EgoNight: Towards Egocentric Vision Understanding at Night with a Challenging Benchmark

Anonymous authors
Paper under double-blind review

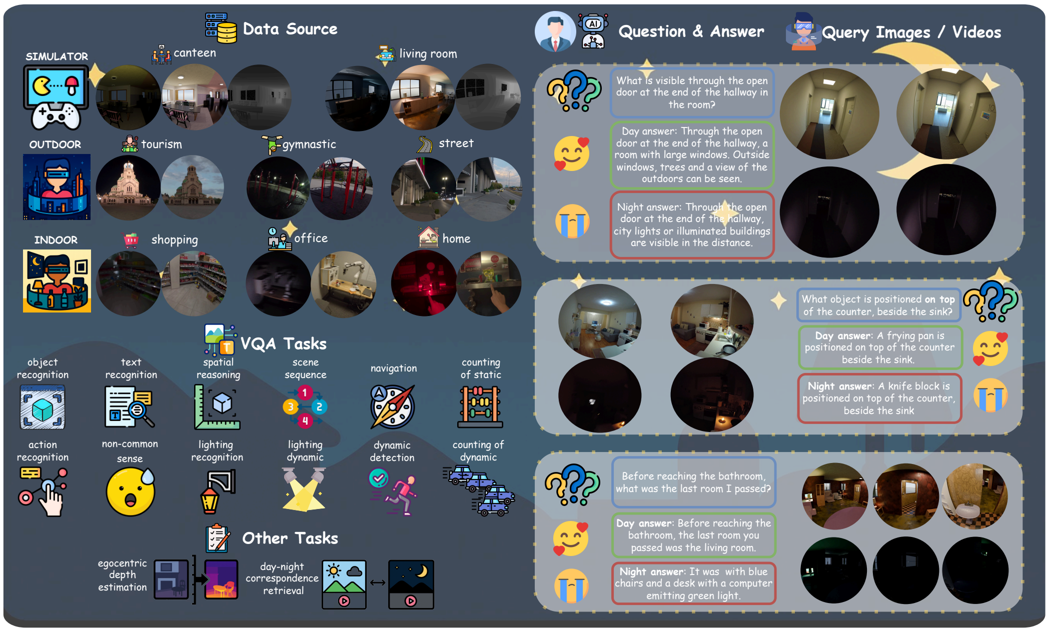

Data Source
SIMULATOR
canteen
living room
OUTDOOR
tourism
gymnastic
street
INDOOR
shopping
office
home
VQA Tasks
object recognition
text recognition
spatial reasoning
scene sequence
navigation
counting of static
action recognition
non-common sense
lighting recognition
lighting dynamic
dynamic detection
counting of dynamic
Other Tasks
egocentric depth estimation
day-night correspondence retrieval
Question & Answer
Query Images / Videos
What is visible through the open door at the end of the hallway in the room?
Day answer: Through the open door at the end of the hallway, a room with large windows. Outside windows, trees and a view of the outdoors can be seen.
Night answer: Through the open door at the end of the hallway, city lights or illuminated buildings are visible in the distance.
What object is positioned on top of the counter, beside the sink?
Day answer: A frying pan is positioned on top of the counter beside the sink.
Night answer: A knife block is positioned on top of the counter, beside the sink
Before reaching the bathroom, what was the last room I passed?
Day answer: Before reaching the bathroom, the last room you passed was the living room.
Night answer: It was with blue chairs and a desk with a computer emitting green light.

Figure 1: **Overview of the EgoNight.** EgoNight integrates diverse video sources spanning synthetic environments, real-world indoor and outdoor scenes, recorded under both daytime and nighttime conditions, with spatial and temporal alignment. It consists of three benchmarks: (i) *egocentric VQA* as the primary focus, (ii) *day–night correspondence retrieval*, and (iii) *egocentric depth estimation*, all targeting the challenges of low-light egocentric vision. The day–night alignment (illustrated on the right with VQA examples) enables rigorous analysis of illumination gaps in MLLMs.

Abstract

Most existing benchmarks for egocentric vision understanding focus primarily on daytime scenarios, overlooking the low-light conditions that are inevitable in real-world applications. To investigate this gap, we present **EgoNight**, the first comprehensive benchmark for nighttime egocentric vision, with visual question answering (VQA) as the core task. A key feature of EgoNight is the introduction of day–night aligned videos, which enhance night annotation quality using the daytime data and reveal clear performance gaps between lighting conditions. To achieve this, we collect both synthetic videos rendered by Blender and real-world recordings, ensuring that scenes and actions are visually and temporally aligned. Leveraging these paired videos, we construct **EgoNight-VQA**, supported by a novel day-augmented night auto-labeling engine and refinement through extensive human verification. Each QA pair is double-checked by annotators for reliability. In total, EgoNight-VQA contains 3658 QA pairs across 90 videos, spanning 12 diverse QA types, with more than 300 hours of human work. Evaluations of the state-of-the-art multimodal large language models (MLLMs) reveal substantial performance drops when transferring from day to night, underscoring the challenges of reasoning under low-light conditions. Beyond VQA, EgoNight also introduces two auxiliary tasks, *day–night correspondence retrieval* and *egocentric depth estimation at night*, that further explore the boundaries of existing models. We believe EgoNight-VQA provides a strong foundation for advancing application-driven egocentric vision research and for developing models that generalize across illumination domains. All the data and code will be made available upon acceptance.

1 Introduction

With the rapid development of wearable devices, egocentric vision understanding has become increasingly important. Unlike third-person vision, egocentric perception naturally aligns with the way humans perceive, understand, and interact with the world. A robust egocentric vision system can not only serve as an intelligent assistant in daily activities Yang et al. (2025) but also play a crucial role in embodied AI and robotic learning Li et al. (2025a); Kareer et al. (2025). Beyond these general applications, egocentric vision holds unique potential for assisting specific user groups such as people who are blind or visually impaired Xiao et al. (2025), or physically disabled Zhang et al. (2023a), enabling technologies that enhance navigation, accessibility, and real-time scene understanding.

Significant efforts have been made to advance egocentric vision understanding, including the construction of large-scale ego-centric datasets such as EPIC-KITCHENS Damen et al. (2020), Ego4D Grauman et al. (2022), and Ego-Exo4D Grauman et al. (2024); the design of diverse and challenging benchmarks such as EgoTaskQA Jia et al. (2022), EgoSchema Mangalam et al. (2023), and EgoTempo Plizzari et al. (2025); and the development of egocentric multimodal large language models (MLLMs) such as EgoVLPv2 Pramanick et al. (2023), EgoGPT Yang et al. (2025), and Exo2Ego Zhang et al. (2025a). Despite these advances, almost all prior works focus on daytime scenarios with favorable lighting. In contrast, real-world egocentric systems, for example, intelligent personal assistants for navigation, must operate at night, under low light, uneven illumination, and severely limited visibility. This motivates us to investigate egocentric vision at night, focusing on complex scene understanding and reasoning tasks.

A central challenge in constructing such a benchmark lies in obtaining suitable video sources that capture the characteristics of nighttime environments and developing annotation methods that ensure high labeling quality. To address this, we place particular emphasis on *day–night aligned videos*, which not only allow us to leverage daytime data to annotate nighttime videos, but also enable rigorous performance comparisons across day and night lighting conditions. However, in practice, collecting perfectly aligned day–night pairs in the real world is highly non-trivial. To overcome this, we leverage Blender Iraci (2013), where scene layouts, camera trajectories, and lighting can be precisely controlled, enabling the synthesis of the desired videos. This produces **EgoNight-Synthetic**, a collection of 50 ideally aligned egocentric pairs spanning diverse and complex indoor scenarios with varying illumination levels. To complement synthetic data with real-world evidence, we design a *video-guided recording protocol* to construct **EgoNight-Sofia**, which contains 20 pairs of real-world egocentric videos with spatially and temporally aligned day–night counterparts. These videos cover realistic use cases (e.g., “Where did I put my keys?”, “How much is the item I saw in the grocery shop?”), spanning both indoor and outdoor environments under diverse illumination sources such as streetlights, flashlights, and candles. Finally, we incorporate 20 nighttime videos from the Oxford Day-and-Night dataset Wang et al. (2025b), termed **EgoNight-Oxford**, which serve as an additional testbed despite lacking day-night alignment. Together, these three video sources constitute our **EgoNight** dataset, which is the first egocentric dataset providing day–night aligned correspondences, as mainly summarized in Fig. 1.

