REASONING IN THE DARK: EVENT-RGB ADAPTATION FOR LOW-LIGHT VIDEO MLLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Video multimodal large language models (video MLLMs) rely heavily on RGB appearance cues and degrade sharply in low light, where texture, color, and object boundaries are lost. Event cameras provide complementary motion and temporal information under poor illumination, but integrating event streams into language-based reasoning is bottlenecked by scarce paired event-language supervision, especially under low light. We propose LITE-Event, a sample-efficient framework for low-light video reasoning that fuses event and RGB streams. A lightweight visual-language interface combines modality-specific projectors with gated fusion. A support-conditioned hypernetwork then generates LoRA-style residual updates from a small non-target support set, adapting the interface without updating the frozen encoders or LLM. We also introduce three resources for training and evaluation: NExT-QA-Dark, a large-scale synthetic low-light RGB-event dataset for pretraining, EventLang-Dark, a curated event-language dataset for adaptation, and RealDark-Bench, a real-world low-light RGB-event reasoning benchmark. Trained predominantly on synthetic data, LITE-Event reaches 85.0% on RealDark-Bench, outperforming the strongest enhancement-then-reason baseline by 5.2 points and surpassing both RGB-only and event-only video MLLMs.

1 INTRODUCTION

Recent video multimodal large language models (video MLLMs) achieve strong performance on visual reasoning by extending image-language models to temporally structured inputs (Zhang et al., 2024b; Zhu et al., 2025; Bai et al., 2025; Zhang et al., 2025). These models depend on appearance cues from RGB frames, and this dependence breaks down in low light, where texture, color, and object boundaries are lost. Recent benchmarks have made this gap concrete, reporting substantial accuracy drops under challenging illumination (Park et al., 2025; Zhang et al., 2026b). Low-light enhancement methods can partially restore visibility, but they operate on already degraded inputs and cannot recover temporal evidence such as motion discontinuities, subtle state changes, or interaction dynamics once the signal-to-noise ratio is severely degraded.

Event cameras offer a complementary signal by capturing asynchronous brightness changes rather than intensity, preserving motion, edges, and temporal transitions under extreme illumination. A growing body of work has begun connecting event streams to language models (Liu et al., 2025b; Li et al., 2025; Zhou & Lee, 2025; Liu et al., 2026; Xu et al., 2025; Zhang et al., 2026a), advancing event-language alignment. Closest to our low-light setting, Event-MLLM (Zhang et al., 2026a) fuses a degraded RGB frame with a corresponding event frame and learns illumination correction using paired normal-light references. LITE-Event instead targets temporally synchronized event-RGB video under scarce real event-language supervision, where paired clean references are not required for real-event adaptation and no target-domain support is available. More generally, standard recipes for adapting MLLMs to new modalities through instruction tuning or model fine-tuning (Liu et al., 2023; Li et al., 2023) are difficult to apply in this low-resource regime and can be prone to overfitting.

We propose LITE-Event (Low-light Integrated Temporal Event reasoning), a sample-efficient event-RGB adaptation framework for low-light video reasoning. LITE-Event keeps the visual encoders and LLM frozen and connects them through a lightweight visual-language interface with modality-specific projectors and gated fusion. A hypernetwork generates LoRA-style residual updates from a small non-target support set. This source-conditioned adaptation helps transfer an interface learned

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

predominantly from synthetic data to real low-light event-RGB video. The model requires only limited real event-language supervision and no target-domain support.

In our work, we use a staged pipeline that moves from synthetic pretraining to real-event adaptation and held-out real-world evaluation. We construct NExT-QA-Dark by augmenting a large video-language dataset with synthetic low-light RGB and event streams for projector and hypernetwork pretraining. We then adapt the model on EventLang-Dark, a dataset we curate by combining low-light RGB with real event data. For evaluation, we introduce RealDark-Bench, a real-world low-light RGB-event reasoning benchmark built from multiple datasets originally proposed for event-guided low-light video enhancement. Trained predominantly on synthetic data, LITE-Event reaches 85.0% on RealDark-Bench, outperforming the strongest enhancement-then-reason baseline by 5.2 points and surpassing both RGB-only and event-only video MLLMs. These results support the benefit of combining event-derived temporal cues with RGB appearance through a transferable multimodal interface.

In summary, our contributions are threefold: (1) we introduce LITE-Event, a sample-efficient event-RGB adaptation framework that keeps the visual encoders and LLM frozen while adapting the visual-language interface through support-conditioned low-rank residual updates; (2) we develop a staged low-light event-language learning pipeline, with NExT-QA-Dark for large-scale synthetic pretraining and EventLang-Dark for adaptation with real event data; and (3) we introduce RealDark-Bench, a real-world low-light RGB-event reasoning benchmark for evaluating generalization from predominantly synthetic training to unseen real low-light scenarios.

2 Related Work

Video MLLMs and Low-Light Robustness. Video MLLMs extend image-language models to multi-frame video understanding (Zhang et al., 2024b; Zhu et al., 2025; Bai et al., 2025; Zhang et al., 2025). Their reasoning relies heavily on RGB appearance cues, which degrade under poor illumination. Benchmarks like DarkEQA (Park et al., 2025) and EgoNight (Zhang et al., 2026b) report large performance drops, while adopting enhancement-then-reason as the default solution. We instead address the gap at the sensing level by adding event streams as a complementary input.

Event-Based MLLMs. Event-language modeling has expanded from staged event-language alignment to multimodal fusion and efficient event processing. EventGPT (Liu et al., 2025b) aligns event representations with language, while EventVL (Li et al., 2025) scales the paradigm with large event-image/video-text corpora. LLaFEA (Zhou & Lee, 2025) combines frame and event signals for fine-grained spatiotemporal reasoning, EventFlash (Liu et al., 2026) uses event sparsity for long-sequence processing, and EventSTU (Xu et al., 2025) uses events as a training-free signal for token reduction. Most learned event-language models rely on substantial paired supervision. Closest to our work, Event-MLLM (Zhang et al., 2026a) performs illumination-aware fusion of a degraded RGB frame and a corresponding event frame. Its training uses paired normal-light references to align the fused representation with non-degraded semantics. LITE-Event addresses a different problem. It reasons over synchronized event-RGB video sequences, does not require paired normal-light references during real-event adaptation. Thus, Event-MLLM focuses on illumination-robust frame-level fusion, whereas LITE-Event focuses on sample-efficient temporal reasoning and synthetic-to-real transfer.

Event-Language Datasets and Benchmarks. Existing event-language resources span instruction-oriented corpora such as N-ImageNet-Chat and EventChat (Liu et al., 2025b) and the larger event-image/video-text data of EventVL (Li et al., 2025); task-specific benchmarks for QA and grounding such as EvQA (Lou et al., 2025b) and Talk2Event (Kong et al., 2025); and broader multi-task evaluations such as EventBench (Liu et al., 2025a). Coverage of low-light multimodal reasoning, however, is limited. Most resources assume event-only inputs or treat RGB and event streams independently, and emphasize perception-level rather than temporally extended reasoning. Our RealDark-Bench (§3.3) addresses this gap with paired real-world low-light RGB-event video and questions that require temporal evidence.

