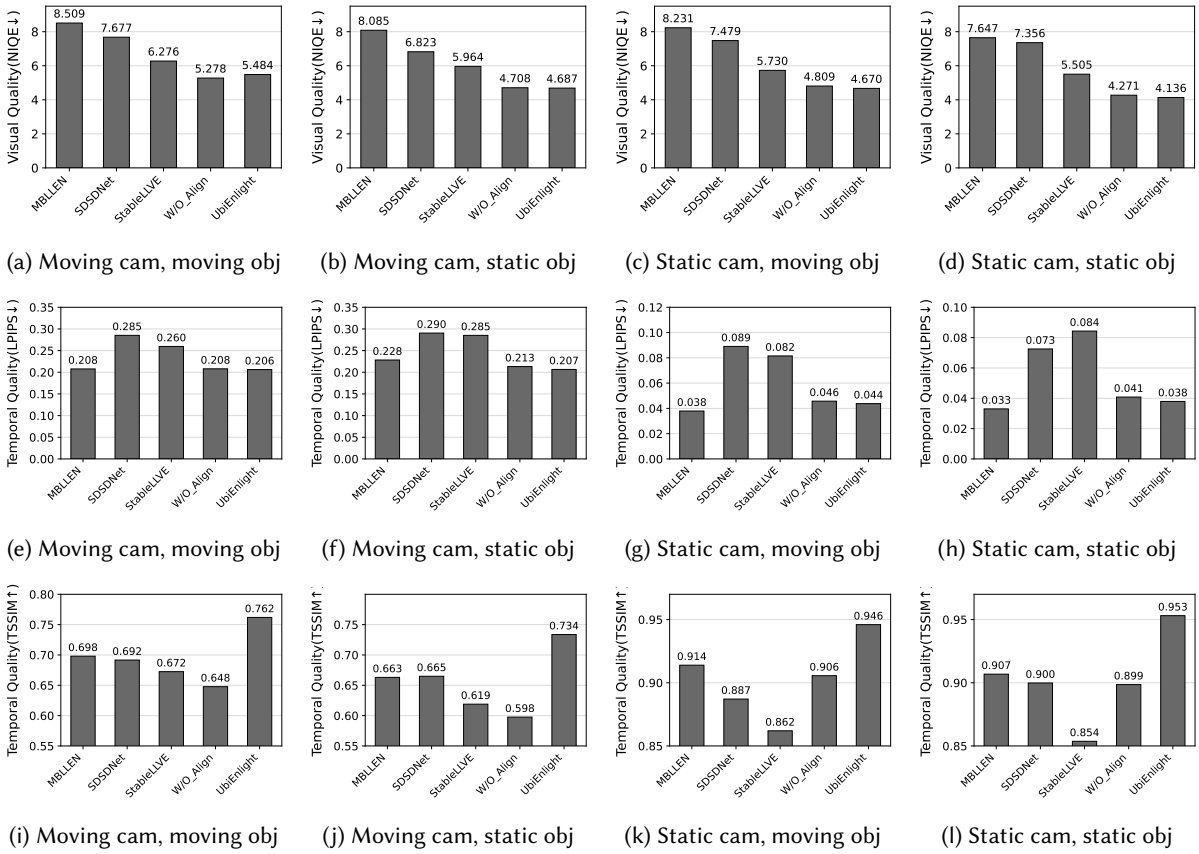
Fig. 8. Performance comparison of existing representative low-light video enhancement methods.

Table 4. Performance comparison of UbiEnlight with representative baselines in terms of visual quality (PSNR/SSIM, ↑), temporal consistency (TSSIM, ↑), and per-frame latency (s, ↓). We also report whether a method requires paired supervision and extra motion modeling.

Methods		No paired video supervision	No extra motion modeling	TSSIM ↑	SDSD-indoor [60]			DRV [6]		
					PSNR ↑	SSIM ↑	Latency (s) ↓	PSNR ↑	SSIM ↑	Latency (s) ↓
Frame-wise (single-image) methods	**Zero-DCE++** [36]	✓	✓	–	14.54	0.51	0.05	14.47	0.42	0.06
	LLFlow [61]	✓	✓	–	16.13	0.60	0.66	14.06	0.59	0.55
	URetinex-Net [67]	✓	✓	–	15.00	0.55	1.39	15.43	0.64	1.39
	Diff-retinex [71]	✓	✓	–	14.36	0.57	2.86	16.40	0.67	2.81
	FourierDiff [46]	✓	✓	–	14.18	0.53	17.02	14.77	0.56	17.78
Frame-wise + post-hoc temporal modeling	**MBLLEN** [45]	×	✓	0.79	14.52	0.67	1.45	14.90	0.57	1.46
	FastLLVE [39]	×	✓	0.68	14.00	0.64	1.95	18.05	0.62	1.68
	StableLLVE [72]	×	×	0.75	15.33	0.63	0.22	11.78	0.54	0.21
	AdaEnlight [40]	✓	×	0.79	15.52	0.63	0.05	15.23	0.52	0.05
Video-native methods	**SDSDNet** [60]	×	✓	0.78	21.26	0.56	3.39	17.41	0.68	3.41
	BRVE [74]	×	✓	0.76	15.86	0.52	2.62	17.12	0.60	2.15
	RetinexMCNet [59]	×	✓	0.81	20.84	0.71	1.99	18.62	0.64	1.86
	UbiEnlight	✓	✓	**0.86**	16.22	0.75	0.07	19.34	0.67	0.06

modeling (*e.g.*, optical flow). Metrics include PSNR/SSIM, TSSIM, and per-frame latency on Jetson AGX Orin, together with no-reference usability metrics (VPR@τ, MGE, and MF@τ) for challenging scenarios.

Results are summarized in Tab. 4. *First,* UbiEnlight achieves an excellent quality-efficiency trade-off, running at only 0.06–0.07 s/frame while maintaining competitive reconstruction quality (Fig. 8). It achieves the best SSIM (0.75) on SDSD-indoor, strong PSNR (19.34) on DRV, and the highest temporal stability (TSSIM 0.86), indicating robust structural recovery with reduced flicker. PSNR measures pixel-wise fidelity to paired references, whereas SSIM emphasizes structural similarity. Since UbiEnlight is designed for spectrum-guided structural preservation and temporal consistency rather than strict pixel-level reconstruction, it achieves stronger SSIM/TSSIM while showing slightly lower PSNR than fully supervised video methods trained with paired ground truth. *Second,* UbiEnlight achieves this quality–stability trade-off without paired video supervision in the core enhancement stage or explicit motion modeling, unlike several video-native baselines. This avoids costly alignment and reduces dependence on scarce paired video data, making UbiEnlight more practical for mobile deployment. Compared with high-quality diffusion baselines, UbiEnlight is over 20× faster than FourierDiff and over 25× faster than SDSDNet on DRV. Compared with real-time methods, it improves PSNR by 4.87 over Zero-DCE++ while maintaining comparable latency. We further evaluate two challenging scenarios: (i) dynamic illumination with intense user motion and (ii) extreme low light with intense user motion. As shown in Tab. 3, UbiEnlight is the only method satisfying all usability criteria (visibility, detail recovery, and temporal stability) in both scenarios, demonstrating reliable real-time enhancement under severe conditions.