The videos in EgoNight pave the way for constructing challenging benchmarks to evaluate the capabilities of existing models. Among many egocentric tasks, we focus on the egocentric video question answering, a flagship task that best reflects high-level understanding in egocentric vision. Specifically, to comprehensively evaluate model abilities, we first propose a diverse set of QA types, spanning well-studied tasks (e.g., object recognition, spatial reasoning, action recognition, counting, text recognition) as well as several underexplored dimensions (e.g., temporal scene sequence understanding, navigation, lighting recognition, and non–common-sense reasoning). These are further organized into paired and unpaired QA types, depending on whether day–night counterparts share the same questions and answers. To construct the benchmark at scale, we then develop a *novel three-stage day-augmented auto-labeling pipeline* that leverages daytime videos to assist in generating question–answer pairs for nighttime clips, followed by extensive human verification to ensure accuracy and reliability. Building EgoNight and annotating VQA required **over 300 hours of human effort**, with each QA pair verified by at least one expert annotator. This process results in the high-quality **EgoNight-VQA** dataset, comprising 3,658 QA pairs. Beyond VQA, we introduce two auxiliary tasks with dedicated testbeds: day–night correspondence retrieval, which evaluates cross-illumination matching, and egocentric depth estimation at night, which is crucial for navigation and interaction in embodied AI. These two tasks further broaden the benchmark and expose new challenges for existing models.

Our extensive experiments across three video sources, three tasks, and 10 state-of-the-art multimodal large language models reveal that nearly all models (including closed-source models such as GPT and Gemini) struggle on this challenging benchmark, with a clear and consistent performance gap between day and night. This highlights the unsolved challenges of egocentric vision at nighttime and calls for more robust models that generalize across illumination conditions. Besides, we highlight that our newly introduced QA types, covering lighting recognition/dynamic, scene sequence reasoning, navigation, and non–common-sense reasoning, are substantially more challenging than well-studied categories, revealing fresh difficulties for MLLMs. We further prove synthetic data is highly correlated with real data and effectively boosts real-world performance through fine-tuning. Our pilot studies further show that fine-tuning on specialized subset of data improves model performance through adapting vision encoder into low-light domain and aligning the language model to the uncertain features during night.

Our main contributions are threefold: i) **EgoNight Dataset**: We present the first egocentric dataset that systematically addresses nighttime conditions, featuring day–night aligned videos from synthetic (EgoNight-Synthetic), real-world (EgoNight-Sofia), and existing (EgoNight-Oxford) sources. ii) **Benchmark Suite**: We build a comprehensive benchmark centered on egocentric VQA with diverse QA types and 3658 fully human-verified QA pairs, complemented by egocentric depth estimation at night and day–night correspondence retrieval tasks. iii) **Empirical Insights**: Extensive evaluations reveal clear day–night performance gaps, underscoring illumination robustness as a key challenge; our newly proposed QA types are also validated to pose practical difficulties for current MLLMs.

2 Related Works

2.1 Egocentric Datasets and VQA Benchmarks

A series of large-scale egocentric datasets, such as EPIC-KITCHENS Damen et al. (2020), Ego4D Grauman et al. (2022), Ego-Exo4D Grauman et al. (2024), and EgoExoLearn Huang et al. (2024), have laid the foundation for a wide range of tasks, including action recognition Sudhakaran et al. (2019), object detection Ren & Gu (2010), pose estimation Luo et al. (2021), video generation Liu et al. (2021), Ego-Exo correspondence Fu et al. (2025). Among these, we are particularly interested in egocentric visual question answering (VQA) Fan (2019), which provides a natural and human-like framework for comprehensively evaluating model performance through question–answer interactions. In recent years, several egocentric VQA benchmarks have been proposed, including EgoVQA Fan (2019), EgoTaskQA Jia et al. (2022), EgoSchema Mangalam et al. (2023), EgoThink Cheng et al. (2024), EgoTempo Plizzari et al. (2025), EgoCross Li et al. (2025b), EgoBlind Xiao et al. (2025), EgoMemoria Ye et al. (2024), HourVideo Chandrasegaran et al. (2024), EgoLifeQA Yang et al. (2025) with different focuses. However, nearly all of them are confined to daytime or well-lit scenarios, leaving model performance in low-light or nighttime conditions largely unexplored. The Oxford Day-and-Night dataset Wang et al. (2025b) is a partial exception but was not designed for VQA and lacks day–night alignment. This makes EgoNight and EgoNight-VQA fundamentally distinct from prior benchmarks.

2.2 MLLMs for Video Understanding

The rapid development of multimodal large language models (MLLMs) has substantially advanced the frontier of video understanding. Prominent open-source models include Qwen-VL Bai et al. (2023), InternVL Chen et al. (2024b), Video-LLaMA Zhang et al. (2023b), LLaVA-NeXT-Video Li et al. (2024), and GLM-V Hong et al. (2025), while closed-source commercial systems such as GPT-4V Achiam et al. (2023) and Gemini Comanici et al. (2025) demonstrate even stronger capabilities in video captioning, summarization, and open-ended visual question answering. Building on these advances, egocentric MLLMs have emerged to adapt foundation models from exocentric to first-person perspectives. Representative examples include EgoVLPv2 Pramanick et al. (2023) for improved video–language cross-modal fusion, EgoGPT Yang et al. (2025) fine-tuned with egocentric captioning and QA, MM-Ego Ye et al. (2024) with a memory mechanism for long videos, and Exo2Ego Zhang et al. (2025a) leveraging exocentric data for egocentric generalization. These works highlight the potential of MLLMs as egocentric assistants. However, nearly all of them are developed and tested under well-lit daytime conditions, leaving their robustness in low-light or nighttime scenarios unexplored. Nevertheless, almost all existing MLLMs are developed and evaluated under well-lit daytime conditions, with little consideration of low-light or nighttime videos, leaving their robustness in low-light or nighttime scenarios unexplored.

2.3 Cross-Domain Generalization

Domain generalization Zhou et al. (2022) is a long-standing challenge in computer vision, where models trained on one distribution must adapt to another. Shifts can arise from semantic drift, style changes, or variations in weather and lighting. Many algorithms have been validated across tasks such as image classification Li et al. (2017); Zhou et al. (2021), object detection Fu et al. (2024); Li et al. (2025c), action recognition Pan et al. (2020); Bian et al. (2011), few-shot learning Guo et al. (2020); Fu et al. (2021), and autonomous driving Li et al. (2022; 2023a). In contrast, cross-domain transfer for MLLMs, especially in video understanding, remains underexplored, with only a few recent attempts (e.g., CL-CrossVQA Zhang et al. (2025b), VQA-GEN Unni et al. (2023), Super-CLEVR Li et al. (2023c)). However, none of them are targeted for egocentric video, which is naturally different from exocentric videos in terms of recorded images, camera motion, and contained information. The most relevant effort to us is EgoCross Li et al.

(2025b), an egocentric VQA benchmark that moves beyond daily activities to evaluate model generalization across distinct long-tail, specialized domains such as surgery and industrial settings. In this paper, however, we investigate MLLMs from a different perspective, robustness under nighttime conditions, a common and ubiquitous scenario in daily life, yet previously overlooked dimension of domain generalization in egocentric video understanding.