Support-Conditioned Modality Integration. A common approach to integrating new modalities into LLMs is to keep the encoder and language model frozen and learn a lightweight connector like an MLP projector (Gao et al., 2023; Liu et al., 2023), Q-Former (Li et al., 2023), Perceiver resampler (Jaegle et al., 2021), or cross-attention interface (Alayrac et al., 2022). Recent work makes these interfaces conditional. HyMA (Singh et al., 2025) generates connectors for pretrained model

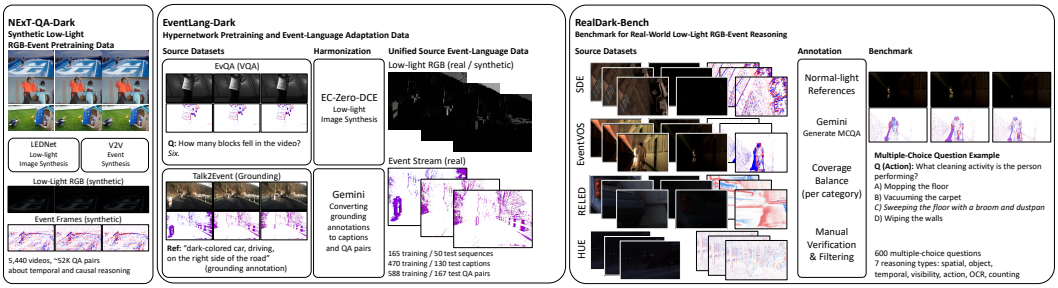


Figure 1: **Synthetic-to-real data pipeline for low-light RGB–event reasoning.** (A) NExT-QA-Dark: large-scale pretraining data with synthetic low-light RGB and event streams. (B) EventLang-Dark: adaptation data with harmonized low-light inputs and real event signals. (C) RealDark-Bench: real-world low-light RGB–event benchmark with semi-automatic annotation for evaluation.

pairs, HyperTuning (Phang, 2024) compresses few-shot support examples into adaptation parameters, and HyperLLaVA (Zhang et al., 2024a) produces input-conditioned parameter shifts. SEMI (İnce et al., 2025) uses a support-conditioned hypernetwork to generate low-rank projector updates for sample-efficient modality integration. LITE-Event builds on these aforementioned ideas but targets *synthetic-to-real transfer for synchronized event-RGB video without target-domain support*. It jointly models aligned event and RGB streams with modality-specific projectors, gated fusion, temporal support prototypes, and a dual-stream support encoder. Residual updates are generated from non-target event-RGB support and applied to unseen real low-light data. Our focus is source-conditioned adaptation of paired temporal modalities for synthetic-to-real low-light reasoning.

3 Datasets and Benchmark

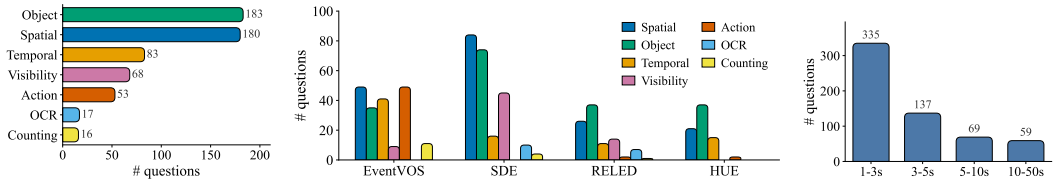
We introduce three novel resources for low-light event-RGB video reasoning (Fig. 1): NExT-QA-Dark for large-scale synthetic pretraining (§3.1), EventLang-Dark for adaptation with real event data (§3.2), and RealDark-Bench for evaluation on real-world low-light RGB-event data (§3.3).

3.1 NExT-QA-Dark: Synthetic Low-Light RGB-Event Pretraining

NExT-QA-Dark augments NExT-QA (Xiao et al., 2021), a video-language dataset with 5,440 videos and ∼52K QA pairs focused on temporal and causal reasoning, with synthetic low-light RGB and event streams. We synthesize low-light RGB using the exposure-conditioned Zero-DCE procedure of Zhou et al. (2022), which applies a reversed, spatially varying curve adjustment conditioned on exposure, and generate aligned events from the original RGB sequences using V2V (Lou et al., 2025a). The resulting data pairs synthetic low-light RGB with simulated events but does not fully capture real low-light event statistics, motivating Stage 3 adaptation on EventLang-Dark. Original splits and QA annotations are preserved. NExT-QA-Dark retains the temporal and causal emphasis, encouraging the interface to exploit motion and state changes when RGB evidence is degraded.

3.2 EventLang-Dark: Event-Language Adaptation Data

EventLang-Dark provides paired event-language supervision for adapting LITE-Event from synthetic to real event data. It is curated from EvQA (Lou et al., 2025b), which provides event-based QA, and Talk2Event (Kong et al., 2025), which provides language-grounded event annotations. We construct 165 training sequences with 470 captions and 588 multiple-choice QA pairs. We apply two harmonization steps: (i) for normal-light sequences, we synthesize low-light RGB counterparts again using the approach suggested in (Zhou et al., 2022) to match the event/RGB-dark format used during pretraining; and (ii) because Talk2Event is not originally a QA dataset, we use Gemini to convert its grounding annotations into captions and multiple-choice QA pairs with category-consistent distractors, followed by manual filtering of ambiguous or unsupported items. EvQA annotations are kept unchanged. The resulting dataset provides compact, real-event-grounded supervision that complements synthetic pretraining without requiring large-scale event-language collection.



(a) Question types. (b) Question types by dataset. (c) Clip durations.

Figure 2: **Statistics of RealDark-Bench.** The benchmark contains 600 multiple-choice questions across seven reasoning categories (a), drawn in comparable proportions from four sources (b). Clip durations mostly fall within 1-5 seconds (c), reflecting the short-form nature of low-light event-RGB video and emphasizing fine-grained temporal reasoning over short interactions.

3.3 REALDARK-BENCH: REAL-WORLD LOW-LIGHT RGB-EVENT REASONING

We introduce RealDark-Bench, a multiple-choice benchmark for reasoning over paired real-world low-light RGB and event video. It contains 600 questions spanning seven reasoning types: spatial, object, temporal, visibility, action, OCR, and counting (Fig. 2). The clips are drawn from EventVOS (Li et al., 2024), SDE (Liang et al., 2024), RELED (Kim et al., 2024), and HUE (Ercan et al., 2024). EventVOS, SDE, and RELED were originally proposed for event-guided low-light video enhancement and provide aligned low-light RGB and event streams. HUE includes high-resolution real-world sequences spanning diverse indoor, urban, twilight, night-time, driving, and controlled conditions. Existing event-language resources either use event-only inputs, lack paired low-light RGB-event video, or do not provide language annotations suitable for multimodal reasoning. RealDark-Bench addresses this gap by evaluating questions that require complementary appearance, spatial, and temporal cues that degraded RGB alone may not reliably provide. Notably, EventLang-Dark and RealDark-Bench are also strictly disjoint, sharing no source dataset, recording sequence, or video clip. No RealDark-Bench sample is used for training, support construction, model selection, or tuning.

Annotation pipeline. Annotating severely degraded low-light video is difficult since object identities, actions, and fine-grained scene details can be hard to infer reliably. For EventVOS, SDE, and RELED, we use provided normal-light references only to assist candidate annotation. These references are never available during evaluation. HUE provides no paired normal-light references, so DarkIR-enhanced frames serve as an auxiliary annotation aid. Every retained question must remain answerable from the actual evaluation inputs: low-light RGB frames, event frames, or both. The annotation pipeline has four stages. (i) *Generation:* Gemini produces dense per-clip captions and candidate multiple-choice questions across the seven reasoning categories, with category-consistent distractors. (ii) *Coverage:* We balance candidate generation across categories while preserving each source dataset's natural scene distribution. (iii) *Independent verification:* Three annotators review each question-video pair independently. They assess visual grounding, answer uniqueness, distractor plausibility, ambiguity, reliance on external knowledge, and whether the answer is supported by the evaluation inputs rather than annotation-only references. We retain an item only if at least two annotators judge it valid. (iv) *Correction and filtering:* We remove ambiguous or unsupported items, revise correctable answer-choice errors, and recheck them against the original evaluation inputs.

Evaluation protocol. Although normal-light references are used for annotation, models receive *only* degraded low-light RGB frames and aligned event streams at test time. This ensures that RealDark-Bench evaluates low-light multimodal reasoning generalization from predominantly synthetic training with limited real-event adaptation to unseen real scenes. The benchmark contains 600 multiple-choice questions: 194 from EventVOS, 233 from SDE, 98 from RELED, and 75 from HUE, emphasizing temporally grounded reasoning over real-world low-light video.