(a) Moving cam, moving obj (b) Moving cam, static obj (c) Static cam, moving obj (d) Static cam, static obj

(e) Moving cam, moving obj (f) Moving cam, static obj (g) Static cam, moving obj (h) Static cam, static obj

(i) Moving cam, moving obj (j) Moving cam, static obj (k) Static cam, moving obj (l) Static cam, static obj

Fig. 9. Temporal stability quantified by NIQE, tLPIPS, and TSSIM, in four mobile scenarios. Scenario A: moving camera and moving objects {a,e,i}, Scenario B: moving camera and static objects {b,f,j}, Scenario C: static camera and moving objects {c,g,k}, and Scenario D: static camera and static objects {d,h,l}. A low NIQE and tLPIPS and a high TSSIM indicate good perceptual quality and temporal stability.

Summary. UbiEnlight offers the best overall balance of visual quality, temporal stability, efficiency, and deployability for real-time mobile and wearable video enhancement.

5.3 System Performance under Real-world Diversity and Dynamics

We evaluate UbiEnlight at the system level, focusing on end-to-end robustness and efficiency under diverse mobile conditions.

5.3.1 Performance under Diverse Camera/Object Motion. We evaluate UbiEnlight on our self-collected *UbiVideo* dataset, covering four representative mobile scenarios with different combinations of camera and object motion. We report NIQE (lower is better) for perceptual quality and temporal LPIPS / TSSIM (lower / higher is better) for temporal consistency. Results are shown in Fig. 9. *First,* UbiEnlight maintains strong perceptual quality, achieving the lowest NIQE across all four scenarios (Fig. 9a–9d). Although **W/O_Align** is marginally better in a few cases, the difference is small. *Second,* UbiEnlight consistently improves temporal consistency. It achieves the lowest temporal LPIPS in all moving-camera scenarios and remains best or near-best in static-camera settings

Table 5. UbiEnlight's performance under diverse mobile energy budgets with the CPU on device 1.

No.	Diverse energy supply	Power (W)	Energy per frame (mJ)	NIQE (↓)	Latency per frame (ms)	Optimal Hyperparameters		
						θ_s	θ_a	θ_f
1	20%	3	312.76	5.2841	76.47	2	1	0
2	43%	3.2	331.78	5.2607	103.68	6	1/3	0
3	68%	3.5	420.18	5.2558	120.05	10	1/2	1/5
4	74%	3.7	472.60	5.2368	127.73	4	1/2	3/5
5	96%	4.2	639.88	5.1859	159.97	10	1/5	1

Table 6. UbiEnlight's performance under diverse mobile energy budgets with the GPU on device 2.

No.	Diverse energy supply	Power (W)	Energy per frame (J)	NIQE (↓)	Latency per frame (ms)	Optimal hyperparameters		
						θ_s	θ_a	θ_f
1	18%	6.21	0.8379	4.7725	66.16	4	3/4	0
2	33%	6.54	0.8473	4.7708	68.33	10	1/5	1/5
3	58%	6.79	0.8691	4.7561	70.49	16	1/4	1/5
4	78%	6.84	0.8951	4.7530	73.31	12	1/4	1/2
5	94%	6.88	0.9144	4.7509	75.51	18	1/3	1

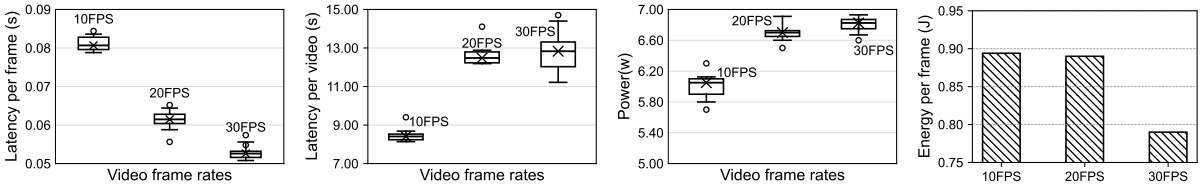
(Fig. 9e–Fig. 9h), indicating reduced flicker and perceptual jitter. *Third,* UbiEnlight achieves the highest TSSIM across all scenarios (Fig. 9i–Fig. 9l). Compared with **W/O_Align**, it consistently improves both temporal LPIPS and TSSIM, demonstrating the effectiveness of the Spectrum-Conditioned Diffusion Transformer in stabilizing streaming enhancement.

Summary. Across diverse camera and object motion patterns, UbiEnlight provides the best overall balance of perceptual quality and temporal stability, enabling robust real-time low-light video enhancement.

5.3.2 Performance over Dynamic Energy Supply. We evaluate UbiEnlight under dynamic energy constraints during continuous streaming. According to the current battery level, the controller adaptively adjusts the sampling-step budget (θ_s), attention reuse ratio (θ_a), and keyframe refresh rate (θ_f) to balance enhancement quality and system cost. Experiments are conducted on UbiVideo using Google Pixel 7 (CPU, device 1) and Jetson AGX Orin (GPU, device 2) with battery levels ranging from ∼20% to 90%. We report NIQE (lower is better), latency/frame, and energy/frame. Results are summarized in Tab. 5 and Tab. 6. As energy becomes available, UbiEnlight progressively allocates more computation by increasing the denoising budget (θ_s), reducing attention reuse (θ_a), and refreshing keyframes more frequently (θ_f). For example, on the GPU, θ_s increases from 4 (18%) to 18 (94%), θ_a decreases from 3/4 to 1/3, and θ_f increases from 0 to 1, enabling higher reconstruction quality. Similar adaptation is observed on the CPU, where θ_s increases from 2 (20%) to 10 (96%).

Summary. Across both platforms, UbiEnlight continuously adapts ($\theta_s, \theta_a, \theta_f$) according to battery availability, improving visual quality when energy is sufficient while maintaining stable latency and energy efficiency. In our continuous-use setting on Google Pixel 7, UbiEnlight depletes the battery in 2∼4.5 hours. This duration is specific to the phone-class continuous-use evaluation and should not be interpreted as a uniform result across different platforms.

5.3.3 Performance across Diverse Video Frame Rates. We distinguish *input frame rate* from *enhancement throughput*. The former is the source video rate, whereas the latter is the number of enhanced frames produced per second under the selected runtime configuration, which may be lower than the input rate on resource-constrained devices. We evaluate whether UbiEnlight maintains responsive streaming enhancement under varying input frame rates (10, 20, and 30 FPS), which commonly arise from different camera settings, scene complexity, and runtime contention. Experiments are conducted on 20 LoLi-Phone clips (10 s each) using Jetson AGX Orin (GPU,



(a) Latency per frame (b) Latency per video (c) Power (d) Energy cost per frame

Fig. 10. UbiEnlight 's performance across diverse inputs with varying video frame rates (FPS), *i.e.*, 10 FPS, 20FPS, and 30FPS.

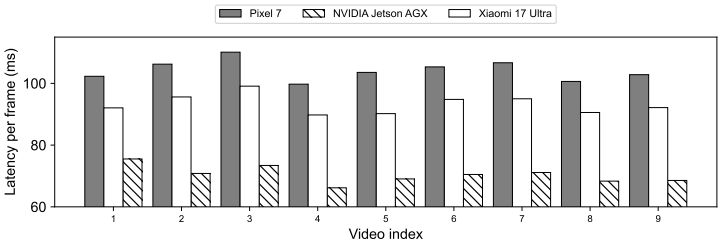


Fig. 11. Performance across three typical camera-embedded platforms.