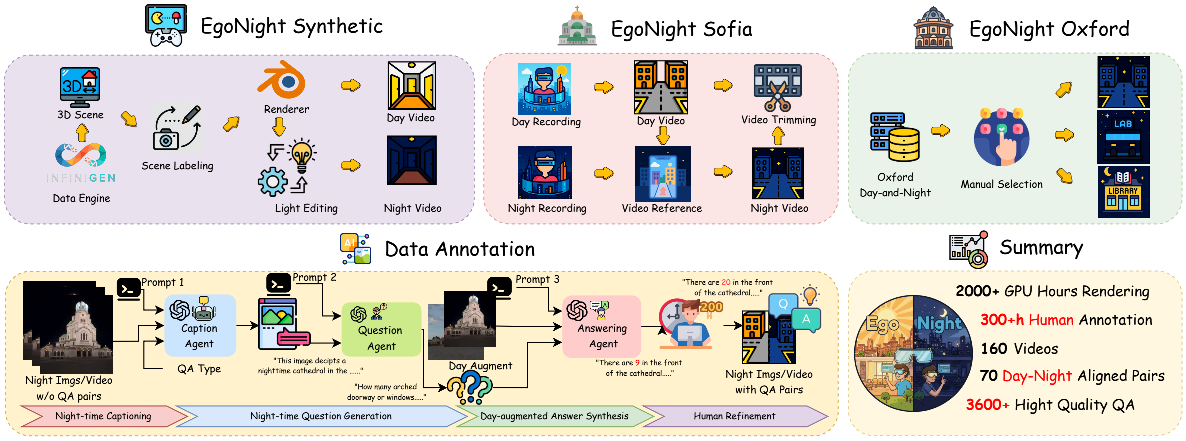


Figure 2: **EgoNight construction and EgoNight-VQA annotation.** EgoNight integrates EgoNight-Synthetic, EgoNight-Sofia, and EgoNight-Oxford sources. Annotation is achieved via a novel three-stage day-augmented Auto QA generation pipeline with 300+ hours of human refinement, resulting in over 3600 high-quality QA pairs.

3 EgoNight Dataset & Benchmarks

3.1 Video Source Collection

Overview & Design Principles. EgoNight is built to systematically evaluate MLLMs under challenging nighttime conditions, which are critical for developing robust intelligent assistants. The collection of video sources follows four principles: ① *Reflect real-world challenges*, such as walking on dimly lit streets or navigating indoors during power outages; ② Involve *natural camera movements* and preferably capture *actions and interactions* with the environment to evaluate both static perception and dynamic understanding; ③ Ensure *diversity of scenarios, illumination, and task difficulties*, spanning indoor, outdoor, office, and grocery settings, lighting from streetlights, flashlights, headlights, and candles, and task levels from easy (relatively clear), through medium (partially visible), to hard (barely visible). ④ Enable rigorous analysis through *day–night paired videos*, where scenes, trajectories, and actions remain consistent across conditions so that differences can be attributed solely to illumination. To meet these requirements, EgoNight integrates three complementary video sources, as illustrated in the upper part of Fig. 2 and detailed below.

EgoNight-Synthetic. To obtain perfectly aligned day–night video pairs, we used a simulation environment where every element can be precisely controlled, including the scene layout, camera path, and lighting. This ensures that the day and night videos match exactly at the pixel and frame level, with lighting being the only difference. We first employ Infinigen Raistrick et al. (2023) to generate diverse indoor 3D scenes. Human annotators cleaned and refined these scenes, then simulated walking through the space at a normal speed (1.2 m/s), recording the camera trajectory. We replayed the same trajectory under different lighting conditions. Daytime videos were rendered using Blender Iraci (2013), and we adjusted the lighting to create the corresponding nighttime versions.

In total, EgoNight-Synthetic contains 50 pairs of egocentric videos, covering more than 100 environment assets. These include indoor scenes such as kitchens, bathrooms, and living rooms, populated with over 50 diverse object categories (e.g., windows, tables, beds, chairs, lamps, bookshelves, plates). We design multiple illumination setups, ranging from uniformly lit rooms to sparsely localized lighting, and incorporate three difficulty levels with a different range of motion blur, sensor noise, and illumination level. Besides RGB frames, Blender also allows us to generate ground-truth depth and normals (see Appendix Sec. A.1), making EgoNight-Synthetic richer and more versatile.

EgoNight-Sofia. To include realistic human–environment interactions missing from synthetic videos, we also recorded our own day–night paired videos. Capturing perfectly aligned real-world pairs is challenging, so we designed a practical video-guided recording strategy with post-trimming for better alignment.

We first record a daytime video with an ego-wearer exploring an environment while viewing the live camera feed on a phone screen. For the nighttime version, the same person, device, and viewpoint are kept unchanged. Instead of using live preview, the recorded daytime video is played back on the phone, serving as visual guidance to help the ego-wearer match walking speed, viewpoints, and actions. After brief practice, this approach proved more stable and reliable than methods like using landmarks or memorized trajectories. Post-trimming is applied to further refine spatial and temporal consistency.

Our real-world dataset, EgoNight-Sofia, contains 20 day–night paired videos recorded in Sofia, Bulgaria. Despite its modest size, it is a rare resource capturing diverse real-world everyday scenarios, including apartments, offices, grocery stores, streets, tourist spots, and outdoor fitness areas. The recordings include natural actions such as drinking water, locking doors, placing keys, charging devices, or checking price labels—leading to realistic VQA cases (e.g., "Where did I put my keys?", "How much was the drink?", "Did I turn left?"). Illumination sources include street lights, lamps, flashlights, and candles.

EgoNight-Oxford. Oxford Day–Night Wang et al. (2025b) is a notable exception that also includes egocentric videos captured under both daytime and nighttime conditions. Although it was originally designed for 3D vision tasks such as novel view synthesis, it offers illumination variations across five representative locations in Oxford. However, the day and night videos are not spatially or temporally aligned.

To enrich EgoNight with more realistic nighttime content—particularly for urban outdoor scenes—we manually select 20 nighttime segments to form EgoNight-Oxford, based on two criteria: (i) minimal overlap in trajectories and locations, and (ii) genuinely low-light conditions. These segments serve as a complementary testbed for evaluating model generalization under illumination changes when paired alignment is unavailable.

For both EgoNight-Sofia and EgoNight-Oxford, human annotators categorize each video into easy, medium, or hard levels. Together, these sources provide EgoNight with a balanced combination of precise alignment, natural dynamics, and broad real-world diversity.

3.2 EgoNight-VQA Benchmark Reconstruction

QA Task Taxonomy. To thoroughly assess models from multiple perspectives, we define a diverse taxonomy of 12 QA tasks. Some of these categories are well-studied and have been explored in previous egocentric VQA benchmarks, such as object/action/text recognition, counting, and spatial reasoning. Others are much less studied or newly proposed in EgoNight-VQA, including scene sequence and navigation (which require not only visual perception but also memory and spatial reasoning), illumination recognition and illumination change (designed to test models' understanding of lighting concepts), and non–common-sense reasoning (e.g., detecting abnormal cases such as a door inserted into a wall in the synthetic data). More detailed explanations of QA types can be found in the Appendix Sec. A.4.1. We further organize these categories into *paired* and *unpaired* QA types, depending on whether the same questions can be consistently applied across day–night counterparts: 1) **Paired QA Types.** These cover contexts that remain unchanged across day and night, allowing the same QA pairs to be used for both videos and thus providing a clean testbed for measuring performance gaps. Specifically, we include: ① *object recognition*, ② *text recognition*, ③ *spatial reasoning*, ④ *scene sequence*, ⑤ *navigation*, ⑥ *counting of static*, ⑦ *action recognition*, and ⑧ *non–common-sense reasoning*. 2) **Unpaired QA Types.** These include categories that are impractical to pair across day and night, or are only meaningful in the nighttime condition. We consider: ① *lighting recognition*, ② *lighting dynamic*, ③ *dynamic detection*, and ④ *counting of dynamic*. We control QA clip duration by task type. For static or spatial tasks (e.g., object recognition, lighting recognition), we use short clips of 3 seconds to minimize redundancy; For dynamic or temporal tasks (e.g., action recognition, navigation), the entire video is used to capture the complete context. Following recent works Plizzari et al. (2025); Xiao et al. (2025), we adopt the *open-ended QA* setting over the closed-form multiple-choice format, as it better reflects natural human–AI interactions.

