ICLR 2026

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

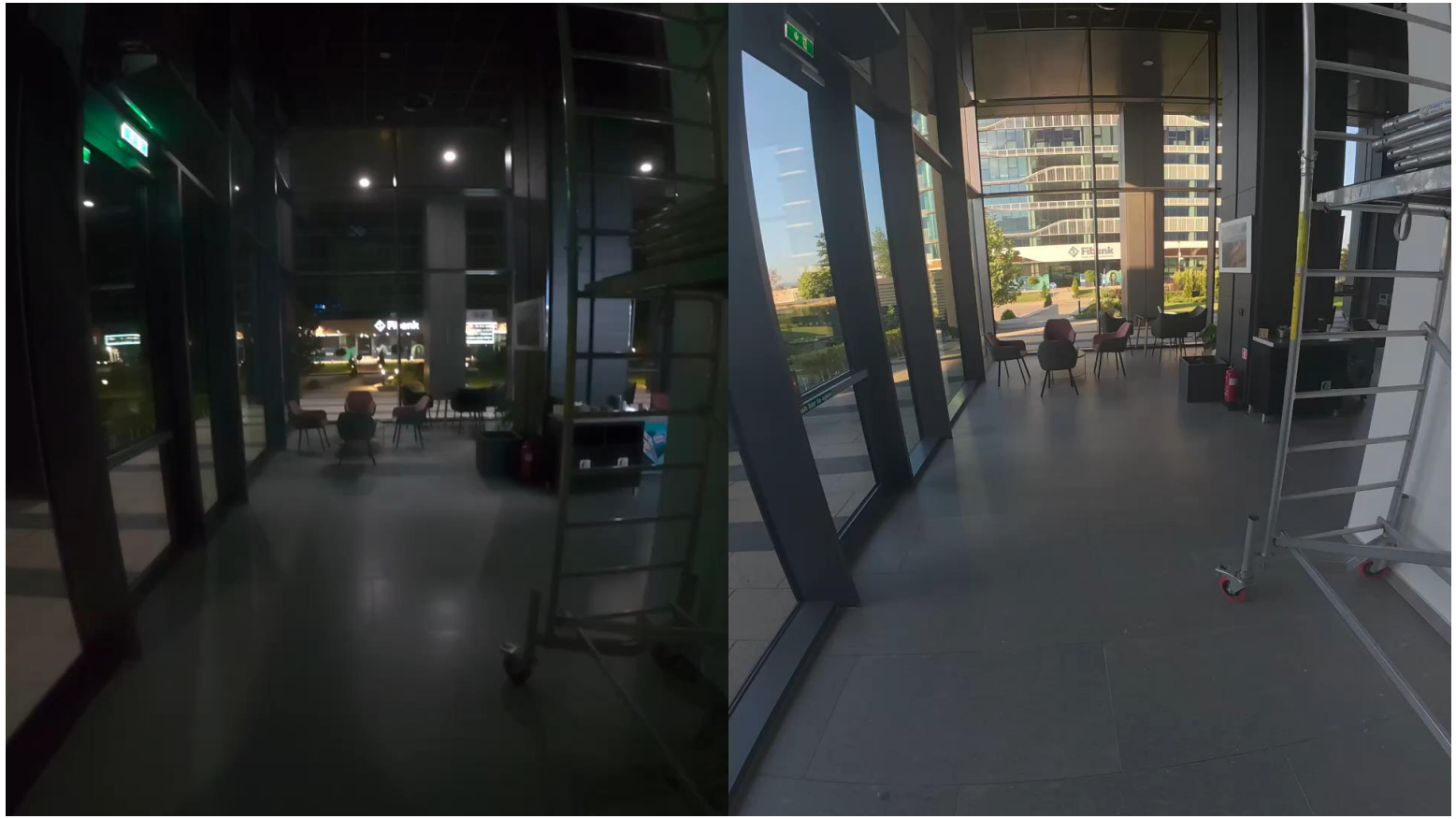
Deheng Zhang[1*], Yuqian Fu[1*], Runyi Yang[1], Yang Miao[1], Tianwen Qian[2], Xu Zheng[1,3], Guolei Sun[4], Ajad Chhatkuli[1], Xuanjing Huang[5], Yu-Gang Jiang[5], Luc Van Gool[1], Danda Pani Paudel[1]

1 INSAIT, Sofia University **2** East China Normal University **3** HKUST(GZ) **4** Nankai University **5** Fudan University