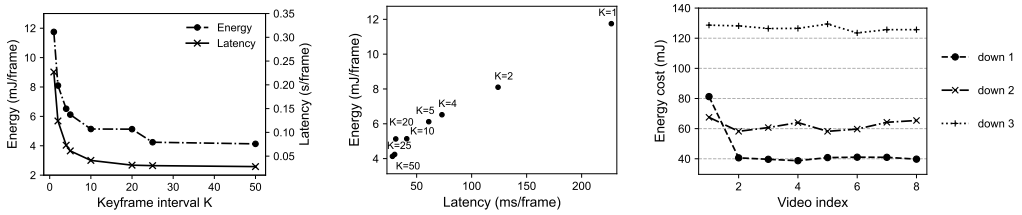
device 2). In adaptive mode, UbiEnlight dynamically adjusts sampling steps, attention reuse, and frame caching according to the input rate. We report per-frame latency, clip-level processing time, power, and energy/frame (Fig. 10). Results show that UbiEnlight scales efficiently with increasing input rates. As FPS increases from 10 to 30, per-frame latency decreases from 0.081 s to 0.053 s, indicating more aggressive computation reuse under higher temporal redundancy. Although clip-level processing time increases with more input frames, its growth remains sublinear (Fig. 10b). Meanwhile, power rises only modestly (5.9 W→6.8 W), and the 30 FPS setting achieves the lowest energy/frame (Fig. 10d).

Summary. UbiEnlight adapts effectively to diverse input frame rates, sustaining responsive streaming while reducing per-frame latency and energy/frame through adaptive computation reuse.

5.3.4 Performance on Diverse Mobile Platforms. We evaluate UbiEnlight on three representative camera-enabled platforms with different compute budgets and deployment stacks: Google Pixel 7 (CPU; NNAPI, device 1), NVIDIA Jetson AGX Orin (GPU; PyTorch, device 2), and Xiaomi 17 Ultra (CPU; PyTorch Mobile, device 3). Using eight 30 s UbiVideo clips, we measure per-frame enhancement latency (Fig. 11). Across all platforms, UbiEnlight consistently delivers low-latency diffusion enhancement. On the phone-class platforms (Pixel 7 and Xiaomi 17 Ultra), selected enhanced frames are processed in 90∼100 ms, while Jetson AGX Orin further reduces latency to ∼ 66 ms, enabling higher effective throughput. Here, "near-real-time" refers to maintaining a bounded end-to-end display latency through the adaptive runtime controller, rather than enhancing every input frame at the raw input frame rate. These results demonstrate stable and practical deployment across heterogeneous mobile and edge platforms.

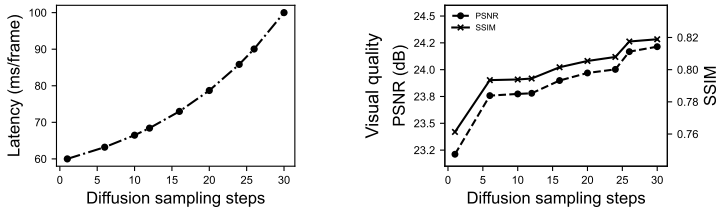
5.4 Ablation Studies

Unless otherwise stated, each ablation is conducted on the platform that best exposes the effect of the mechanism under study. We use Jetson AGX Orin (GPU;device 2) for ablations that require stable measurement over a wider quality–latency operating range, and Google Pixel 7 (CPU;device 1)) for ablations that emphasize lightweight runtime efficiency on a practical mobile device. Therefore, these studies are intended to validate the trend of each control knob rather than to compare absolute runtime values across different hardware platforms.



(a) keyframe interval K (b) Latency–energy tradeoff (c) Impact of frame resolution

Fig. 12. Micro-benchmark of key system parameters.



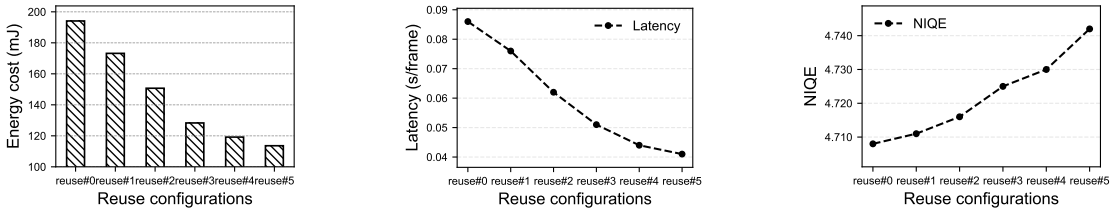
(a) Latency *vs.* sampling steps (b) Quality *vs.* sampling steps

Fig. 13. Benchmark of diffusion sampling steps on system performance.

5.4.1 Impact of Keyframe Selection and Frame Caching. We test the efficiency of UbiEnlight's keyframe selection and frame caching. Using LoLi-Phone, we vary the keyframe interval K from enhancing every frame (K=1) to enhancing one keyframe every K frames, while keeping all other settings fixed. Experiments are conducted on Jetson AGX Orin (GPU; device 2), reporting per-frame latency and energy (Fig. 12). As shown in Fig. 12a, increasing K substantially reduces both latency and energy, with the largest gains at small intervals before gradually saturating. The latency–energy trade-off (Fig. 12b) further shows that intermediate intervals (*e.g.*, K=10–50) achieve most of the efficiency benefits, while larger intervals provide only marginal improvements. **Summary.** Keyframe selection and frame caching provide an effective runtime knob for reducing latency and energy while preserving streaming efficiency.

5.4.2 Impact of Frame Resolution. We study how input resolution affects UbiEnlight 's on-device energy cost on Google Pixel 7 (CPU; device 1) using Ubivideo. Fig. 12c reports per-frame energy under three settings: full resolution, down-sampled to 1/2, and down-sampled to 1/3. Results show a monotonic trend, *i.e.*, higher resolutions consistently incur higher energy across all clips, while down-sampling substantially reduces per-frame energy. **Summary**. This indicates that resolution scaling provides a simple, effective knob to tune UbiEnlight 's energy demand under resource constraints.

5.4.3 Impact of Diffusion Sampling Steps. We evaluate the impact of diffusion sampling steps on the quality–latency trade-off of UbiEnlight. Using DID, we vary the sampling-step budget while keeping all other components fixed, and measure per-frame latency and PSNR/SSIM on NVIDIA Jetson AGX Orin (GPU; device 2). Results are shown in Fig. 13. Latency increases monotonically with the number of sampling steps (Fig. 13a), from ∼60 ms/frame to ∼100 ms/frame, with a steeper increase at higher step counts. Meanwhile, reconstruction quality improves with diminishing returns (Fig. 13b): PSNR increases from 23 dB to 24 dB and SSIM from 0.7 to ∼0.8, while



(a) Energy vs. attention reuse (b) Latency vs. attention reuse (c) NIQE vs. attention reuse

Fig. 14. Benchmark of attention reuse on UbiEnlight's efficiency and visual quality, *i.e.*, (a) per-frame energy cost, (b) per-frame latency, and (c) visual quality under different reuse configurations.

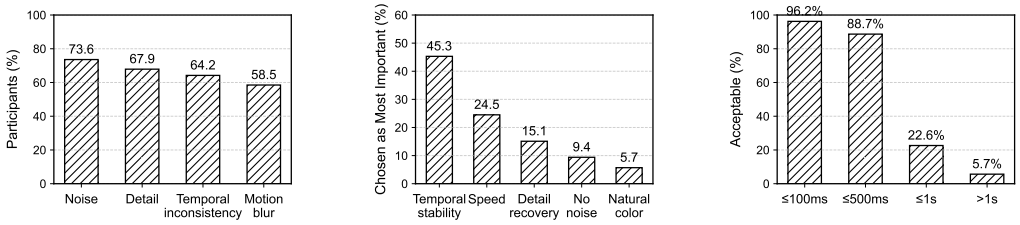
Table 7. Ablation study of the diffusion-stage design under two representative scenarios.