A detailed summary of each QA type, including whether it is paired or unpaired, clip duration, and example questions, is provided in Fig. 3. This taxonomy makes EgoNight-VQA not only diverse and well-structured but also novel, introducing illumination reasoning and other challenges uniquely tied to nighttime egocentric vision.

Day-Augmented Auto QA Generation. Constructing large-scale QA pairs for nighttime videos is particularly challenging due to low visibility, which makes direct annotation both time-consuming and error-prone. To address this, as illustrated in the lower part of Fig. 2, we design a novel three-stage *day-augmented auto QA generation* pipeline that leverages aligned daytime videos as a strong prior for annotating their nighttime counterparts.

Specifically, the pipeline is tailored to each QA type and consists of three stages:

1) **Nighttime captioning.** For each clip, we prompt advanced MLLMs to generate detailed captions with an explicit focus on the target QA type (e.g., highlighting object-related attributes for object recognition or text/logos for text recognition). This ensures that the captions capture the most key information or construct relevant QA pairs.

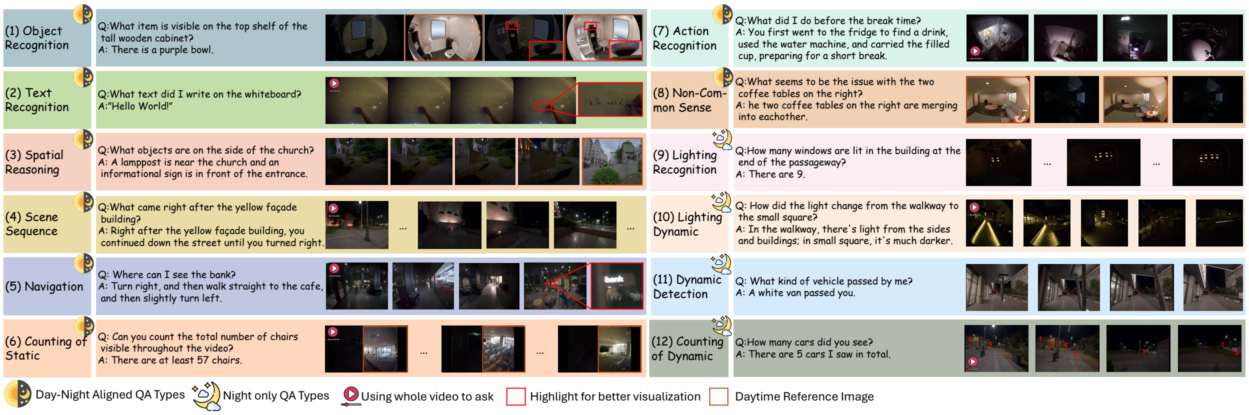

(1) Object Recognition
Q:What item is visible on the top shelf of the tall wooden cabinet?
A: There is a purple bowl.
(2) Text Recognition
Q:What text did I write on the whiteboard?
A:"Hello World!"
(3) Spatial Reasoning
Q:What objects are on the side of the church?
A: A lamppost is near the church and an informational sign is in front of the entrance.
(4) Scene Sequence
Q:What came right after the yellow façade building?
A: Right after the yellow façade building, you continued down the street until you turned right.
(5) Navigation
Q: Where can I see the bank?
A: Turn right, and then walk straight to the cafe, and then slightly turn left.
(6) Counting of Static
Q: Can you count the total number of chairs visible throughout the video?
A: There are at least 57 chairs.
(7) Action Recognition
Q:What did I do before the break time?
A: You first went to the fridge to find a drink, used the water machine, and carried the filled cup, preparing for a short break.
(8) Non-Common Sense
Q:What seems to be the issue with the two coffee tables on the right?
A: he two coffee tables on the right are merging into eachother.
(9) Lighting Recognition
Q:How many windows are lit in the building at the end of the passageway?
A: There are 9.
(10) Lighting Dynamic
Q: How did the light change from the walkway to the small square?
A: In the walkway, there's light from the sides and buildings; in small square, it's much darker.
(11) Dynamic Detection
Q: What kind of vehicle passed by me?
A: A white van passed you.
(12) Counting of Dynamic
Q:How many cars did you see?
A: There are 5 cars I saw in total.
Day-Night Aligned QA Types
Night only QA Types
Using whole video to ask
Highlight for better visualization
Daytime Reference Image

Figure 3: **QA types with examples.** The first eight are *paired* types, where the same question–answer applies to both day and night clips; the last four are *unpaired*, evaluated only at night. QA Types have various durations, with static or spatial tasks (e.g., 1 and 3) using short clips, while dynamic or temporal tasks (e.g., 4 and 5) use full videos.

2) **Nighttime question generation.** The caption, together with the corresponding night clip, is then fed into the same MLLM to produce diverse question candidates centered on the given QA type. This step encourages variety in phrasing and perspective while maintaining fidelity to the visual content.

3) **Day-augmented pseudo answer synthesis.** For paired QA types, pseudo answers are generated by consulting the aligned daytime clip, where content is more visible and less ambiguous. For unpaired QA types or datasets without alignment (e.g., EgoNight-Oxford), answers are instead derived directly from the nighttime clip.

All three stages are powered by GPT-4.1. Empirically, we find that both the QA-type-specific prompting and the inclusion of daytime videos substantially improve the quality and reliability of the generated QA pairs. More examples of VQA pairs and caption generation can be found in Appendix A.4.2 and A.7.1.

Human Annotator Refinement. Finally, human annotators refine QA pairs via three operations: i) **delete**, when QA pairs are meaningless, vague, duplicated, or inconsistent across day–night counterparts (for paired QA types); ii) **modify**, when the question is valid but the answer is wrong (or vice versa), or to resolve ambiguity; iii) **add**, when many pairs are removed or when important, challenging questions, especially about dynamic concepts, are missing.

After the first labeling round, we performed a random double-check to refine low-quality annotations. Thus, although our pipeline combines model generation with human refinement, *every QA pair (3,658 in total) is manually verified at least once*. In total, ∼200 hours of human effort were invested, ensuring the quality and reliability of EgoNight-VQA.

Dataset Statistics. EgoNight-VQA comprises 3,658 high-quality, fully human-verified QA pairs across 12 task types, sourced from EgoNight-Synthetic, EgoNight-Sofia, and EgoNight-Oxford, with an average of 40 pairs per video. Detailed statistics on QA distribution, video durations, task difficulties, scenarios, and illumination are shown in Fig. 4. The number of videos across the three subsets—Synthetic (50), Sofia (20), and Oxford (20)—is proportionally reflected in the VQA annotations (2029 : 813 : 816). This results in an approximately 1:1 balance between synthetic and real (Sofia + Oxford) VQA samples, ensuring that our benchmark is not dominated by synthetic content. We provide more comparison of Egocentric VQA datasets in Appendix A.3. Overall, EgoNight-VQA provides a diverse and comprehensive benchmark for evaluating egocentric vision models under nighttime conditions.

3.3 Benchmarks Beyond Egocentric VQA

Day-Night Correspondence Retrieval. To further assess model capabilities beyond VQA, we introduce *day–night correspondence retrieval*, which evaluates a model's ability to match visual content across illumination conditions. Specifically, we define two subtasks: **i) Spatial Retrieval (Scene Recognition).** Spatial retrieval, or scene recognition, is a long-standing vision task Arandjelovic et al. (2016); Miao et al. (2024). Here, it is extended: given a query clip and a set of N candidate clips of equal duration s, the model must retrieve the one depicting the same scene. This evaluates a model's ability to capture and relate spatial relations in egocentric videos, e.g., distinguishing a bedroom from a bathroom or another bedroom. We built this benchmark with 1000 randomly generated meta-tasks. Each task samples a query clip, and the candidate set includes its temporally aligned counterpart (with a temporal shift for added difficulty) plus $N-1$ negatives from other scenes. Performance is measured by Top-1 accuracy across all tasks. In our setup, we use $N = 10$, $s = 10$ seconds, and a temporal shift of [10, 20] frames. Both *Day (query)* → *Day (database)* and *Day* → *Night* settings are evaluated.