4 METHOD: LITE-EVENT

We propose LITE-Event, an event-RGB adaptation model for low-light video reasoning (Fig. 3). Given a low-light RGB video V^r, an event stream V^e, and a query q, the model predicts an answer y via frozen event/RGB encoders and an instruction-tuned LLM (Grattafiori et al., 2024), connected by a lightweight visual-language interface. A support-conditioned hypernetwork reads source-domain

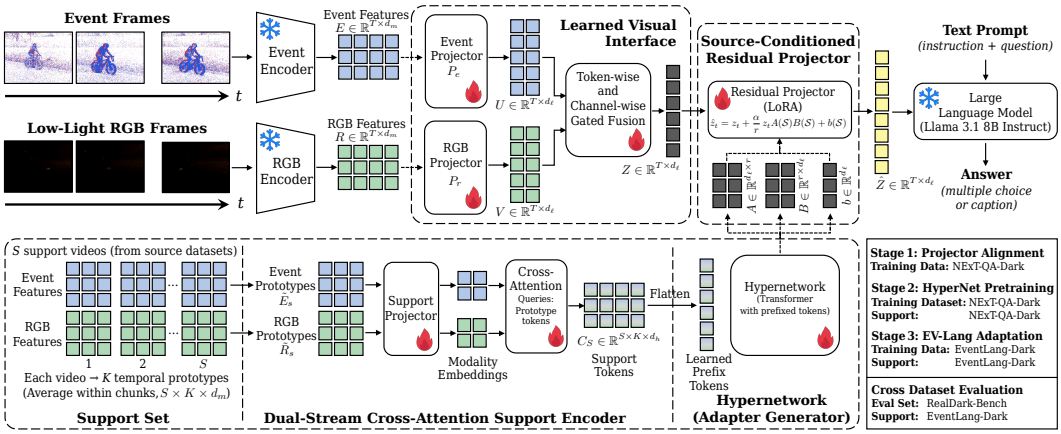


Figure 3: **Overview of LITE-Event.** Frozen encoders extract event and RGB features, which are projected, fused via a learned gate, and adapted by a support-conditioned hypernetwork that generates LoRA-style residual updates. The adapted visual tokens are decoded by the frozen LLM.

event-RGB support examples and generates low-rank residual updates for this interface, enabling source-conditioned transfer to unseen real low-light data without target-domain support. Given event and RGB features, LITE-Event operates in four steps: (i) project both modalities into the LLM embedding space using modality-specific projectors, (ii) fuse them into visual tokens via a learned gate, (iii) adapt the fused tokens using the support-conditioned residual projector, and (iv) feed the adapted tokens with the text prompt to the frozen LLM for decoding (Grattafiori et al., 2024).

4.1 Input Representation

Each video is divided into $T = 16$ equal-duration temporal segments, with synchronized event and RGB content per segment. We extract per-segment features using frozen CM3AE encoders (Wu et al., 2025), ϕ_e and ϕ_r, which share an embedding space:

$$e_t = \frac{1}{|\mathcal{E}_t|} \sum_{x \in \mathcal{E}_t} \phi_e(x), \qquad r_t = \frac{1}{|\mathcal{I}_t|} \sum_{I \in \mathcal{I}_t} \phi_r(I), \tag{1}$$

producing aligned sequences $e_t, r_t \in \mathbb{R}^{d_m}$. $\mathcal{E}_t$ is either the event polarity frame from the t-th temporal bin (raw event input) or the corresponding event packet chunk (aligned frame/event input). $\mathcal{I}_t$ contains the aligned RGB frames. This yields a unified encoder-space interface across data sources: event features capture brightness changes and motion; RGB features retain appearance and scene context.

4.2 Event-RGB Visual Interface

Modality-specific MLP projectors map features into the LLM embedding space, $u_t = P_e(e_t)$ and $v_t = P_r(r_t)$ with $u_t, v_t \in \mathbb{R}^{d_\ell}$. A learned per-token gate fuses the two streams:

$$z_t = g_t \odot u_t + (1 - g_t) \odot v_t, \qquad g_t = \sigma(W_g[u_t; v_t] + b_g)\,. \tag{2}$$

Because $g_t \in \mathbb{R}^{d_\ell}$ is predicted per token, the gate is both token-wise and channel-wise: different tokens and different channels can rely on event or RGB evidence to different degrees. The fused sequence $Z = (z_1, \ldots, z_T)$ forms the query-side visual tokens before adaptation.

4.3 Support-Conditioned Adaptation

A fixed projector is sensitive to variations in illumination and event statistics. We instead condition the projector on a small support set $\mathcal{S}$ at inference time and generate residual updates per batch.

Temporal prototypes. For each of S support items, we compress the event and RGB feature sequences into K temporal prototypes by averaging contiguous chunks, yielding $E_s, R_s \in \mathbb{R}^{S \times K \times d_m}$. This preserves a coarse temporal signature per item while keeping the support context small.

Dual-stream support encoding. Prototypes are projected into the hypernetwork space and tagged with learned modality embeddings, giving $\tilde{E}_s, \tilde{R}_s$. For each support item, keys/values concatenate both modalities, $K_s = V_s = [\tilde{E}_s; \tilde{R}_s]$. Query sequence is initialized from the per-prototype average of the two modalities, augmented with a sample-level summary token (the mean of all prototypes for that item) and a learned sample-identity token. A cross-attention module then computes

$$C_s = \mathrm{CrossAttn}(Q_s, K_s, V_s), \qquad C_s \in \mathbb{R}^{S\times K\times d_h}, \tag{3}$$

keeping one output token per temporal prototype. The result is flattened into SK context tokens.

Hypernetwork and residual update. Context tokens are prepended with learned prefix tokens, processed by a transformer with positional encodings, and decoded into LoRA-style parameters

$$A(\mathcal{S}) \in \mathbb{R}^{d_\ell \times r}, \quad B(\mathcal{S}) \in \mathbb{R}^{r\times d_\ell}, \quad b(\mathcal{S}) \in \mathbb{R}^{d_\ell}. \tag{4}$$

Treating z_t as a row vector, the residual update is

$$\hat{z}_t = z_t + \frac{\alpha}{r}\, z_t A(\mathcal{S}) B(\mathcal{S}) + b(\mathcal{S}), \tag{5}$$

with $b(\mathcal{S})$ broadcast across tokens. We use $d_h = 768$, $r = 32$, and $\alpha = 32$. The adapted sequence $\hat{Z}$ is prepended to the prompt embeddings and decoded by the frozen LLM.

4.4 Training Protocol

LITE-Event is trained in three stages, using the data resources introduced in §3. The encoders and LLM are frozen throughout; only the visual-language interface and the hypernetwork that conditions it are updated. Across all stages, we optimize the causal language-modeling loss over answer tokens,

$$\mathcal{L} = -\sum_{i\in\mathcal{Y}} \log p_\Theta(y_i \mid y_{<i}, X), \tag{6}$$

where Θ denotes the frozen LLM parameters, $\mathcal{Y}$ is the set of answer positions, and $X = [Z; q]$ in Stage 1 (no adaptation) or $X = [\hat{Z}; q]$ in Stages 2 and 3.

Stage 1: Projector alignment. We train the modality-specific projectors P_e, P_r and the fusion gate on NExT-QA-Dark (§3.1), learning a stable mapping from event-RGB features into the LLM embedding space. The hypernetwork is not used at this stage.

Stage 2: Hypernetwork pretraining. We freeze the projectors and gate, and train the support encoder and hypernetwork on NExT-QA-Dark. Each step samples separate query and support sets, enabling the hypernetwork to produce residual updates conditioned on varying synthetic contexts.

Stage 3: Event-language adaptation. We adapt the support encoder and hypernetwork on EventLang-Dark (§3.2). The projectors and gate remain frozen, preserving the interface learned in Stage 1 while specializing adaptation to real event data.

4.5 Cross-Dataset Evaluation Protocol

At evaluation time, the model receives only RealDark-Bench inputs $(E_{\mathrm{real}}, R_{\mathrm{real}})$ and prompt q; no target-domain data is used for training or support. We sample M support batches from EventLang-Dark training data, $\mathcal{S}^{(m)}_{\mathrm{ELD}} \subset \mathcal{D}^{\mathrm{train}}_{\mathrm{ELD}}$, and average the generated factors and biases independently:

$$\bar{A}_\ell = \frac{1}{M}\sum_m A^{(m)}_\ell, \quad \bar{B}_\ell = \frac{1}{M}\sum_m B^{(m)}_\ell, \quad \bar{b}_\ell = \frac{1}{M}\sum_m b^{(m)}_\ell. \tag{7}$$

The averaged factors form a single fixed residual projector $\bar{P}_{\mathrm{res}}$, which adapts Z_{real} before frozen-LLM decoding.