Method	Moving cam., moving obj.		Static cam., static obj.	
	NIQE↓	TSSIM↑	NIQE↓	TSSIM↑
UbiEnlight w/o Spectrum Prior	9.214	0.621	8.736	0.772
UbiEnlight w/o Align	5.278	0.648	4.271	0.899
UbiEnlight	5.484	0.762	4.136	0.953

the incremental gains gradually diminish. **Summary.** Sampling steps provide an effective quality-latency control knob. We therefore adopt 10~20 steps by default to balance reconstruction quality and real-time throughput.

5.4.4 Impact of Attention Reuse. We evaluate attention reuse as a way to remove redundant computation across similar neighboring diffusion steps and frames. On Google Pixel 7 (CPU; device 1), we fix the diffusion budget to 20 steps and compare six policies: *reuse#0* recomputes attention at every step, while *reuse#1–#5* reuse cached attention outputs every 2/4/6/8/10 steps. We also report NIQE on LoLi-Phone to measure perceptual quality. As shown in Fig. 14, attention reuse improves efficiency with little quality loss. Energy decreases from 194.1 mJ to 113.6 mJ per frame, achieving a 41.5% reduction. Latency drops from 86 ms/frame to 41 ms/frame, yielding a 52.3% reduction and a 2.1× speedup. Meanwhile, NIQE changes only slightly from 4.708 to 4.742, suggesting negligible perceptual degradation. **Summary.** Attention reuse removes redundant diffusion computation and substantially improves on-device latency and energy efficiency while preserving visual quality.

5.4.5 Impact of Spectrum Conditioning. We evaluate the effect of spectrum conditioning by comparing the full model with two variants under two representative scenarios: moving camera with moving objects and static camera with static objects. Specifically, *UbiEnlight w/o Spectrum Prior* removes the amplitude and phase priors, while *UbiEnlight w/o Align* removes the second-stage Diffusion Transformer. As shown in Table 7, removing the spectrum prior causes the largest performance degradation, increasing NIQE to 9.214/8.736 and reducing TSSIM to 0.621/0.772 across the two scenarios. This indicates that spectrum conditioning provides stable guidance for diffusion by jointly leveraging amplitude cues for illumination adjustment and phase cues for structural anchoring, rather than relying on diffusion alone. Removing the second-stage Diffusion Transformer mainly reduces temporal stability, while the full model achieves the highest TSSIM (0.762/0.953) and the best NIQE in the static scenario. **Summary.** Spectrum conditioning provides stable guidance for diffusion-based low-light video enhancement, with amplitude improving illumination restoration and phase preserving structural consistency under motion and illumination changes.



(a) Perceived degradations (b) Most important factor (c) Latency acceptance

Fig. 15. User study results with 53 participants.

Table 8. Subjective evaluation of three operating modes.

Mode	Temporal Score	Responsiveness Score	Detail Score	Perceptual Utility U_{pref}	Preference Ratio
Fixed High-Quality	4.16	2.23	**4.68**	3.37	5/20 (25.0%)
Fixed Low-Cost	2.84	**4.51**	2.37	2.96	2/20 (10.0%)
Adaptive Mode	**4.57**	4.42	4.19	**3.92**	**13/20 (65.0%)**

5.5 User Study

We conduct a two-part user study to understand user requirements for low-light mobile video enhancement and evaluate whether UbiEnlight's adaptive design improves perceived usability. The study includes (i) a 53-participant survey on user needs and (ii) a 20-participant comparison of three operating modes.

User requirements. We first surveyed 53 participants (age 21–68; 29 male/24 female) who regularly capture or view low-light videos on mobile or wearable devices. The survey covered usage scenarios, perceived degradations, and preferences regarding latency, energy consumption, and adaptive behavior. Results (Fig. 15) show that low-light video enhancement is both common and safety-critical. Participants frequently encountered low-light scenarios, including night commuting (69.8%), home monitoring (60.4%), and outdoor activities (52.8%), while reporting visible noise (73.6%), detail loss (67.9%), temporal inconsistency (64.2%), and motion blur (58.5%). These degradations affected usability (90.6%), safety (84.9%), and viewing comfort (81.1%). Users consistently prioritized temporal stability (45.3%) over latency (24.5%) and detail (15.1%); 71.7% considered flicker more objectionable than minor frame artifacts, 96.2% accepted delays below 100 ms, and 79.2% preferred on-device enhancement for responsiveness and privacy. These findings directly motivate the user-centric adaptive design of UbiEnlight. Visual examples (Fig. 16) further illustrate that UbiEnlight improves dark-region visibility while maintaining stable illumination and reducing flicker across consecutive frames.

Adaptive mode evaluation. We further recruited 20 participants (age 22–65; 11 male/9 female) from the original cohort to compare three operating modes: *Fixed High-Quality*, *Fixed Low-Cost*, and *Adaptive Mode*. Participants evaluated 12 representative low-light clips (5–8 s) covering dynamic illumination, fast motion, handheld shake, and static high-detail scenes. To reduce ordering bias, both video clips and operating modes were randomized and counterbalanced across participants. As this study aims to validate perceptual benefits rather than provide a fully powered statistical analysis, we report descriptive subjective scores, preference ratios, and normalized perceptual utility. Participants rated temporal stability, responsiveness, detail preservation, and overall preference on five-point scales. Based on the preference survey, we define a preference-weighted perceptual utility $U_{\text{pref}} = 0.453S + 0.245R + 0.151D$, where S, R, and D denote normalized ratings for temporal stability, responsiveness, and detail preservation. We further define the unit-energy perceptual utility UVE $= \frac{U_{\text{pref}}}{\text{Energy}}$ to measure perceived utility per unit energy. Table 8 summarizes the results. The *Adaptive Mode* achieves the highest overall utility ($U_{\text{pref}} = 3.92$), outperforming *Fixed High-Quality* (3.37) and *Fixed Low-Cost* (2.96), while receiving

Fig. 16. **Qualitative results on continuous low-light video frames.** UbiEnlight improves dark-region visibility and text details at Frame t, t+3, and t+6, while maintaining temporally consistent illumination.

the highest user preference (65.0%). Using measured energy on Google Pixel 7, it also improves UVE by 61.3% and 32.7% over the two fixed modes, respectively.

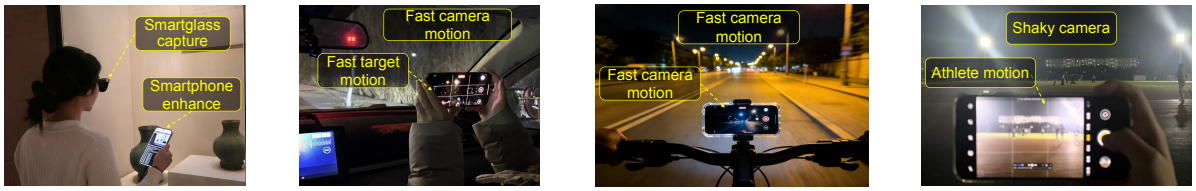
Summary. The results show that UbiEnlight adaptively allocates computation toward the aspects users value most, *e.g.*, temporal stability, responsiveness, and usable detail, thereby improving perceived quality without sacrificing energy efficiency.