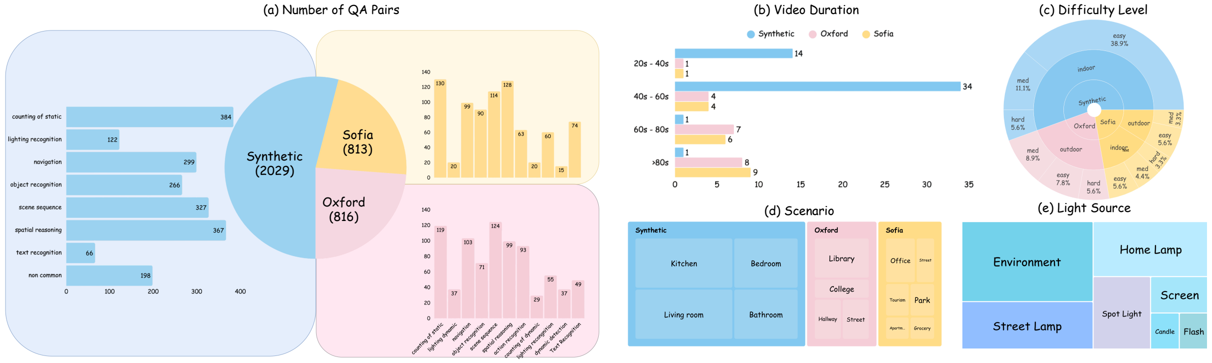


Figure 4: **Statistics of EgoNight-VQA benchmark.** (a) Distribution of QA pairs across QA types and sources. (b) Video duration distribution. (c) Task difficulty levels cross scenarios. (d) Scenario coverage. (e) Illumination coverage.

ii) Temporal Localization. We further design a temporal localization task to test whether models can align video segments across dynamics. Given a query clip of duration s, the model must localize it within the corresponding full video by predicting its start and end timestamps (t_i, t_j), directly evaluating temporal reasoning (e.g., grounding "The door is being locked" to 10–20s). We construct 1000 meta-tasks, each generated by randomly sampling one clip from its parent full video that is also randomly selected. Inspired by temporal grounding literature Xin et al. (2024), we adopt mean Intersection-over-Union (mIoU) between the predicted interval (t_i, t_j) and the ground-truth interval (t_i^*, t_j^*) as the evaluation metric. Consistent with spatial retrieval, we set $s = 10$ seconds and evaluate both *Day* $\rightarrow$ *Day* and *Night* $\rightarrow$ *Day* settings.

Egocentric Depth Estimation at Night. Depth estimation is a fundamental component of computer vision. On the one hand, extensive research Yang et al. (2024a;b); Wang et al. (2025a) has focused on depth estimation in non-egocentric settings (typically not with fisheye cameras), while egocentric depth estimation remains largely underexplored, especially under nighttime conditions. On the other hand, recent works Chen et al. (2024a); Liu et al. (2025) suggest that incorporating depth can enhance models' spatial reasoning abilities. These two observations motivate us to construct an auxiliary benchmark for *egocentric depth estimation at night*. Specifically, we use EgoNight-Synthetic as the testbed, where ground-truth depth maps are provided by the rendering engine. Thanks to the day–night aligned design, we can quantitatively evaluate models under both controlled daytime and nighttime conditions. For evaluation, we adopt standard depth estimation metrics, including absolute relative error (AbsRel), $\delta_1(1.25)$, $\delta_2(1.25^2)$, and $\delta_3(1.25^3)$, where δ_k measures the percentage of predicted pixels whose relative error is within a threshold of 1.25^k.

4 Experiments

4.1 Evaluated MLLMs & Metrics

We evaluate a broad set of state-of-the-art MLLMs on the proposed benchmarks. i) For **EgoNight-VQA**, we include two closed-source commercial models, GPT-4.1 Achiam et al. (2023) and Gemini 2.5 Pro Comanici et al. (2025); eight open-source models, Qwen2.5-VL (3B, 7B, 72B) Bai et al. (2023), VideoLLaMA3 (7B) Zhang et al. (2023b), InternVL3 (8B) Chen et al. (2024b), GLM-4.1V (9B-Base) Hong et al. (2025), and LLaVA-NeXT-Video (7B) Li et al. (2024); as well as EgoGPT Yang et al. (2025), one of the few open-source egocentric models tailored for open-ended QA. Following prior work Plizzari et al. (2025); Fan (2019), we adopt an *LLM-as-a-Judge* strategy to assess semantic consistency between predictions and ground truth, and report average accuracy across the test sets. ii)We provide further in-depth analysis on synthetic data quality, and potential model improvements. iii) For **day–night correspondence retrieval**, we benchmark feature-based retrieval methods, DINOv2 Oquab et al. (2023) and Perception Encoder (Percep. Enc.) Bolya et al. (2025), alongside MLLM-based methods, GPT-4.1 and InternVL3 (8B). As described in Sec. 3.3, Top-1 accuracy (Acc-R@1, %) and mIoU (%) are used for evaluating the spatial and temporal subtasks, respectively. iv) For **egocentric depth estimation**, we test a general monocular depth model (Depth Anything Yang et al. (2024a;b)), a 3D reconstruction-based method (VGGTStream Zhuo et al. (2025); Wang et al. (2025a)), and two egocentric fisheye-specific models (DAC Guo et al. (2025) and UniK3D Piccinelli et al. (2025)). For Depth Anything and VGGTStream, input fisheye RGB frames and depth maps are undistorted prior to inference for fair comparison. Additional implementation details (e.g., fps for frame extraction, prompts, and model settings) and discussion about LLM-as-a-Judge strategy are provided in the Appendix Sec. A.5.

Models	EgoNight-Synthetic			EgoNight-Sofia			EgoNight-Oxford			Avg.
	Easy	Medium	Hard	Easy	Medium	Hard	Easy	Medium	Hard	-
Closed-Source MLLMs										
GPT-4.1	29.30	**26.87**	**18.87**	32.04	**29.35**	**31.69**	**39.72**	**37.13**	**40.72**	**30.93**
Gemini 2.5 Pro	**31.05**	24.81	16.51	**38.24**	26.81	28.87	36.75	36.81	27.88	30.60
Open-source MLLMs										
InternVL3-8B	20.21	15.50	16.98	24.03	21.74	20.42	22.90	20.85	16.36	20.06
Qwen2.5-VL-72B	18.39	15.25	12.26	24.03	17.03	20.42	24.81	22.80	16.36	18.99
Qwen2.5-VL-7B	13.01	13.95	13.68	15.44	12.68	12.68	13.74	13.36	12.73	13.44
Qwen2.5-VL-3B	14.69	10.34	7.08	15.50	13.04	12.68	17.18	11.40	12.12	13.41
GLM-4.1V-9B-Base	19.09	13.70	15.57	18.60	18.48	16.20	17.15	22.15	18.79	18.20
VideoLLaMA3-7B	16.85	13.44	14.62	11.11	10.87	9.15	12.26	10.46	9.15	13.64
LLaVA-NeXT-Video-7B	6.36	11.37	1.89	13.95	9.78	14.79	3.05	2.61	3.03	7.28
Egocentric MLLMs										
EgoGPT	15.79	13.55	12.04	12.41	12.13	10.36	12.37	13.58	13.68	14.29

Table 1: **Comparison results on EgoNight-VQA.** Accuracies (%) of OpenQA results across three datasets and three difficulty levels. We compare closed-source models, open-source models, and egocentric-specific models.