5 Experiments

We evaluate LITE-Event from three perspectives: (1) whether event-language reasoning can be learned with limited real-event supervision, (2) whether the learned interface transfers to held-out real-world low-light RGB-event data, and (3) whether the improvements persist when compared with RGB-only models trained with matched supervision. We evaluate source-domain reasoning on held-out EventLang-Dark and cross-dataset generalization on RealDark-Bench.

5.1 Implementation Details

Model configuration. All experiments use CM3AE features with $d_m = 768$ and Llama-3.1-8B-Instruct as the frozen LLM. Each video is represented by $T = 16$ temporal tokens. The support encoder uses $S = 32$ examples with $K = 4$ temporal prototypes per item. The hypernetwork dimension is 768, with LoRA rank $r = 32$ and scaling $\alpha = 32$.

Training schedule. Training proceeds in three stages. Stage 1 trains the event and RGB projectors and fusion module on synthetic data from NExT-QA-Dark for 5 epochs. Stage 2 freezes this interface and trains the hypernetwork for 5 epochs. Stage 3 adapts the model on EventLang-Dark for 50 epochs. We use AdamW with weight decay 5×10^{-6}. The learning rate is 5×10^{-5} in Stages 1-2 and 1×10^{-4} in Stage 3. Stage 1 uses batch size 2 (with accumulation 2), while Stages 2-3 use query batch size 1 and support batch size 32 with accumulation 32. All training and evaluation experiments were run on a single NVIDIA RTX A6000 GPU.

Evaluation setup. All models use deterministic decoding and identical generation limits. For RGB baselines, we sample the same number of frames per video. For enhancement baselines, frames are first enhanced and then passed to the same video MLLM. Unless otherwise stated, general video MLLM baselines are evaluated off-the-shelf. For data-controlled comparisons, we additionally finetune InternVL3 and Qwen3-VL with the same supervision used by LITE-Event. For RealDark-Bench evaluation, we form a fixed residual projector by averaging the generated LoRA factors and biases over $M = 31$ EventLang-Dark support batches.

5.2 Baselines

We compare LITE-Event with four baseline groups covering complementary aspects of the problem[1].

- **RGB-only video MLLMs.** We evaluate LLaVA-NeXT-Video (Zhang et al., 2024b), InternVL3 (Zhu et al., 2025), Qwen3-VL-8B-Instruct (Bai et al., 2025), and VideoLLaMA3 (Zhang et al., 2025) using only low-light RGB frames. These models test whether strong RGB-based reasoning alone is sufficient under severe illumination degradation.
- **Supervised RGB-only models.** To test whether additional post-training supervision alone explains the gains of LITE-Event, we finetune InternVL3 and Qwen3-VL via LoRA using the corresponding RGB-dark training data. DarkSFT denotes tuning on NExT-QA-Dark, while DarkSFT+EventLang further adapts the models on EventLang-Dark.
- **Enhancement-then-reason.** We enhance low-light RGB frames with DarkIR (Feijoo et al., 2025) or LEDNet (Zhou et al., 2022) before passing them to InternVL3. These baselines test whether improving RGB visibility can substitute for complementary event information.
- **Event-only MLLM.** We evaluate EventGPT-Plus (Liu et al., 2025a), which operates solely on event streams, to test whether event-only reasoning suffices without RGB appearance cues.

5.3 Event-Language Evaluation under Low-Light Inputs

We first evaluate LITE-Event on the held-out EventLang-Dark split to assess event-language reasoning under low-light RGB-event inputs before cross-dataset evaluation. Table 1 reports results for multiple-choice question-answering (MCQA) and captioning. LITE-Event performs best on both tasks. The supervised RGB-only controls show that matched post-training data alone does not account for the gain. DarkSFT lowers MCQA accuracy for both InternVL3 and Qwen3-VL. Adding EventLang-Dark supervision raises InternVL3 to 71.26%, still below LITE-Event at 80.24%. Qwen3-VL shows a similar pattern, with both supervised variants remaining below LITE-Event. EventGPT-Plus also underperforms LITE-Event, showing that events alone are not enough. The results show the benefit of combining event-derived temporal cues, RGB appearance, and support-conditioned adaptation.

5.4 Generalization to Real Low-Light RGB-Event Data

[1]EventFlash (Liu et al., 2026) lacks a complete public fine-tuning/checkpoint release. Event-MLLM (Zhang et al., 2026a) provides code, but its checkpoint was inaccessible at submission time, and faithful retraining is not directly compatible with our setting as it uses frame-level RGB-event pairs and paired normal-light references.

Table 1: **Event-language evaluation on the held-out EventLang-Dark split.** LITE-Event outperforms off-the-shelf, enhancement-based, supervised RGB-only, and event-only baselines on both MCQA and captioning tasks.

MLLM Type	Model	MCQA	Captioning				
		Acc. (%)	ROUGE-1	ROUGE-L	BLEU	METEOR	CIDEr
General Models	LLaVA-NeXT-Video-7B	52.10	0.3841	0.2757	0.0865	0.3370	0.1592
	VideoLLaMA3-7B	63.47	0.2996	0.2379	0.0429	0.1998	0.0550
	InternVL3-8B	68.26	0.2855	0.2436	0.0205	0.1690	0.0197
	Qwen3-VL-8B-Instruct	66.47	0.2887	0.2192	0.0491	0.2614	0.1440
Enhancement-then-reason	InternVL3-8B + DarkIR	67.66	0.2936	0.2489	0.0179	0.1721	0.0174
	InternVL3-8B + LEDNet	67.96	0.2896	0.2463	0.0192	0.1706	0.0186
Supervised RGB-only	InternVL3-DarkSFT	59.88	0.2954	0.2773	0.0533	0.1694	0.0134
	InternVL3-DarkSFT+EventLang	71.26	0.3944	0.3420	0.1195	0.2804	0.2170
	Qwen3-VL-DarkSFT	61.08	0.3003	0.2216	0.0591	0.2513	0.1795
	Qwen3-VL-DarkSFT+EventLang	62.28	0.3070	0.2241	0.0584	0.2653	0.1741
Event-only Model	EventGPT-Plus	43.71	0.3877	0.2704	0.0625	0.3150	0.1133
Event + RGB Model	LITE-Event (Ours)	**80.24**	**0.4857**	**0.4096**	**0.1821**	**0.4288**	**0.4882**

Table 2 shows MCQA results on RealDark-Bench, where models are evaluated exclusively on real low-light RGB-event data. LITE-Event obtains the highest accuracy, 85.00%, outperforming InternVL3+DarkIR by 5.2 points. The gap over the enhancement-then-reason baseline shows that restoring low-light RGB frames does not fully replace event-RGB adaptation. LITE-Event also outperforms EventGPT-Plus, indicating that event streams alone are insufficient because RGB frames still provide appearance and scene context. The results show that combining event-derived temporal cues with RGB appearance improves reasoning under severe illumination and transfers well from predominantly synthetic training to real low-light data.

Table 2: **MCQA accuracy on RealDark-Bench.** LITE-Event achieves the highest accuracy, outperforming all baselines on real-world low-light RGB-event data.

Type	Model	Acc. (%)
Generic	LLaVA-NeXT-Video	65.00
	VideoLLaMA3	74.33
	InternVL3	78.17
	Qwen3-VL-8B-Instruct	78.17
Illumination-enhanced	InternVL3-8B + LEDNet	79.17
	InternVL3-8B + DarkIR	<u>79.83</u>
Supervised RGB-only	InternVL3-DarkSFT	77.50
	InternVL3-DarkSFT+EventLang	78.00
	Qwen3-VL-DarkSFT	77.50
	Qwen3-VL-DarkSFT+EventLang	78.50
Event-only Model	EventGPT-Plus	73.83
Event+RGB Model	LITE-Event (Ours)	**85.00**

5.5 Ablations

We ablate training stages, support aggregation, modality usage, fusion, temporal support, and adaptation strategy in Table 3. Removing Stage 3 causes a large drop in performance, confirming that synthetic pretraining alone is insufficient. A single support batch also lowers accuracy, while averaging residual updates improves stability. RGB-dark-only and event-only variants both underperform the full model, showing the benefit of combining event-derived temporal structure with RGB appearance. Replacing gated fusion with cross-attention reduces both MCQA and captioning performance, suggesting that the lightweight gate is better suited to this low-resource setting. For temporal support, $K = 4$ outperforms $K = 1$, while larger values reduce performance, possibly because they add redundant or noisy support context. Fixed projector variants also underperform the support-conditioned HyperNet. The LoRA-adapted projector improves over direct projector fine-tuning but remains 2.40 points below the full model, showing that support-conditioned adaptation adds gains beyond EventLang supervision alone.