5.6 Real-world Case Study

To stress-test UbiEnlight under real-world low-light videos with mixed, non-stationary illumination and motion, we conduct in-the-wild case studies on mobile and wearable devices. As shown in Fig. 17, we curate four representative scenarios with increasing temporal complexity, each recorded as a 2-hour continuous video to evaluate long-term temporal stability and online adaptation. Fig. 18 presents qualitative results, showing that UbiEnlight consistently improves dark-region visibility while preserving structural consistency across diverse mobile and wearable low-light scenes. Together, these scenarios cover representative challenges in everyday use, including illumination variation, abrupt scene transitions, and camera motion.

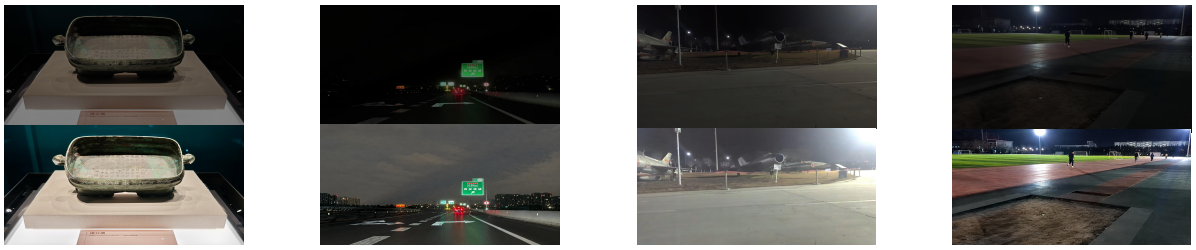
- **Museum/Heritage Smartglass Capture (wearable, perception-sensitive).** We capture dim, no-flash exhibits using XREAL One Pro AR glasses with smartphone-assisted enhancement (Xiaomi 17 Ultra), emulating a night-blind user wearing smart glasses. Repeated *stop-pan-stop* motion makes flicker and exposure drift readily perceptible on near-eye displays.
- **Parking Garage Entrance/Exit (abrupt illumination transitions).** Using a Pixel 7, we record continuous entry and exit sequences with rapid dark–bright–dark transitions caused by lamps, reflective surfaces, and headlights, challenging exposure consistency.
- **Night E-bike Riding (fast motion + dynamic lighting).** A smartphone mounted on an e-bike records night scenes with rapidly changing illumination, vibration, rolling-shutter artifacts, and motion blur, representing a demanding real-time enhancement scenario.
- **Handheld Walking/Running (camera shake).** We record handheld outdoor videos under low light, where strong camera shake and stable illumination isolate motion-induced flicker and structural drift.

We evaluate time-series trends of temporal consistency (TSSIM), no-reference visual quality (NIQE), and per-frame latency (Fig. 19). *First,* UbiEnlight maintains *consistent visual quality* across diverse dynamics: NIQE stays within a narrow band over time for each scenario (Fig. 19b). *Second,* UbiEnlight achieves *stable temporal consistency* over the full 2-hour streams: TSSIM curves are largely flat with only small fluctuations, indicating suppressed flicker and no long-term drift (Fig. 19a). *Third,* UbiEnlight sustains *bounded real-time latency* without systematic growth, suggesting robust online adaptation that does not destabilize throughput (Fig. 19c). Unlike the controlled battery-budget experiment in Section 5.3.2, this latency trend reflects real in-the-wild runtime variation rather than battery level alone. Latency may increase in harder segments with fast motion, abrupt illumination changes, or fewer reuse/caching opportunities, where the controller allocates more computation to



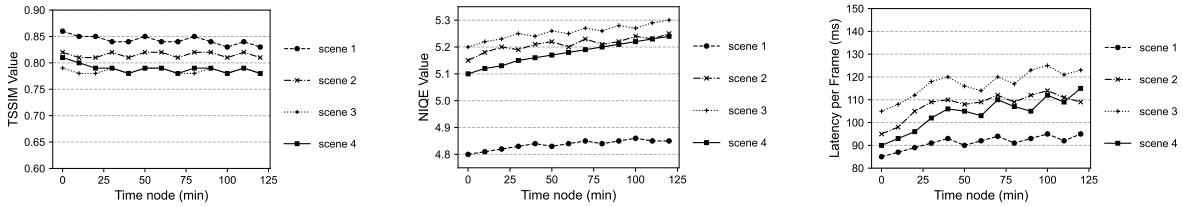
(a) Museum/Night-blind (b) In-vehicle capture (c) Night e-bike (d) Handheld walk/run

Fig. 17. Case study scenarios for real-world low-light videos with non-stationary illumination and motion.



(a) Museum (b) In-vehicle (c) E-bike (d) Walk/run

Fig. 18. Qualitative enhancement results for the four real-world case-study scenarios in Fig. 17. Each example compares the low-light input and the corresponding output, demonstrating improved visibility and structural consistency across museum/smartglass, in-vehicle, night e-bike, and handheld walk/run scenarios.



(a) Temporal consistency over time (b) Perceptual quality over time (c) Per-frame latency over time

Fig. 19. Long-stream stability in in-the-wild low-light videos. Time-series of TSSIM, NIQE, and per-frame latency over *2-hour* streams across four cases with non-stationary mixed illumination, fast/shaky motion, and extreme low light; the smartglasses case (scene 1) highlights benefits for users *with night-blindness*.

preserve stability and perceptual quality. Thus, the trend does not contradict the battery-level experiment; it shows that long-running performance is jointly shaped by scene dynamics, adaptation decisions, and system conditions. **Summary.** Across four 2-hour in-the-wild scenarios, UbiEnlight maintains high visual quality, stable temporal consistency, and bounded real-time latency under mixed illumination, fast motion, shaky motion, and extreme low-light. The smartglasses case study further suggests practical benefits for users with night blindness.

6 RELATED WORK

6.1 User-Centric Mobile Visual Sensing Application

Smartphones and wearables have become *video-centric* sensing platforms for human-environment interaction, enabled by high-quality cameras, connectivity, and on-device AI. Prior work mainly spans *media creation* (photography/short video [20, 66]), *interactive perception* (video communication [25, 31], AR navigation [10, 14,

33]), and *safety/assistive sensing* (assistive perception [26, 37], mobile robotics [3, 11], smart-home monitoring [4, 5]). These applications run over *long streams*, are directly perceived by users, and must meet strict on-device latency/energy budgets. UbiEnlight targets robust long-duration processing by explicitly modeling perceptual tolerance, temporal stability, and runtime variability.

6.2 Low-light Image/Video Enhancement

We group prior work into **three types** and position UbiEnlight accordingly. **First, frame-centric enhancement.** Videos are processed frame-by-frame using physical priors (*e.g.*, Retinex [32], 1977) or learning-based single-image models [17, 36, 53, 54, 58, 64, 67, 76]. Curve-based methods (Zero-DCE [17] and its accelerated variant Zero-DCE++ [36]) estimate image-adaptive exposure curves with (little or) no paired supervision, enabling lightweight deployment. These methods often deliver strong per-frame fidelity with lightweight models [36, 45, 55], but independent optimization readily induces flicker/jitter and motion-related artifacts. **Second, frame-centric enhancement + post-hoc temporal modeling.** Frame-centric pipelines are augmented with explicit temporal constraints (*e.g.*, flow-based warping, recurrent propagation, or short-term temporal modules) to suppress flicker. A representative milestone is [73] (2021), which introduces inter-frame consistency constraints to learn temporal stability for low-light video enhancement even from *single-image* training data. While such constraints improve coherence with minimal changes to image backbones, they are *fragile* when alignment becomes unreliable (low SNR, blur, fast motion) and incur high latency/energy overhead. **Third, video-native spatiotemporal restoration.** Videos are optimized *directly* with video-centric data, models, and objectives, evolving from *(III-A)* paired dynamic-video supervision [29, 60], to *(III-B)* pushing stronger spatiotemporal guidance [13], and further to *(III-C)* unpaired/generative learning [81]. Diffusion models [23, 46, 62, 71] belong to *Stage III-C* (generative priors): strong perceptual quality, but typically too heavy for real-time on-device use due to iterative sampling. By exploiting cross-frame redundancy, video-native methods make temporal stability largely an *endogenous* property; however, III-C must still balance consistency with efficiency for real-time deployment. This taxonomy clarifies two key questions:

- *Q1: why video-native beats frame-wise + post-hoc consistency.* Low-light degradation and motion/illumination changes are *spatiotemporally coupled*, so enforcing consistency *after* independent per-frame restoration often leaves residual flicker and motion artifacts. Video-native methods restore jointly across frames, turning consistency from an external constraint into an *endogenous* outcome.
- *Q2: our leap within Stage III-C.* We move beyond reliance on paired dynamic-video supervision (costly, narrow coverage, and weaker generalization under dynamic lighting/fast motion) toward *unpaired/generative* video-native learning that still yields *stable* spatiotemporal enhancement in real scenes, while remaining *resource-aware* on ubiquitous devices.

6.3 Adaptive Vision Tasks on Mobile Devices

Mobile vision increasingly adopts *adaptive computation* to balance perceptual quality with latency and energy under dynamic runtime conditions. Prior work mainly falls into two lines. *Architectural adaptivity* (*e.g.*, early-exit [57, 68], MoE [12, 70]) skips computation via conditional execution, but is often *runtime-agnostic* (battery/thermals/contention), limiting long-duration mobile use. *System-level adaptivity* (*e.g.*, scheduling [69], content-aware control [9]) adjusts runtime execution, but typically treats models as black boxes and lacks *perceptual-quality control.* UbiEnlight closes this gap by coupling *human perceptual tolerance* with *runtime feedback* to scale enhancement cost and sustain always-on mobile video processing.

7 CONCLUSION

We presented UbiEnlight, a real-time on-device low-light video enhancement system that makes diffusion-based enhancement practical for mobile interactive vision. Rather than pursuing visual quality alone, UbiEnlight jointly optimizes enhancement fidelity, temporal stability, and interactive responsiveness through a perception-aware and resource-aware design. By tightly coupling diffusion inference with runtime resource adaptation, UbiEnlight enables stable real-time enhancement under highly dynamic illumination and user motion on commodity smartphones and smartglasses. More broadly, this work demonstrates that diffusion models can become practical building blocks for mobile interactive vision when algorithm design is co-optimized with user perception and system constraints. More effort and insight are needed in perception-driven scheduling, adaptive generative inference, and resource-aware foundation models for always-on mobile intelligence.

ACKNOWLEDGMENTS

This work was partially supported by the National Natural Science Foundation of China (62522215, 62532009, 6247235462272368) and Key Talent Project of Xidian University (No.QTZX24004). Zimu Zhou's research is supported by CityU APRC grant (No. 9610633). The authors thank the anonymous reviewers for their constructive feedback that has made the work stronger.

REFERENCES

[1] Aswir Aswir, Muhammad Sofian Hadi, and Fatimah Rosiana Dewi. 2021. Google meet application as an online learning media for descriptive text material. *Jurnal studi guru dan pembelajaran* 4, 1 (2021), 189–194.

[2] Oshri Bar-Gil. 2020. Clipping us together: The case of the Google Clips camera. (2020).

[3] Ryan M Bena, Chongbo Zhao, and Quan Nguyen. 2023. Safety-aware perception for autonomous collision avoidance in dynamic environments. *IEEE Robotics and Automation Letters* 8, 12 (2023), 7962–7969.

[4] Marco Buzzelli, Alessio Albé, and Gianluigi Ciocca. 2020. A vision-based system for monitoring elderly people at home. *Applied Sciences* 10, 1 (2020), 374.

[5] Alexandros Andre Chaaraoui, José Ramón Padilla-López, Francisco Javier Ferrández-Pastor, Mario Nieto-Hidalgo, and Francisco Flórez-Revuelta. 2014. A vision-based system for intelligent monitoring: Human behaviour analysis and privacy by context. *Sensors* 14, 5 (2014), 8895–8925.

[6] Chen Chen, Qifeng Chen, Minh N Do, and Vladlen Koltun. 2019. Seeing motion in the dark. In *Proceedings of the IEEE/CVF International conference on computer vision*. 3185–3194.

[7] Chen Chen, Qifeng Chen, Jia Xu, and Vladlen Koltun. 2018. Learning to see in the dark. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3291–3300.

[8] Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, et al. 2023. Pixart-\alpha:Fast training of diffusion transformer for photorealistic text-to-image synthesis. *arXiv preprint arXiv:2310.00426* (2023).

[9] Ting-Wu Chin, Ruizhou Ding, and Diana Marculescu. 2019. Adascale: Towards real-time video object detection using adaptive scaling. *Proceedings of machine learning and systems* 1 (2019), 431–441.

[10] Biqin Deng. 2024. The application of AI video generation technology in virtual reality (VR) and augmented reality (AR). In *2024 International Conference on Artificial Intelligence, Deep Learning and Neural Networks (AIDLNN)*. IEEE, 168–173.

[11] Davide Falanga, Suseong Kim, and Davide Scaramuzza. 2019. How fast is too fast? the role of perception latency in high-speed sense and avoid. *IEEE Robotics and Automation Letters* 4, 2 (2019), 1884–1891.

[12] William Fedus, Barret Zoph, and Noam Shazeer. 2022. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research* 23, 120 (2022), 1–39.

[13] Huiyuan Fu, Wenkai Zheng, Xicong Wang, Jiaxuan Wang, Heng Zhang, and Huadong Ma. 2023. Dancing in the dark: A benchmark towards general low-light video enhancement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 12877–12886.

[14] Georg Gerstweiler, Emanuel Vonach, and Hannes Kaufmann. 2015. HyMoTrack: A mobile AR navigation system for complex indoor environments. *Sensors* 16, 1 (2015), 17.

[15] Christina Granquist, Susan Y Sun, Sandra R Montezuma, Tu M Tran, Rachel Gage, and Gordon E Legge. 2021. Evaluation and comparison of artificial intelligence vision aids: Orcam myeye 1 and seeing ai. *Journal of Visual Impairment & Blindness* 115, 4 (2021), 277–285.

[16] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. 2022. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 18995–19012.

[17] Chunle Guo, Chongyi Li, Jichang Guo, Chen Change Loy, Junhui Hou, Sam Kwong, and Runmin Cong. 2020. Zero-reference deep curve estimation for low-light image enhancement. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 1780–1789.

[18] Jiang Hai, Zhu Xuan, Ren Yang, Yutong Hao, Fengzhu Zou, Fang Lin, and Songchen Han. 2023. R2rnet: Low-light image enhancement via real-low to real-normal network. *Journal of Visual Communication and Image Representation* 90 (2023), 103712.

[19] Tae Young Han and Byung Cheol Song. 2016. Night vision pedestrian detection based on adaptive preprocessing using near infrared camera. In *2016 IEEE International Conference on Consumer Electronics-Asia (ICCE-Asia)*. IEEE, 1–3.

[20] Samuel W Hasinoff, Dillon Sharlet, Ryan Geiss, Andrew Adams, Jonathan T Barron, Florian Kainz, Jiawen Chen, and Marc Levoy. 2016. Burst photography for high dynamic range and low-light imaging on mobile cameras. *ACM Transactions on Graphics (ToG)* 35, 6 (2016), 1–12.