4.2 RESULTS ON EGONIGHT-VQA

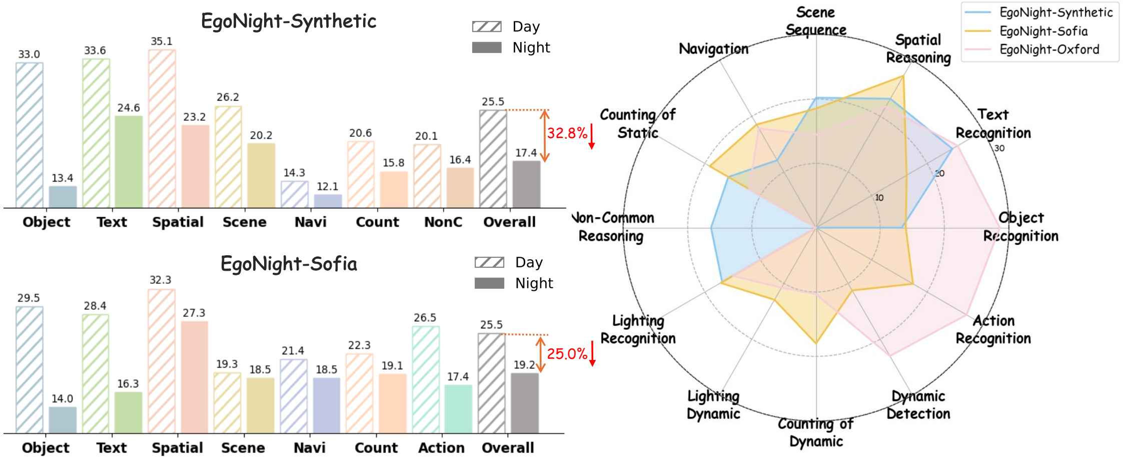
The main results of all MLLMs are shown in Tab. 1. In addition, we provide per-QA performance comparisons between day (striped bars) and night (solid bars) for paired QA types (Fig. 5(a)) and report nighttime performance across all QA types (Fig. 5(b)), based on averages across all models. Note that non–common case detection is available only in EgoNight-Synthetic, while dynamic events and actions are included only in the real-world data.

From the results in Tab. 1, we observe that almost all MLLMs struggle on our benchmark, with maximum averaged accuracies of 30.93% from the closed-source GPT-4.1, 20.06% from the open-source InternVL3-8B, and 14.29% from the egocentric EgoGPT. The wide performance spread also confirms that our dataset is sufficiently challenging and effective for distinguishing model capabilities. Fig. 5(a) further highlights the performance gap, showing declines of 32.8% and 25.0% on EgoNight-Synthetic and EgoNight-Sofia, respectively. Together, these results underscore the substantial challenges posed by our benchmark, exposing the limitations of current MLLMs under nighttime scenarios and highlighting the need for more illumination-robust models. Beyond the overall trends, we note three additional insights from Tab. 1: i) Closed-source models perform best. Within open-source models, Qwen2.5-VL generally improves with scale, yet InternVL outperforms the larger Qwen2.5-VL (72B), suggesting that size alone is insufficient. The relatively low results of EgoGPT further emphasize the need for more robust egocentric models. ii) EgoNight-Oxford achieves the highest scores, but its illumination conditions are more challenging than those in EgoNight-Synthetic and EgoNight-Sofia (Sec. A.4.2, Appendix). This indicates that without paired day videos and our day-augmented auto-labeling strategy, even human annotators face difficulties generating challenging QA pairs, underscoring the practical value of our dataset design; iii) Overall, performance declines across task levels (easy, medium, hard), validating the diversity and difficulty of our benchmark.

From the per-QA results in Fig. 5(a) and Fig. 5(b), we further observe three key trends: i) Models perform better on perception-oriented tasks (e.g., object recognition, text recognition, scene sequence) than reasoning-oriented tasks (e.g., navigation, counting, non-common-sense reasoning cases) under daytime conditions. However, at night, perception tasks suffer larger performance drops, indicating their higher sensitivity to illumination, whereas reasoning tasks, though harder overall, are relatively less affected since they rely more on temporal and contextual cues. ii) MLLMs achieve substantially lower accuracy on our newly proposed tasks, such as lighting recognition, lighting dynamics, scene sequence, dynamic detection, navigation, and non-common-sense reasoning, suggesting that existing MLLMs generalize poorly to novel tasks compared with well-studied ones like object recognition. iii) Each dataset in Fig. 5(b) emphasizes distinct aspects of nighttime challenges, together providing complementary perspectives that ensure EgoNight spans a balanced range of perception–reasoning difficulties under low-light conditions.

4.3 MORE IN-DEPTH ANALYSIS

What is the Quality of EgoNight Synthetic. More examples of synthetic visualization can be found in Appendix A.4.2 and A.1. To further show the quality of our synthetic dataset, we calculate the Pearson correlation of the average score per-model shown in Appendix A.6 between synthetic and Sofia (0.9359 with p-value 6.84710^{-5}), synthetic and Oxford (0.8588 with p-value 1.462×10^{-3}). These strong and statistically significant correlations in-



(a) Day-Night Performance Gap on Paired QA Types **(b) Performance Comparison on All QA Types**

Figure 5: **Performance analysis of MLLMs on EgoNight-VQA.** (a) Day–night performance gap across paired QA types, showing consistent degradation at night. (b) Nighttime performance across all 12 QA types. NonC means non–common-sense reasoning.

dicate that performance on synthetic data is highly predictive of performance on real-world data, further validating its representativeness. We also finetune Qwen2.5-VL-7B model with Supervised-Fine-Tuning(SFT) using only the synthetic data and evaluate it on the real dataset. This improves the model accuracy from 14.83% to 20.57%, which demonstrates that our synthetic data can effectively enhance model performance in real-world scenarios.

Could fine-tuning help to enhance performance? To explore potential solutions for improving low-light egocentric QA, we proactively conduct pilot studies using the EgoNight benchmark. We split EgoNight into 70% training and 30% testing subsets, and fine-tune Qwen2.5-VL-7B using supervised fine-tuning (SFT) under three configurations: i) Full Fine-Tuning. ii) Fine-tuning vision encoder only. iii) Fine-tuning LLM only. As shown in Tab. 2, our observations are as follows:

- Fine-tuning on EgoNight leads to substantial performance improvements, demonstrating that EgoNight-style nighttime data effectively helps models adapt to low-light egocentric scenarios.
- Both vision-encoder tuning and LLM tuning independently contribute to performance gains. Interestingly, fine-tuning the LLM only yields even larger improvements, suggesting that visual representation is not the only bottleneck. LLM fine-tuning plays a crucial role in aligning uncertain visual features to language space.
- Full fine-tuning consistently outperforms partial fine-tuning, indicating that EgoNight requires both strong visual perception and robust visual–language alignment.

How perception vs. reasoning-oriented tasks benefit from fine-tuning? To further dive into the impact of fine-tuning, we compare perception-oriented tasks (Object, Text Recognition) and reasoning-oriented tasks (Navigation, Counting) accuracy in Tab. 3. We can observe that:

- Perception-oriented tasks are significantly easier to enhance through fine-tuning compared to reasoning-oriented tasks, indicating that visual feature learning benefits more from data adaptation than higher-level reasoning.
- Fine-tuning the vision encoder improves perception-oriented tasks by 2–4 times but provides limited gains for reasoning-oriented tasks, revealing the substantial challenges posed by reasoning-centric scenarios.
- Fine-tuning the language model yields improvements for both perception and reasoning-oriented tasks, with larger boosts for perception tasks. This suggests that the main benefit arises from aligning uncertain visual features with language space, while enhancement of true reasoning ability remains limited.

We provide more failure case analysis in Appendix. A.9.

4.4 Results on Day-Night Correspondence Retrieval

Setting	Synthetic	Real
Qwen7B (Base)	23.23	16.40
Enc. FT	29.74	20.92
LLM. FT	35.50	22.26
Full FT	36.25	25.61

Table 2: Fine-tuning performance comparison across datasets. FT means Fine-tuning.

Task	Qwen7B	Enc. FT	LLM FT
Object	8.435	34.718	35.855
Text	18.440	49.890	50.988
Navigation	17.870	19.495	19.918
Counting	16.558	16.945	24.275

Table 3: Fine-tuning performance comparison across tasks.