5.6 Qualitative Results

Fig. 4 shows representative qualitative examples. In the multiple-choice example, LITE-Event correctly localizes the chair relative to the walking subject, identifying it as directly in front of him,

Table 3: **Ablation studies and adaptation baselines on the held-out EventLang-Dark split.** The main model uses dual event-RGB-dark input and $K = 4$ temporal support prototypes.

Group	Setting	VQA Acc.	ROUGE-L	BLEU	METEOR
Main model	Full LITE-Event	**80.24**	**0.4096**	**0.1821**	**0.4288**
Training stage	w/o Stage 3 adaptation	71.86	0.3587	0.1452	0.3724
Support aggregation	$M = 1$ source support batch	78.44	0.3978	0.1756	0.4160
Fusion strategy	Cross-attention fusion	77.25	0.3920	0.1715	0.4100
Modality	RGB-Dark only	75.45	0.3820	0.1683	0.4073
	Event only	77.25	0.3937	0.1740	0.4121
Temporal prototypes	$K = 1$	75.45	0.3803	0.1667	0.3951
	$K = 8$	78.44	0.3990	0.1765	0.4154
	$K = 16$	77.25	0.3905	0.1700	0.4052
Adaptation baseline	Projector from scratch	70.66	0.3440	0.1390	0.3520
	Fine-tuned projector	76.05	0.3661	0.1546	0.3745
	LoRA-adapted projector	77.84	0.3848	0.1659	0.4005

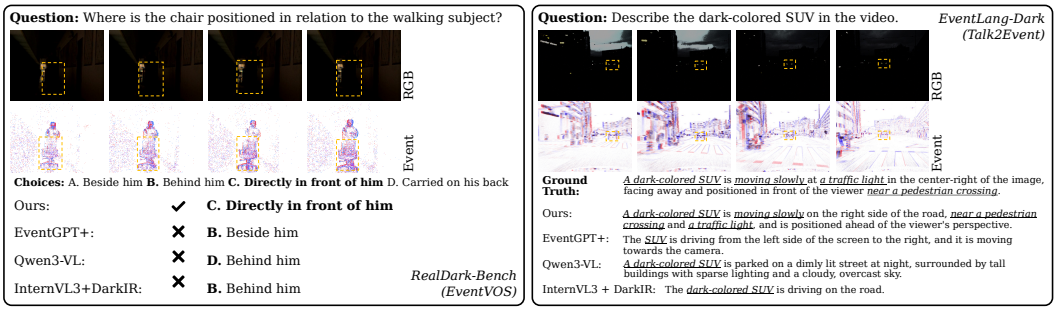


Figure 4: **Qualitative results on RealDark-Bench and EventLang-Dark.** Left: an MCQA example where LITE-Event correctly localizes the chair relative to the walking subject under severe low-light, while baselines confuse the spatial relation. Right: a captioning example where LITE-Event captures the identity, motion, and context of a dark SUV, whereas baselines give generic or incorrect descriptions. Bounding boxes are included for illustrative purposes.

while strong RGB-only and event-only baselines predict incorrect spatial relations under severe low-light degradation. In the captioning example, LITE-Event produces a more faithful description of the dark SUV, including its motion and scene context, whereas the baselines tend to generate generic, static, or partially hallucinated descriptions. These examples are consistent with the quantitative results: combining low-light RGB appearance with event-derived motion and contour cues improves both object recognition and scene understanding in challenging low-light settings.

6 Conclusion

We studied low-light RGB-event video reasoning in a setting where training relies largely on synthetic data and evaluation uses real low-light scenes. We introduced LITE-Event, a sample-efficient framework that combines event and RGB signals through a support-conditioned multimodal interface, together with a staged training pipeline and RealDark-Bench for real-world evaluation. LITE-Event outperforms RGB-only, enhancement-based, and event-only baselines on both EventLang-Dark and RealDark-Bench. The results show that event-derived temporal cues and residual RGB appearance provide complementary information, and that source-conditioned adaptation helps transfer this interface beyond synthetic training data. The proposed datasets and benchmark also provide a testbed for future work on low-light multimodal reasoning.

AI USE STATEMENT

Generative AI tools were used in this work in two roles. First, as part of the research pipeline, we used a multimodal language model to assist with generating and transforming language annotations for our datasets, including converting existing grounding annotations into captions and multiple-choice questions and generating candidate annotations for benchmark construction. All such automatically generated annotations were subsequently reviewed and filtered by the authors according to the procedures described in the paper. Second, generative AI tools were used in a limited supporting role for coding assistance and improving the clarity and presentation of the manuscript. They were not used to formulate the research ideas, design the methodology or experiments, or prepare citations. All AI-assisted code, annotations, and text were reviewed and verified by the authors. The authors take full responsibility for the final manuscript, datasets, experimental results, claims, and other content produced with the assistance of generative AI tools.

REPRODUCIBILITY STATEMENT

We provide detailed descriptions of the model architecture, training protocol, hyperparameters, datasets, and evaluation setup. The Appendix includes the full three-stage training and cross-dataset evaluation algorithm, together with additional implementation and analysis details to facilitate reproducibility. We plan to release the code and evaluation resources upon completion of the review process.

REFERENCES

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: A visual language model for few-shot learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Burak Ercan, Onur Eker, Aykut Erdem, and Erkut Erdem. Hue dataset: High-resolution event and frame sequences for low-light vision. In *European Conference on Computer Vision*, pp. 174–191. Springer, 2024.

Daniel Feijoo, Juan C Benito, Alvaro Garcia, and Marcos V Conde. Darkir: Robust low-light image restoration. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 10879–10889, 2025.

Peng Gao, Jiaming Han, Renrui Zhang, Ziyi Lin, Shijie Geng, Aojun Zhou, Wei Zhang, Pan Lu, Conghui He, Xiangyu Yue, et al. Llama-adapter v2: Parameter-efficient visual instruction model. *arXiv preprint arXiv:2304.15010*, 2023.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Osman Batur İnce, André F. T. Martins, Oisin Mac Aodha, and Edoardo M. Ponti. Sample-efficient integration of new modalities into large language models. *arXiv preprint arXiv:2509.04606*, 2025.

Andrew Jaegle, Felix Gimeno, Andy Brock, Oriol Vinyals, Andrew Zisserman, and Joao Carreira. Perceiver: General perception with iterative attention. In *International Conference on Machine Learning (ICML)*, pp. 4651–4664, 2021.

Taewoo Kim, Jaeseok Jeong, Hoonhee Cho, Yuhwan Jeong, and Kuk-Jin Yoon. Towards real-world event-guided low-light video enhancement and deblurring. In *European Conference on Computer Vision*, pp. 433–451. Springer, 2024.

Lingdong Kong, Dongyue Lu, Ao Liang, Rong Li, Yuhao Dong, Tianshuai Hu, Lai Xing Ng, Wei Tsang Ooi, and Benoit R. Cottereau. Talk2event: Grounded understanding of dynamic scenes from event cameras. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Hebei Li, Jin Wang, Jiahui Yuan, Yue Li, Wenming Weng, Yansong Peng, Yueyi Zhang, Zhiwei Xiong, and Xiaoyan Sun. Event-assisted low-light video object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3250–3259, 2024.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning (ICML)*, 2023.

Pengteng Li, Yunfan Lu, Pinghao Song, Wuyang Li, Huizai Yao, and Hui Xiong. Eventvl: Understand event streams via multimodal large language model. *arXiv preprint arXiv:2501.13707*, 2025.

Guoqiang Liang, Kanghao Chen, Hangyu Li, Yunfan Lu, and Lin Wang. Towards robust event-guided low-light image enhancement: a large-scale real-world event-image dataset and novel approach. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23–33, 2024.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Shaoyu Liu, Jianing Li, Guanghui Zhao, Yunjian Zhang, and Xiangyang Ji. Eventbench: Towards comprehensive benchmarking of event-based mllms. *arXiv preprint arXiv:2511.18448*, 2025a.