[21] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. 2022. Video diffusion models. *Advances in neural information processing systems* 35 (2022), 8633–8646.

[22] Zhe Hu, Sunghyun Cho, Jue Wang, and Ming-Hsuan Yang. 2014. Deblurring low-light images with light streaks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3382–3389.

[23] Yi Huang, Jiancheng Huang, Jianzhuang Liu, Mingfu Yan, Yu Dong, Jiaxi Lv, Chaoqi Chen, and Shifeng Chen. 2024. Wavedm: Wavelet-based diffusion models for image restoration. *IEEE Transactions on Multimedia* 26 (2024), 7058–7073.

[24] Balu N Ilag and Arun M Sabale. 2022. Microsoft teams overview. In *Troubleshooting Microsoft Teams: Enlisting the Right Approach and Tools in Teams for Mapping and Troubleshooting Issues*. Springer, 17–74.

[25] Varun Jain, Zongwei Wu, Quan Zou, Louis Florentin, Henrik Turbell, Sandeep Siddhartha, Radu Timofte, Qifan Gao, Linyan Jiang, Qing Luo, et al. 2025. NTIRE 2025 challenge on video quality enhancement for video conferencing: Datasets, methods and results. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 1184–1194.

[26] Bin Jiang, Jiachen Yang, Zhihan Lv, and Houbing Song. 2018. Wearable vision assistance system based on binocular sensors for visually impaired users. *IEEE Internet of Things Journal* 6, 2 (2018), 1375–1383.

[27] Hai Jiang, Ao Luo, Haoqiang Fan, Songchen Han, and Shuaicheng Liu. 2023. Low-light image enhancement with wavelet-based diffusion models. *ACM Transactions on Graphics (TOG)* 42, 6 (2023), 1–14.

[28] Hai Jiang, Ao Luo, Xiaohong Liu, Songchen Han, and Shuaicheng Liu. 2024. Lightendiffusion: Unsupervised low-light image enhancement with latent-retinex diffusion models. In *European Conference on Computer Vision*. Springer, 161–179.

[29] Haiyang Jiang and Yinqiang Zheng. 2019. Learning to see moving objects in the dark. In *Proceedings of the IEEE/CVF international conference on computer vision*. 7324–7333.

[30] Shuangping Jin, Bingbing Yu, Minhao Jing, Yi Zhou, Jiajun Liang, and Renhe Ji. 2022. Darkvisionnet: Low-light imaging via rgb-nir fusion with deep inconsistency prior. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. 1104–1112.

[31] Demetrios Karis, Daniel Wildman, and Amir Mané. 2016. Improving remote collaboration with video conferencing and video portals. *Human–Computer Interaction* 31, 1 (2016), 1–58.

[32] Edwin H Land. 1977. The retinex theory of color vision. *Scientific american* 237, 6 (1977), 108–129.

[33] Micheal Lanham. 2018. *Learn ARCore-Fundamentals of Google ARCore: Learn to build augmented reality apps for Android, Unity, and the web with Google ARCore 1.0.* Packt Publishing Ltd.

[34] Chongyi Li, Chunle Guo, Linghao Han, Jun Jiang, Ming-Ming Cheng, Jinwei Gu, and Chen Change Loy. 2021. Low-light image and video enhancement using deep learning: A survey. *IEEE transactions on pattern analysis and machine intelligence* 44, 12 (2021), 9396–9416.

[35] Chongyi Li, Chunle Guo, Linghao Han, Jun Jiang, Ming-Ming Cheng, Jinwei Gu, and Chen Change Loy. 2021. Low-light image and video enhancement using deep learning: A survey. *IEEE transactions on pattern analysis and machine intelligence* 44, 12 (2021), 9396–9416.

[36] Chongyi Li, Chunle Guo, and Chen Change Loy. 2021. Learning to enhance low-light image via zero-reference deep curve estimation. *IEEE transactions on pattern analysis and machine intelligence* 44, 8 (2021), 4225–4238.

[37] Guoxin Li, Jiaqi Xu, Zhijun Li, Chao Chen, and Zhen Kan. 2022. Sensing and navigation of wearable assistance cognitive systems for the visually impaired. *IEEE Transactions on Cognitive and Developmental Systems* 15, 1 (2022), 122–133.

[38] Stan Z Li, RuFeng Chu, ShengCai Liao, and Lun Zhang. 2007. Illumination invariant face recognition using near-infrared images. *IEEE Transactions on pattern analysis and machine intelligence* 29, 4 (2007), 627–639.

[39] Wenhao Li, Guangyang Wu, Wenyi Wang, Peiran Ren, and Xiaohong Liu. 2023. Fastllve: Real-time low-light video enhancement with intensity-aware look-up table. In *Proceedings of the 31st ACM International Conference on Multimedia*. 8134–8144.

[40] Sicong Liu, Xiaochen Li, Zimu Zhou, Bin Guo, Meng Zhang, Haocheng Shen, and Zhiwen Yu. 2023. AdaEnlight: Energy-aware low-light video stream enhancement on mobile devices. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 4 (2023), 1–26.

[41] X Liu, Boyd Fowler, Hung Do, Steve Mims, Dan Laxson, and Brett Frymire. 2007. High performance CMOS image sensor for low light imaging. In *International Image Sensor Workshop*. 327–330.

[42] Kin Gwn Lore, Adedotun Akintayo, and Soumik Sarkar. 2017. LLNet: A deep autoencoder approach to natural low-light image enhancement. *Pattern Recognition* 61 (2017), 650–662.

[43] Gang Luo. 2020. How 16,000 people used a smartphone magnifier app in their daily lives. *Clinical and Experimental Optometry* 103, 6 (2020), 847–852.

[44] Yun Luo, Jeffrey Remillard, and Dieter Hoetzer. 2010. Pedestrian detection in near-infrared night vision system. In *2010 IEEE Intelligent Vehicles Symposium*. IEEE, 51–58.

[45] Feifan Lv, Feng Lu, Jianhua Wu, and Chongsoon Lim. 2018. MBLLEN: Low-light image/video enhancement using cnns.. In *Bmvc*, Vol. 220. Northumbria University, 4.

[46] Xiaoqian Lv, Shengping Zhang, Chenyang Wang, Yichen Zheng, Bineng Zhong, Chongyi Li, and Liqiang Nie. 2024. Fourier priors-guided diffusion for zero-shot joint low-light enhancement and deblurring. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 25378–25388.

[47] Xin Ma, Yaohui Wang, Xinyuan Chen, Gengyun Jia, Ziwei Liu, Yuan-Fang Li, Cunjian Chen, and Yu Qiao. 2024. Latte: Latent diffusion transformer for video generation. *arXiv preprint arXiv:2401.03048* (2024).

[48] William Peebles and Saining Xie. 2023. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*. 4195–4205.

[49] Bo Peng, Xuanyu Zhang, Jianjun Lei, Zhe Zhang, Nam Ling, and Qingming Huang. 2022. LVE-S2D: Low-light video enhancement from static to dynamic. *IEEE Transactions on Circuits and Systems for Video Technology* 32, 12 (2022), 8342–8352.

[50] Aashish Sharma and Robby T Tan. 2021. Nighttime visibility enhancement by increasing the dynamic range and suppression of light effects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 11977–11986.