Models	Spatial Retrieval (Acc - R@1 % ↑)				Temporal Localization (mIoU % ↑)			
	EgoNight-Synthetic		EgoNight-Sofia		EgoNight-Synthetic		EgoNight-Sofia	
	Day→Day	Night→Day	Day→Day	Night→Day	Day→Day	Night→Day	Day→Day	Night→Day
DINOv2	45.7	28.7	84.5	74.5	-	33.7	-	33.1
Percep. Enc.	65.4	41.6	89.8	80.9	-	32.9	-	33.4
GPT-4.1	75.6	54.1	92.5	84.5	14.7	10.0	21.2	15.5
InternVL3-8B	39.4	27.7	73.9	56.3	10.2	9.9	12.5	13.3

Table 4: **Night-to-Day retrieval performance.** Each dataset is evaluated on both Day→Day and Night→Day settings.

The results of day–night retrieval are reported in Tab. 4. The gap between Night–Day and Day–Day shows that cross-illumination retrieval remains highly challenging compared with in-domain retrieval. For spatial retrieval, GPT-4.1 consistently outperforms other methods, achieving over 80% accuracy. This suggests that Retrieval-Augmented Generation methods could further improve performance, as Fig. 5(a) already shows that daytime inputs significantly benefit the models. For temporal retrieval, however, GPT-4.1, despite its strong results on egocentric VQA (Tab. 1) and spatial retrieval, shows a substantial drop compared with feature-based methods (DINOv2 and Perception Encoder). A similar degradation is observed for InternVL3-8B. These findings suggest that while MLLMs excel at spatial semantic understanding, they struggle with temporal reasoning, such as timestamp prediction, which is critical for temporal localization. Further results on temporal limitations are provided in Appendix A.6.

4.5 Results on Depth Estimation

Results for depth estimation are reported in Tab. 5. The relatively low scores across all models highlight the difficulty of our EgoNight dataset, which combines egocentric motion, complex geometry, and extreme lighting variations. A clear gap between daytime and nighttime performance again underscores the challenges of low-light conditions. Among the methods, fisheye-based methods (DAC and UniK3D) outperform general depth estimators, suggesting the need for egocentric-specific algorithms. Additional qualitative results are provided in Sec. A.7.3.

Method	Abs Rel ↓		δ_1 (1.25) ↑		δ_2 (1.25^2) ↑		δ_3 (1.25^3) ↑	
	Day	Night	Day	Night	Day	Night	Day	Night
Depth Anything (U)	0.297	0.302	0.249	0.237	0.463	0.447	0.622	0.60
VGGTStream (U)	0.293	0.298	0.234	0.232	0.447	0.442	0.615	0.609
DAC (F)	0.245	0.292	0.255	0.216	0.495	0.425	0.684	0.602
UniK3D (F)	0.224	0.253	0.280	0.254	0.524	0.481	0.706	0.658

Table 5: Depth estimation results on EgoNight-Synthetic. **U**: undistorted input; **F**: fisheye input.

5 Conclusion

In this work, we introduced EgoNight, the first benchmark suite designed to systematically evaluate egocentric multimodal large language models (MLLMs) under challenging nighttime conditions. EgoNight integrates synthetic and real-world videos with day–night alignment, enabling rigorous analysis of illumination effects. Building upon this data, we proposed EgoNight-VQA, spanning 12 QA types with 3,658 human-verified pairs, alongside two complementary benchmarks: day–night correspondence retrieval and egocentric depth estimation. Experiments reveal that even state-of-the-art MLLMs struggle under low-light conditions, with performance dropping substantially compared to daytime. This highlights that nighttime egocentric vision remains far from being solved, motivating future research into illumination-robust egocentric perception and reasoning. More discussion about contribution and future works is provided in Appendix. A.9.1 and A.9.2. We believe EgoNight provides a valuable and timely benchmark that will drive progress toward more reliable egocentric AI assistants.

REFERENCES

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Relja Arandjelovic, Petr Gronat, Akihiko Torii, Tomas Pajdla, and Josef Sivic. NetVLAD: CNN architecture for weakly supervised place recognition. 2016.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023.

Wei Bian, Dacheng Tao, and Yong Rui. Cross-domain human action recognition. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 2011.

Daniel Bolya, Po-Yao Huang, Peize Sun, Jang Hyun Cho, Andrea Madotto, Chen Wei, Tengyu Ma, Jiale Zhi, Jathushan Rajasegaran, Hanoona Rasheed, Junke Wang, Marco Monteiro, Hu Xu, Shiyu Dong, Nikhila Ravi, Daniel Li, Piotr Dollár, and Christoph Feichtenhofer. Perception encoder: The best visual embeddings are not at the output of the network, 2025. URL `https://arxiv.org/abs/2504.13181`.

Keshigeyan Chandrasegaran, Agrim Gupta, Lea M Hadzic, Taran Kota, Jimming He, Cristóbal Eyzaguirre, Zane Durante, Manling Li, Jiajun Wu, and Li Fei-Fei. Hourvideo: 1-hour video-language understanding. *NeurIPS*, 2024.

Boyuan Chen, Zhuo Xu, Sean Kirmani, Brain Ichter, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In *CVPR*, 2024a.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *CVPR*, 2024b.

Sijie Cheng, Zhicheng Guo, Jingwen Wu, Kechen Fang, Peng Li, Huaping Liu, and Yang Liu. Egothink: Evaluating first-person perspective thinking capability of vision-language models. In *CVPR*, 2024.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. The epic-kitchens dataset: Collection, challenges and baselines. *TPAMI*, 2020.

Chenyou Fan. Egovqa-an egocentric video question answering benchmark dataset. In *ICCV Workshop*, 2019.

Yuqian Fu, Yanwei Fu, and Yu-Gang Jiang. Meta-fdmixup: Cross-domain few-shot learning guided by labeled target data. In *ACM Multimedia*, 2021.

Yuqian Fu, Yu Wang, Yixuan Pan, Lian Huai, Xingyu Qiu, Zeyu Shangguan, Tong Liu, Yanwei Fu, Luc Van Gool, and Xingqun Jiang. Cross-domain few-shot object detection via enhanced open-set object detector. In *ECCV*, 2024.

Yuqian Fu, Runze Wang, Yanwei Fu, Danda Pani Paudel, Xuanjing Huang, and Luc Van Gool. Objectrelator: Enabling cross-view object relation understanding in ego-centric and exo-centric videos. *ICCV*, 2025.

Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *CVPR*, 2022.

Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, Bikram Boote, et al. Ego-exo4d: Understanding skilled human activity from first-and third-person perspectives. In *CVPR*, 2024.

Yuliang Guo, Sparsh Garg, S. Mahdi H. Miangoleh, Xinyu Huang, and Liu Ren. Depth any camera: Zero-shot metric depth estimation from any camera. In *CVPR*, 2025.

Yunhui Guo, Noel C Codella, Leonid Karlinsky, James V Codella, John R Smith, Kate Saenko, Tajana Rosing, and Rogerio Feris. A broader study of cross-domain few-shot learning. In *ECCV*, 2020.

Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang, Jiale Cheng, Ji Qi, Junhui Ji, Lihang Pan, et al. Glm-4.1 v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning. pp. arXiv–2507, 2025.

Yifei Huang, Guo Chen, Jilan Xu, Mingfang Zhang, Lijin Yang, Baoqi Pei, Hongjie Zhang, Lu Dong, Yali Wang, Limin Wang, et al. Egoexolearn: A dataset for bridging asynchronous ego-and exo-centric view of procedural activities in real world. In *CVPR*, 2024.

Bernardo Iraci. *Blender cycles: lighting and rendering cookbook*. Packt Publishing Ltd, 2013.

Baoxiong Jia, Ting Lei, Song-Chun Zhu, and Siyuan Huang. Egotaskqa: Understanding human tasks in egocentric videos. *NeurIPS*, 2022.

Simar Kareer, Dhruv Patel, Ryan Punamiya, Pranay Mathur, Shuo Cheng, Chen Wang, Judy Hoffman, and Danfei Xu. Egomimic: Scaling imitation learning via egocentric video. In *ICRA*, 2025.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024.

Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Deeper, broader and artier domain generalization. In *ICCV*, 2017.

Guofa Li, Zefeng Ji, Xingda Qu, Rui Zhou, and Dongpu Cao. Cross-domain object detection for autonomous driving: A stepwise domain adaptative yolo approach. *IEEE Transactions on Intelligent Vehicles*, 2022.