Shaoyu Liu, Jianing Li, Guanghui Zhao, Yunjian Zhang, Xin Meng, Fei Richard Yu, Xiangyang Ji, and Ming Li. Eventgpt: Event stream understanding with multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 29139–29149, 2025b.

Shaoyu Liu, Jianing Li, Guanghui Zhao, Yunjian Zhang, Wen Jiang, Ming Li, and Xiangyang Ji. Eventflash: Towards efficient mllms for event-based vision. *arXiv preprint arXiv:2602.03230*, 2026.

Hanyue Lou, Jinxiu Liang, Minggui Teng, Yi Wang, and Boxin Shi. V2v: Scaling event-based vision through efficient video-to-voxel simulation. *arXiv preprint arXiv:2505.16797*, 2025a.

Hanyue Lou, Jiayi Zhou, Yang Zhang, Boyu Li, Yi Wang, Guangnan Ye, and Boxin Shi. Reconstruction as a bridge for event-based visual question answering. *arXiv preprint arXiv:2512.11510*, 2025b.

Yohan Park, Hyunwoo Ha, Wonjun Jo, and Tae-Hyun Oh. Darkeqa: Benchmarking vision-language models for embodied question answering in low-light indoor environments. *arXiv preprint arXiv:2512.24985*, 2025.

Jason Phang. Investigating the effectiveness of hypertuning via gisting. *arXiv preprint arXiv:2402.16817*, 2024.

Jaisidh Singh, Diganta Misra, Boris Knyazev, and Antonio Orvieto. (almost) free modality stitching of foundation models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 19784–19800, 2025.

Wentao Wu, Xiao Wang, Chenglong Li, Bo Jiang, Jin Tang, Bin Luo, and Qi Liu. Cm3ae: A unified rgb frame and event-voxel/-frame pre-training framework. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 2159–2168, 2025.

Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. Next-qa: Next phase of question-answering to explaining temporal actions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9777–9786, 2021.

Wenhao Xu, Xin Dong, Yue Li, Haoyuan Shi, and Zhiwei Xiong. Eventstu: Event-guided efficient spatio-temporal understanding for video large language models. *arXiv preprint arXiv:2511.18920*, 2025.

Baoheng Zhang, Jiahui Liu, Gui Zhao, Weizhou Zhang, Yixuan Ma, Jun Jiang, Yingxian Chen, Wilton W. T. Fok, Xiaojuan Qi, and Hayden Kwok-Hay So. Learning to see through illumination extremes with event streaming in multimodal large language models. *arXiv preprint arXiv:2603.27558*, 2026a.

Boqiang Zhang, Kehan Li, Zesen Cheng, Zhiqiang Hu, Yuqian Yuan, Guanzheng Chen, Sicong Leng, Yuming Jiang, Hang Zhang, Xin Li, Peng Jin, Wenqi Zhang, Fan Wang, Lidong Bing, and Deli Zhao. Videollama 3: Frontier multimodal foundation models for image and video understanding. *arXiv preprint arXiv:2501.13106*, 2025.

Deheng Zhang, Yuqian Fu, Runyi Yang, Yang Miao, Tianwen Qian, Xu Zheng, Guolei Sun, Ajad Chhatkuli, Xuanjing Huang, Yu-Gang Jiang, Luc Van Gool, and Danda Pani Paudel. Egonight: Towards egocentric vision understanding at night with a challenging benchmark. In *International Conference on Learning Representations (ICLR)*, 2026b.

Wenqiao Zhang, Tianwei Lin, Jiang Liu, Fangxun Shu, Haoyuan Li, Lei Zhang, He Wanggui, Hao Zhou, Zheqi Lv, Hao Jiang, et al. Hyperllava: Dynamic visual and language expert tuning for multimodal large language models. *arXiv preprint arXiv:2403.13447*, 2024a.

Yuanhan Zhang, Bo Li, Haotian Liu, Yong Jae Lee, Liangke Gui, Di Fu, Jiashi Feng, Ziwei Liu, and Chunyuan Li. Llava-next: A strong zero-shot video understanding model, 2024b. URL https://llava-vl.github.io/blog/2024-04-30-llava-next-video/.

Hanyu Zhou and Gim Hee Lee. Llafea: Frame-event complementary fusion for fine-grained spatiotemporal understanding in lmms. *arXiv preprint arXiv:2503.06934*, 2025.

Shangchen Zhou, Chongyi Li, and Chen Change Loy. Lednet: Joint low-light enhancement and deblurring in the dark. In *European conference on computer vision*, pp. 573–589. Springer, 2022.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Yuchen Duan, Hao Tian, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

SUPPLEMENTARY MATERIAL AND TECHNICAL APPENDICES

In this supplementary material, we provide:

- Support-set sensitivity and inference latency analyses (Sec. A).
- Additional qualitative results on RealDark-Bench (Sec. B).
- The full LITE-Event training and cross-dataset evaluation algorithm (Sec. C).
- Detailed analysis of model performances on RealDark-Bench performance (Sec. D).
- Representative failure cases under severe low-light degradation (Sec. E).
- A discussion of limitations, including the dependence on RGB for static appearance cues and possible biases in the semi-automatic benchmark construction (Sec. F).
- Broader-impact considerations for low-light multimodal reasoning systems (Sec. G).

A Support-Set Sensitivity and Inference Latency

A.1 Support-set sampling sensitivity

Table A.1 evaluates whether LITE-Event is sensitive to the choice of source support examples. All strategies use only EventLang-Dark training examples as support; no RealDark-Bench example is used for support construction, model selection, or tuning. A single fixed support batch introduces moderate variance, whereas averaging adapters over multiple random source batches substantially stabilizes performance. Text-similarity-based source selection does not outperform broad random coverage, suggesting that the adaptation gain is not driven by lexical overlap between source support questions and target questions.

Table A.1: **Support-set sensitivity on RealDark-Bench.** All support examples are sampled only from EventLang-Dark training data. Deviations are computed over five support choices or seeds.

Support strategy	RealDark-Bench Acc. (%)
Single fixed source batch, reused for all target questions	82.9 ± 0.8
Random source support, averaging $M = 31$ adapters	85.0 ± 0.4
Text-similarity-based source support selection	84.6 ± 0.3

These results show that the model is not dependent on a lucky support draw. The higher variance of a single small support batch indicates that source support choice does matter, but averaging adapters over $M = 31$ source batches reduces this variance. The similarity-based strategy also remains below random multi-batch coverage, indicating that lexical similarity is not the main driver of adaptation quality.

A.2 Inference latency

We measure per-sample inference latency on the same NVIDIA RTX A6000 GPU with batch size 1, deterministic decoding, identical generation limits, and the same evaluated samples. For LITE-Event, latency includes query-side event/RGB projection, gated fusion, residual adaptation, and frozen LLM decoding. The support-derived residual projector is generated once from EventLang-Dark support batches and reused during evaluation, so this one-time cost is amortized across target queries.

Table A.2: **Inference latency comparison.** Mean latency is measured per sample under the same decoding protocol.

Model	Mean latency / sample
EventGPT-Plus	0.552s
LITE-Event	0.688s
Qwen3-VL	1.150s
LLaVA-NeXT-Video	1.496s
InternVL3	1.815s
InternVL3 + DarkIR	2.568s
VideoLLaMA3	6.015s

For LITE-Event, the median latency is 0.582s and the 95th percentile is 1.129s. Thus, the proposed event-RGB interface does not introduce a substantial inference-time penalty relative to the evaluated video MLLMs. It is slightly slower than the event-only EventGPT-Plus, but faster than RGB video MLLM and enhancement-then-reason baselines under the same decoding protocol.

B ADDITIONAL QUALITATIVE RESULTS

Fig. A.1 and Fig. A.2 provide additional qualitative examples from RealDark-Bench. These examples illustrate cases where low-light RGB frames alone are severely degraded, while event frames preserve complementary structural or temporal evidence. Across the examples, LITE-Event more consistently selects the correct answer by combining residual appearance cues from RGB with contour, motion, and high-contrast structure from the event stream.