[51] Daisuke Sugimura, Takuya Mikami, Hiroki Yamashita, and Takayuki Hamamoto. 2015. Enhancing color images of extremely low light scenes based on RGB/NIR images acquisition with different exposure times. *IEEE Transactions on Image Processing* 24, 11 (2015), 3586–3597.
[52] Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz. 2018. Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 8934–8943.
[53] Li Tao, Chuang Zhu, Jiawen Song, Tao Lu, Huizhu Jia, and Xiaodong Xie. 2017. Low-light image enhancement using CNN and bright channel prior. In *2017 IEEE international conference on image processing (ICIP)*. IEEE, 3215–3219.
[54] Li Tao, Chuang Zhu, Guoqing Xiang, Yuan Li, Huizhu Jia, and Xiaodong Xie. 2017. LLCNN: A convolutional neural network for low-light image enhancement. In *2017 IEEE Visual Communications and Image Processing (VCIP)*. IEEE, 1–4.
[55] Mpilo M Tatana, Mohohlo S Tsoeu, and Rito C Maswanganyi. 2025. Low-Light Image and Video Enhancement for More Robust Computer Vision Tasks: A Review. *Journal of Imaging* 11, 4 (2025), 125.
[56] Thorsten Tatarek, Jan Kronenberger, and Uwe Handmann. 2017. Functionality, advantages and limits of the tesla autopilot. (2017).
[57] Surat Teerapittayanon, Bradley McDanel, and Hsiang-Tsung Kung. 2016. Branchynet: Fast inference via early exiting from deep neural networks. In *2016 23rd international conference on pattern recognition (ICPR)*. IEEE, 2464–2469.
[58] Chenxi Wang, Hongjun Wu, and Zhi Jin. 2023. Fourllie: Boosting low-light image enhancement by fourier frequency information. In *Proceedings of the 31st ACM International Conference on Multimedia*. 7459–7469.
[59] Meiao Wang, Xuejing Kang, Yaxi Lu, and Jie Xu. 2025. RetinexMCNet: A Memory Controller Dominated Network for Low-Light Video Enhancement Based on Retinex. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 9716–9725.
[60] Ruixing Wang, Xiaogang Xu, Chi-Wing Fu, Jiangbo Lu, Bei Yu, and Jiaya Jia. 2021. Seeing dynamic scene in the dark: A high-quality video dataset with mechatronic alignment. In *Proceedings of the IEEE/CVF international conference on computer vision*. 9700–9709.
[61] Yufei Wang, Renjie Wan, Wenhan Yang, Haoliang Li, Lap-Pui Chau, and Alex Kot. 2022. Low-light image enhancement with normalizing flow. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 36. 2604–2612.
[62] Yufei Wang, Yi Yu, Wenhan Yang, Lanqing Guo, Lap-Pui Chau, Alex C Kot, and Bihan Wen. 2023. Exposurediffusion: Learning to expose for low-light image enhancement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 12438–12448.
[63] Yufei Wang, Yi Yu, Wenhan Yang, Lanqing Guo, Lap-Pui Chau, Alex C Kot, and Bihan Wen. 2023. Exposurediffusion: Learning to expose for low-light image enhancement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 12438–12448.
[64] Chen Wei, Wenjing Wang, Wenhan Yang, and Jiaying Liu. 2018. Deep retinex decomposition for low-light enhancement. *arXiv preprint arXiv:1808.04560* (2018).
[65] Ji Wei, Qian Zhijie, Xu Bo, and Zhao Dean. 2018. A nighttime image enhancement method based on Retinex and guided filter for object recognition of apple harvesting robot. *International Journal of Advanced Robotic Systems* 15, 1 (2018), 1729881417753871.
[66] Prayanto Widyo Harsanto and Jalung Wirangga Jakti. 2023. Post-photography: the disruption effect of artificial intelligence on photography for product advertising. *Information Sciences Letters an International Journal* 9, 12 (2023), 2141–2151.
[67] Wenhui Wu, Jian Weng, Pingping Zhang, Xu Wang, Wenhan Yang, and Jianmin Jiang. 2022. Uretinex-net: Retinex-based deep unfolding network for low-light image enhancement. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 5901–5910.
[68] Shuochao Yao, Shaohan Hu, Yiran Zhao, Aston Zhang, and Tarek Abdelzaher. 2017. Deepsense: A unified deep learning framework for time-series mobile sensing data processing. In *Proceedings of the 26th international conference on world wide web*. 351–360.
[69] Zilyu Ye, Zhiyang Chen, Tiancheng Li, Zemin Huang, Weijian Luo, and Guo-Jun Qi. 2025. Schedule on the fly: Diffusion time prediction for faster and better image generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 23412–23422.
[70] Rongjie Yi, Liwei Guo, Shiyun Wei, Ao Zhou, Shangguang Wang, and Mengwei Xu. 2025. Edgemoe: Empowering sparse large language models on mobile devices. *IEEE Transactions on Mobile Computing* (2025).
[71] Xunpeng Yi, Han Xu, Hao Zhang, Linfeng Tang, and Jiayi Ma. 2023. Diff-retinex: Rethinking low-light image enhancement with a generative diffusion model. In *Proceedings of the IEEE/CVF international conference on computer vision*. 12302–12311.
[72] Fan Zhang, Yu Li, Shaodi You, and Ying Fu. 2021. Learning temporal consistency for low light video enhancement from single images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 4967–4976.
[73] Fan Zhang, Yu Li, Shaodi You, and Ying Fu. 2021. Learning temporal consistency for low light video enhancement from single images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 4967–4976.
[74] Gengchen Zhang, Yulun Zhang, Xin Yuan, and Ying Fu. 2024. Binarized low-light raw video enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 25753–25762.
[75] Lixian Zhang, Runmin Dong, Shuai Yuan, Jinxiao Zhang, Mengxuan Chen, Juepeng Zheng, and Haohuan Fu. 2024. Deeplight: Reconstructing high-resolution observations of nighttime light with multi-modal remote sensing data. *arXiv preprint arXiv:2402.15659* (2024).
[76] Lin Zhang, Lijun Zhang, Xiao Liu, Ying Shen, Shaoming Zhang, and Shengjie Zhao. 2019. Zero-shot restoration of back-lit images using deep internal learning. In *Proceedings of the 27th ACM international conference on multimedia*. 1623–1631.

[77] Xin Zhang, Xia Wang, and Changda Yan. 2023. LL-CSFormer: A novel image denoiser for intensified CMOS sensing images under a low light environment. *Remote Sensing* 15, 10 (2023), 2483.

[78] Yonghua Zhang, Xiaojie Guo, Jiayi Ma, Wei Liu, and Jiawan Zhang. 2021. Beyond brightening low-light images. *International Journal of Computer Vision* 129, 4 (2021), 1013–1037.

[79] Zangwei Zheng, Xiangyu Peng, Tianji Yang, Chenhui Shen, Shenggui Li, Hongxin Liu, Yukun Zhou, Tianyi Li, and Yang You. 2024. Open-sora: Democratizing efficient video production for all. *arXiv preprint arXiv:2412.20404* (2024).

[80] Anqi Zhu, Lin Zhang, Ying Shen, Yong Ma, Shengjie Zhao, and Yicong Zhou. 2020. Zero-shot restoration of underexposed images via robust retinex decomposition. In *2020 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE Computer Society, 1–6.

[81] Lingyu Zhu, Wenhan Yang, Baoliang Chen, Hanwei Zhu, Zhangkai Ni, Qi Mao, and Shiqi Wang. 2024. Unrolled decomposed unpaired learning for controllable low-light video enhancement. In *European Conference on Computer Vision*. Springer, 329–347.