Jinlong Li, Runsheng Xu, Jin Ma, Qin Zou, Jiaqi Ma, and Hongkai Yu. Domain adaptive object detection for autonomous driving under foggy weather. In *WACV*, 2023a.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models, 2023b. URL `https://arxiv.org/abs/2301.12597`.

Kailing Li, Qi'ao Xu, Tianwen Qian, Yuqian Fu, Yang Jiao, and Xiaoling Wang. Clivis: Unleashing cognitive map through linguistic-visual synergy for embodied visual reasoning. *arXiv preprint arXiv:2506.17629*, 2025a.

Yanjun Li, Yuqian Fu, Tianwen Qian, Qi'ao Xu, Silong Dai, Danda Pani Paudel, Luc Van Gool, and Xiaoling Wang. Egocross: Benchmarking multimodal large language models for cross-domain egocentric video question answering. *arXiv preprint arXiv:2508.10729*, 2025b.

Yu Li, Xingyu Qiu, Yuqian Fu, Jie Chen, Tianwen Qian, Xu Zheng, Danda Pani Paudel, Yanwei Fu, Xuanjing Huang, Luc Van Gool, et al. Domain-rag: Retrieval-guided compositional image generation for cross-domain few-shot object detection. *arXiv preprint arXiv:2506.05872*, 2025c.

Zhuowan Li, Xingrui Wang, Elias Stengel-Eskin, Adam Kortylewski, Wufei Ma, Benjamin Van Durme, and Alan L Yuille. Super-clevr: A virtual benchmark to diagnose domain robustness in visual reasoning. In *CVPR*, 2023c.

Gaowen Liu, Hao Tang, Hugo M Latapie, Jason J Corso, and Yan Yan. Cross-view exocentric to egocentric video synthesis. In *ACM Multimedia*, 2021.

Yang Liu, Ming Ma, Xiaomin Yu, Pengxiang Ding, Han Zhao, Mingyang Sun, Siteng Huang, and Donglin Wang. Ssr: Enhancing depth perception in vision-language models via rationale-guided spatial reasoning. *arXiv preprint arXiv:2505.12448*, 2025.

Zhengyi Luo, Ryo Hachiuma, Ye Yuan, and Kris Kitani. Dynamics-regulated kinematic policy for egocentric pose estimation. *NeurIPS*, 2021.

Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. Egoschema: A diagnostic benchmark for very long-form video language understanding. *NeurIPS*, 2023.

Yang Miao, Francis Engelmann, Olga Vysotska, Federico Tombari, Marc Pollefeys, and Dániel Béla Baráth. Scenegraphloc: Cross-modal coarse visual localization on 3d scene graphs. 2024.

Maxime Oquab, Timothée Darcet, Theo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Russell Howes, Po-Yao Huang, Hu Xu, Vasu Sharma, Shang-Wen Li, Wojciech Galuba, Mike Rabbat, Mido Assran, Nicolas Ballas, Gabriel Synnaeve, Ishan Misra, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2023.

Boxiao Pan, Zhangjie Cao, Ehsan Adeli, and Juan Carlos Niebles. Adversarial cross-domain action recognition with co-attention. In *AAAI*, 2020.

Luigi Piccinelli, Christos Sakaridis, Mattia Segu, Yung-Hsu Yang, Siyuan Li, Wim Abbeloos, and Luc Van Gool. UniK3D: Universal camera monocular 3d estimation. In *CVPR*, 2025.

Chiara Plizzari, Alessio Tonioni, Yongqin Xian, Achin Kulshrestha, and Federico Tombari. Omnia de egotempo: Benchmarking temporal understanding of multi-modal llms in egocentric videos. In *CVPR*, 2025.

Shraman Pramanick, Yale Song, Sayan Nag, Kevin Qinghong Lin, Hardik Shah, Mike Zheng Shou, Rama Chellappa, and Pengchuan Zhang. Egovlpv2: Egocentric video-language pre-training with fusion in the backbone. In *ICCV*, 2023.

Alexander Raistrick, Lahav Lipson, Zeyu Ma, Lingjie Mei, Mingzhe Wang, Yiming Zuo, Karhan Kayan, Hongyu Wen, Beining Han, Yihan Wang, Alejandro Newell, Hei Law, Ankit Goyal, Kaiyu Yang, and Jia Deng. Infinite photorealistic worlds using procedural generation. In *CVPR*, 2023.

Xiaofeng Ren and Chunhui Gu. Figure-ground segmentation improves handled object recognition in egocentric video. In *CVPR*, 2010.

Swathikiran Sudhakaran, Sergio Escalera, and Oswald Lanz. Lsta: Long short-term attention for egocentric action recognition. In *CVPR*, 2019.

Suraj Jyothi Unni, Raha Moraffah, and Huan Liu. Vqa-gen: A visual question answering benchmark for domain generalization. *arXiv preprint arXiv:2311.00807*, 2023.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *CVPR*, 2025a.

Zirui Wang, Wenjing Bian, Xinghui Li, Yifu Tao, Jianeng Wang, Maurice Fallon, and Victor Adrian Prisacariu. Seeing in the dark: Benchmarking egocentric 3d vision with the oxford day-and-night dataset. *arXiv preprint arXiv:2506.04224*, 2025b.

Junbin Xiao, Nanxin Huang, Hao Qiu, Zhulin Tao, Xun Yang, Richang Hong, Meng Wang, and Angela Yao. Egoblind: Towards egocentric visual assistance for the blind people. *arXiv preprint arXiv:2503.08221*, 2025.

Gu Xin, Fan Heng, Huang Yan, Luo Tiejian, and Zhang Libo. Context-guided spatio-temporal video grounding. 2024.

Jingkang Yang, Shuai Liu, Hongming Guo, Yuhao Dong, Xiamengwei Zhang, Sicheng Zhang, Pengyun Wang, Zitang Zhou, Binzhu Xie, Ziyue Wang, et al. Egolife: Towards egocentric life assistant. In *CVPR*, 2025.

Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *CVPR*, 2024a.

Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *arXiv:2406.09414*, 2024b.

Hanrong Ye, Haotian Zhang, Erik Daxberger, Lin Chen, Zongyu Lin, Yanghao Li, Bowen Zhang, Haoxuan You, Dan Xu, Zhe Gan, et al. Mm-ego: Towards building egocentric multimodal llms for video qa. *arXiv preprint arXiv:2410.07177*, 2024.

Ganlin Zhang, Deheng Zhang, Longteng Duan, and Guo Han. Accessible robot control in mixed reality, 2023a. URL `https://arxiv.org/abs/2306.02393`.

Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*, 2023b.

Haoyu Zhang, Qiaohui Chu, Meng Liu, Yunxiao Wang, Bin Wen, Fan Yang, Tingting Gao, Di Zhang, Yaowei Wang, and Liqiang Nie. Exo2ego: Exocentric knowledge guided mllm for egocentric video understanding. *arXiv preprint arXiv:2503.09143*, 2025a.

Yao Zhang, Haokun Chen, Ahmed Frikha, Denis Krompass, Gengyuan Zhang, Jindong Gu, and Volker Tresp. Cl-cross vqa: A continual learning benchmark for cross-domain visual question answering. In *WACV*, 2025b.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, Zhangchi Feng, and Yongqiang Ma. Llamafactory: Unified efficient fine-tuning of 100+ language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, Bangkok, Thailand, 2024. Association for Computational Linguistics. URL `http://arxiv.org/abs/2403.13372`.

Kaiyang Zhou, Yongxin Yang, Yu Qiao, and Tao Xiang. Domain generalization with mixstyle. *arXiv preprint arXiv:2104.02008*, 2021.

Kaiyang Zhou, Ziwei Liu, Yu Qiao, Tao Xiang, and Chen Change Loy. Domain generalization: A survey. 2022.

Dong Zhuo, Wenzhao Zheng, Jiahe Guo, Yuqi Wu, Jie Zhou, and Jiwen Lu. Streaming 4d visual geometry transformer. *arXiv preprint arXiv:2507.11539*, 2025.