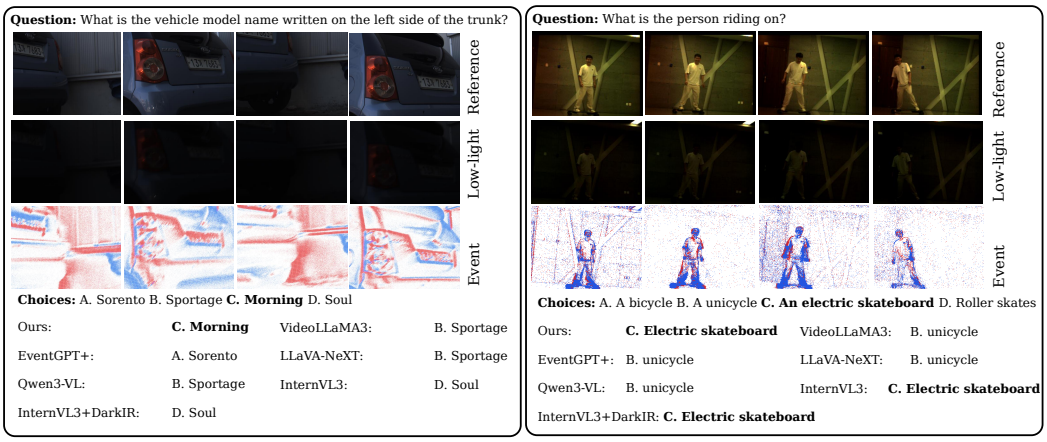


Figure A.1: **Additional qualitative results on RealDark-Bench.** Left: a fine-grained recognition example where LITE-Event correctly identifies the vehicle model name as "Morning" despite severe low-light degradation. Right: an object/action example where LITE-Event correctly recognizes an electric skateboard, while several baselines confuse it with a unicycle.

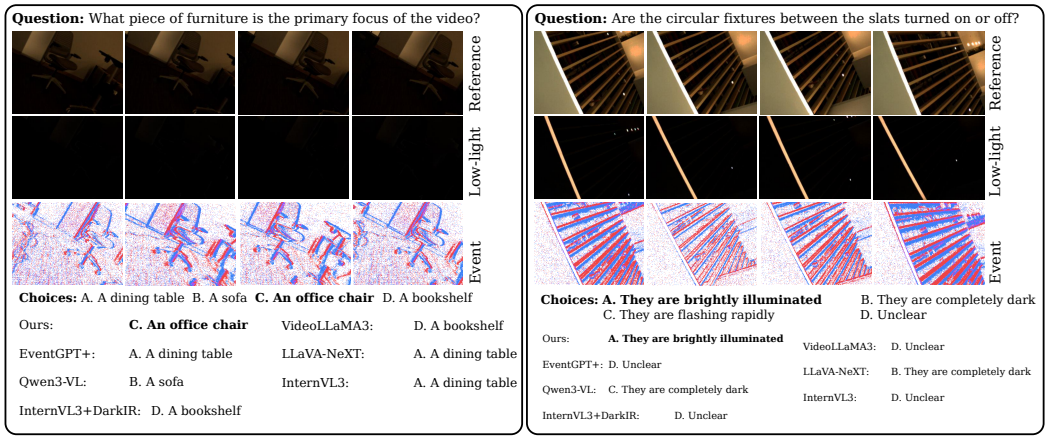


Figure A.2: **Additional qualitative results on RealDark-Bench.** Left: LITE-Event identifies an office chair by combining weak RGB appearance with event-derived object structure. Right: LITE-Event correctly reasons that the circular fixtures are brightly illuminated, whereas several baselines either predict "unclear" or mistake them as dark.

C Training and Cross-Dataset Evaluation Algorithm

Algorithm 1 summarizes the full LITE-Event training and evaluation protocol. The model is trained in three stages: projector alignment on synthetic NExT-QA-Dark features, hypernetwork pretraining on the same synthetic source, and source event-language adaptation on EventLang-Dark. For cross-dataset evaluation, RealDark-Bench is treated strictly as a held-out target benchmark: no RealDark-Bench examples are used for support construction, adapter generation, or parameter updates. Instead, source adapters are generated only from EventLang-Dark training support batches, averaged into a fixed residual projector, and then applied to RealDark-Bench queries.

Algorithm 1 LITE-Event training and cross-dataset evaluation

Require: Source event-language data $\mathcal{D}_{\mathrm{ELD}}$ (EventLang-Dark), target benchmark $\mathcal{D}_{\mathrm{RDB}}$ (RealDark-Bench), support batch size S, temporal prototypes K, number of source adapters M
Ensure: Predictions on $\mathcal{D}_{\mathrm{RDB}}$ without RealDark-Bench support or fine-tuning

Stage 1: Projector alignment
Train event projector P_e, RGB projector P_r, and fusion gate G on NExT-QA-Dark synthetic event/RGB-dark features
Freeze the Stage-1 visual interface (P_e, P_r, G)

Stage 2: Hypernetwork pretraining
for each synthetic update **do**
    Sample query batch $\mathcal{B}$ and support set $\mathcal{S}$ from NExT-QA-Dark
    Build support prototypes $(E_s, R_s) \in \mathbb{R}^{S\times K\times d_m}$
    Encode support with dual-stream cross-attention to obtain $C_s \in \mathbb{R}^{S\times K\times d_h}$
    Flatten C_s into SK support context tokens
    Hypernetwork predicts residual projector parameters $A(\mathcal{S}), B(\mathcal{S}), b(\mathcal{S})$
    Adapt query visual tokens and optimize answer-token LM loss
end for

Stage 3: Source event-language adaptation
Initialize from the Stage-2 checkpoint
for each source update **do**
    Sample query batch $\mathcal{B}_{\mathrm{ELD}}$ and support set $\mathcal{S}_{\mathrm{ELD}}$ from $\mathcal{D}^{\mathrm{train}}_{\mathrm{ELD}}$
    Build support prototypes $(E_s, R_s) \in \mathbb{R}^{S\times K\times d_m}$
    Encode support with dual-stream cross-attention and predict $A(\mathcal{S}_{\mathrm{src}}), B(\mathcal{S}_{\mathrm{src}}), b(\mathcal{S}_{\mathrm{src}})$
    Adapt source query visual tokens and optimize answer-token LM loss
end for

Cross-dataset evaluation
Do not sample support from $\mathcal{D}_{\mathrm{RDB}}$ and **do not** update any model parameters on $\mathcal{D}_{\mathrm{RDB}}$
for $m = 1, \ldots, M$ **do**
    Sample source support batch $\mathcal{S}^{(m)}_{\mathrm{ELD}} \subset \mathcal{D}^{\mathrm{train}}_{\mathrm{ELD}}$
    Build prototypes $(E^{(m)}_s, R^{(m)}_s) \in \mathbb{R}^{S\times K\times d_m}$
    Encode support with dual-stream cross-attention
    Hypernetwork predicts $A^{(m)}, B^{(m)}, b^{(m)}$
end for
Average the generated low-rank parameters across the M source support batches and merge them into a fixed residual projector $\bar{P}_{\mathrm{res}}$
for each RealDark-Bench query $(E_{\mathrm{rdb}}, R_{\mathrm{rdb}}, q) \in \mathcal{D}_{\mathrm{RDB}}$ **do**
    Project event and RGB tokens:

$$U_{\mathrm{rdb}} = P_e(E_{\mathrm{rdb}}), \qquad V_{\mathrm{rdb}} = P_r(R_{\mathrm{rdb}})$$

    Fuse projected tokens:

$$Z_{\mathrm{rdb}} = G(U_{\mathrm{rdb}}, V_{\mathrm{rdb}})$$

    Apply merged residual projector:

$$\hat{Z}_{\mathrm{rdb}} = \bar{P}_{\mathrm{res}}(Z_{\mathrm{rdb}})$$

    Prepend $\hat{Z}_{\mathrm{rdb}}$ to the text prompt q and decode with the frozen LLM
end for

D Detailed Analysis of Models on RealDark-Bench

Table A.3 reports accuracy by question type on RealDark-Bench. LITE-Event achieves the best performance on spatial, object, and visibility questions, and ties for the best action accuracy. Supervised RGB-only and enhancement-based baselines remain competitive on temporal, OCR, and counting questions, while EventGPT-Plus performs strongest on counting. Overall, the breakdown suggests that event-RGB adaptation is most beneficial when low-light reasoning requires complementary appearance, spatial layout, and event-derived structure.

Table A.3: **Accuracy by question type on RealDark-Bench.** Best results are shown in bold and second-best results are underlined.

Model	Spatial ($n = 180$)	Object ($n = 183$)	Temporal ($n = 83$)	Visibility ($n = 68$)	Action ($n = 53$)	OCR ($n = 17$)	Counting ($n = 16$)
LITE-Event (Ours)	**87.8**	**87.4**	73.5	**97.1**	**79.2**	70.6	68.8
LLaVA-NeXT	60.6	68.9	61.4	83.8	66.0	52.9	18.8
VideoLLaMA3	80.6	74.3	69.9	82.4	60.4	52.9	62.5
InternVL3	80.0	<u>81.4</u>	69.9	86.8	69.8	70.6	62.5
Qwen3-VL-8B-Instruct	78.9	<u>81.4</u>	73.5	83.8	73.6	70.6	56.2
InternVL3-8B+DarkIR	<u>83.3</u>	78.1	**75.9**	<u>94.1</u>	67.9	70.6	68.8
InternVL3-DarkSFT	75.0	79.8	<u>74.7</u>	91.2	67.9	70.6	<u>75.0</u>
InternVL3-DarkSFT+EventLang	76.1	80.9	66.3	92.6	77.4	**76.5**	68.8
Qwen3-VL-DarkSFT	78.9	80.9	69.9	82.4	71.7	**76.5**	62.5
Qwen3-VL-DarkSFT+EventLang	79.4	<u>81.4</u>	71.1	82.4	73.6	**76.5**	<u>75.0</u>
EventGPT-Plus	75.0	72.7	68.7	82.4	**79.2**	29.4	**93.8**

Table A.4 reports accuracy separately for each RealDark-Bench source dataset. LITE-Event performs best on EventVOS, SDE, and RELED, while RGB video MLLMs are stronger on HUE. This suggests that HUE contains challenging high-resolution appearance-heavy scenes where event cues alone are less decisive, whereas LITE-Event is particularly effective on sources where low-light RGB and event streams provide complementary temporal and structural evidence.

Table A.4: **Per-source accuracy on RealDark-Bench.** Best results are shown in bold and second-best results are underlined.

Model	EventVOS ($n = 194$)	SDE ($n = 233$)	RELED ($n = 98$)	HUE ($n = 75$)	Overall ($n = 600$)
LLaVA-NeXT	61.9	67.8	71.4	56.0	65.0
VideoLLaMA3	70.1	75.5	78.6	<u>76.0</u>	74.3
InternVL3	74.2	76.8	88.8	**78.7**	78.2
Qwen3-VL-8B-Instruct	73.7	79.0	85.7	77.3	78.2
InternVL3-DarkSFT	73.2	76.8	88.8	<u>76.0</u>	77.5
InternVL3-DarkSFT+EventLang	74.7	79.0	87.8	70.7	78.0
Qwen3-VL-DarkSFT	72.2	77.7	88.8	<u>76.0</u>	77.5
Qwen3-VL-DarkSFT+EventLang	74.2	78.1	<u>89.8</u>	<u>76.0</u>	78.5
InternVL3-8B+DarkIR	71.1	<u>83.3</u>	91.8	<u>76.0</u>	<u>79.8</u>
EventGPT-Plus	<u>74.7</u>	78.5	75.5	54.7	73.8
LITE-Event (Ours)	**81.4**	**90.6**	**93.9**	65.3	**85.0**

Table A.5 ablates the effect of Stage-3 event-language adaptation, temporal support granularity, and source support aggregation on RealDark-Bench. Stage-3 adaptation provides the largest gain, while $K = 4$ gives the best temporal support trade-off. Averaging adapters over $M = 31$ source support batches also improves over using a single support batch.

Table A.5: **Adaptation sensitivity on RealDark-Bench.**

Setting	Acc. (%)
Full LITE-Event, $K = 4$, $M = 31$	**85.0**
w/o Stage-3 adaptation	77.8
$K = 1$	80.7
$K = 8$	83.2
$K = 16$	82.3
$M = 1$ source support batch	83.4

E FAILURE CASES

Fig. A.3 shows representative failure cases where LITE-Event does not outperform a strong RGB-centric video MLLM. In both examples, Qwen3-VL predicts the correct answer, while LITE-Event and the other baselines fail. These cases highlight a limitation of event–RGB reasoning under severe low-light conditions: event streams preserve edges, contours, and temporal changes, but they provide limited information about static appearance cues such as text, material, reflectance, and fine-grained object identity.

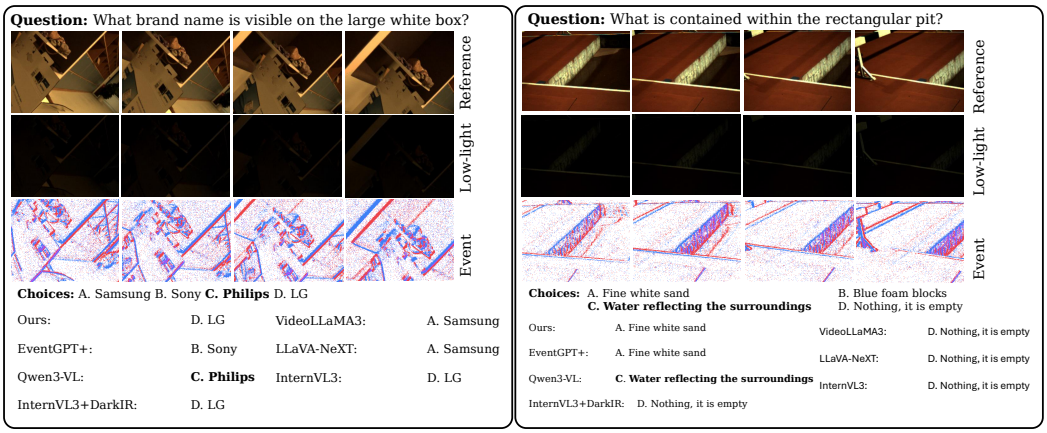


Figure A.3: **Failure cases on RealDark-Bench.** In both examples, Qwen3-VL is the only model that predicts the correct answer, while LITE-Event and the other baselines fail. Left: the question requires recognizing a small brand/logo cue on a low-light object; LITE-Event predicts "LG" instead of the correct answer "Philips". Right: the question requires distinguishing the material inside a pit; LITE-Event predicts "fine white sand" instead of "water reflecting the surroundings". These cases indicate that event streams help preserve geometric and temporal structure, but static appearance cues such as text, logo identity, material, and reflectance remain challenging when the low-light RGB signal is degraded.

F LIMITATIONS

LITE-Event inherits limitations from both sensing modalities. Event streams capture temporal changes and edges but provide limited static appearance information. Tasks such as fine-grained object recognition, color identification, OCR, and detailed scene understanding therefore remain dependent on RGB, especially when the RGB signal is severely degraded.

RealDark-Bench is built through a semi-automatic annotation pipeline that combines model-generated candidate questions with independent human verification. The filtering process removes ambiguous and unsupported items, but the benchmark may still contain biases from the source datasets, prompt design, reasoning-category coverage, and model-generated candidates.

G BROADER IMPACT

Improved low-light multimodal reasoning can benefit safety-critical and socially useful applications such as night-time robotics, assistive perception, search-and-rescue, autonomous navigation, and environmental monitoring. Event cameras are particularly relevant in these settings because they can preserve motion and structural cues when conventional RGB sensors degrade under poor illumination.

At the same time, better perception in low-visibility environments may raise privacy and surveillance concerns. Models that improve reasoning from dark or low-quality video could be misused for covert monitoring or expanded visual surveillance. Our benchmark is constructed from existing research datasets and does not introduce identity labels, biometric annotations, or person-level recognition tasks.

The benchmark annotations are generated through a semi-automatic pipeline and manually validated, but they may still reflect biases in the source datasets, prompt design, and model-generated candidate questions. We therefore view RealDark-Bench as a research benchmark for evaluating multimodal reasoning under low-light conditions, not as a comprehensive measure of real-world safety or fairness. Future work should further study robustness across diverse environments, sensor types, geographic conditions, and human-centered use cases before deployment in high-stakes settings.