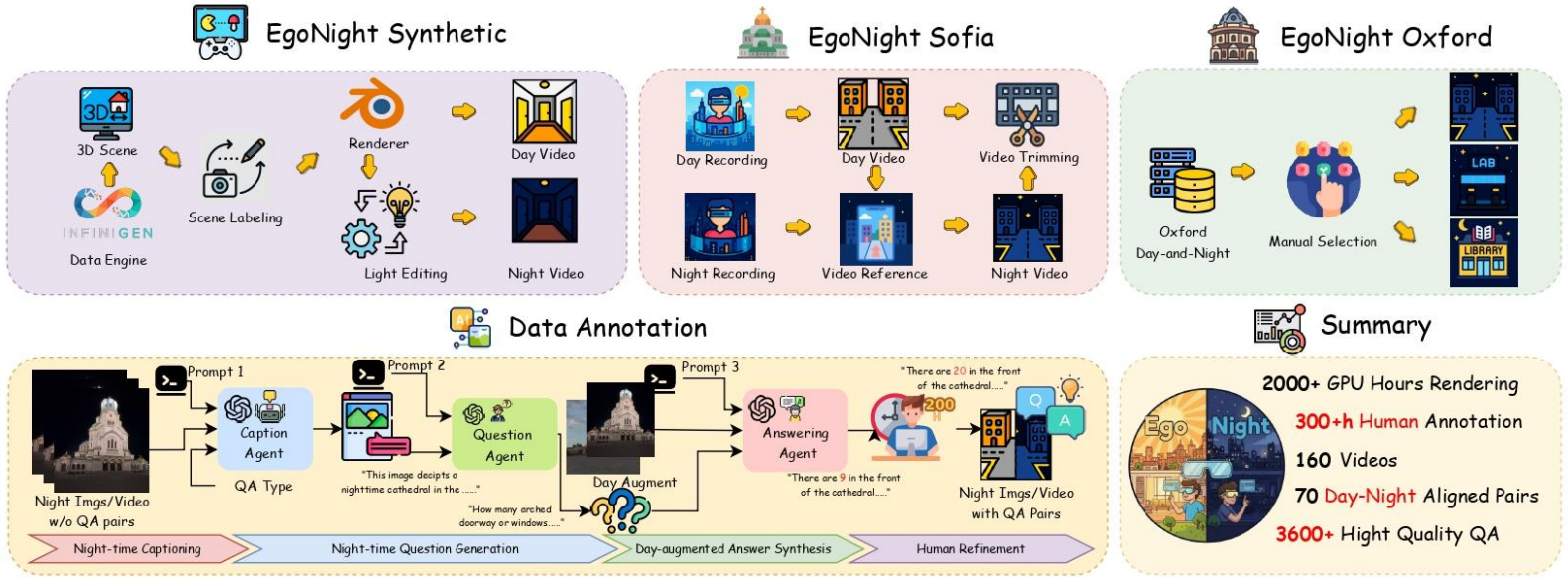
Overview


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.

Annotation Pipeline



Annotation Pipeline



1 Nighttime Captioning: GPT-4.1 generates scene descriptions from nighttime clips.

2 Day-Augmented Answer Synthesis: Daytime reference used to improve answer accuracy.

3 Question Generation: Diverse candidate QA pairs created per task type.

4 Human Refinement: All 3,658 pairs manually verified and refined (300+ hours).

Paired QA Types (8)

Same questions applied to both day and night videos:

Object Recognition · Text Recognition · Spatial Reasoning · Scene Sequence · Navigation · Counting of Static · Action Recognition · Non-Common-Sense Reasoning

Unpaired QA Types (4)

Nighttime-specific or impractical to pair:

Lighting Recognition · Lighting Dynamic · Dynamic Detection · Counting of Dynamic

Dataset Info

01

EgoNight Dataset

160 videos, 70 day-night aligned pairs, combining synthetic and real-world sources.

02

EgoNight-VQA

3,658 human-verified QA pairs across 12 diverse question types.

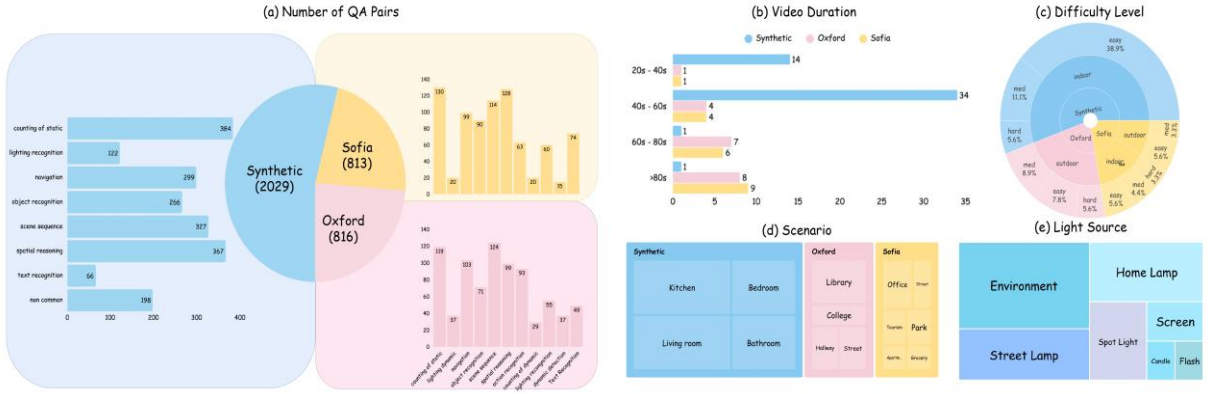
03

Empirical Evaluation

Comprehensive benchmarking of 10+ MLLMs revealing significant day-night performance gaps.

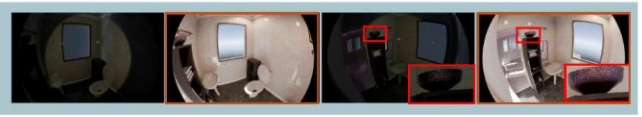
Dataset at a Glance

Subset	Videos	Day-Night Aligned	Source
EgoNight-Synthetic	100	50 pairs	Blender + Infinigen
EgoNight-Sofia	40	20 pairs	Real-world, Sofia
EgoNight-Oxford	20	Night-only	Oxford Day-and-Night



Dataset Info

(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.

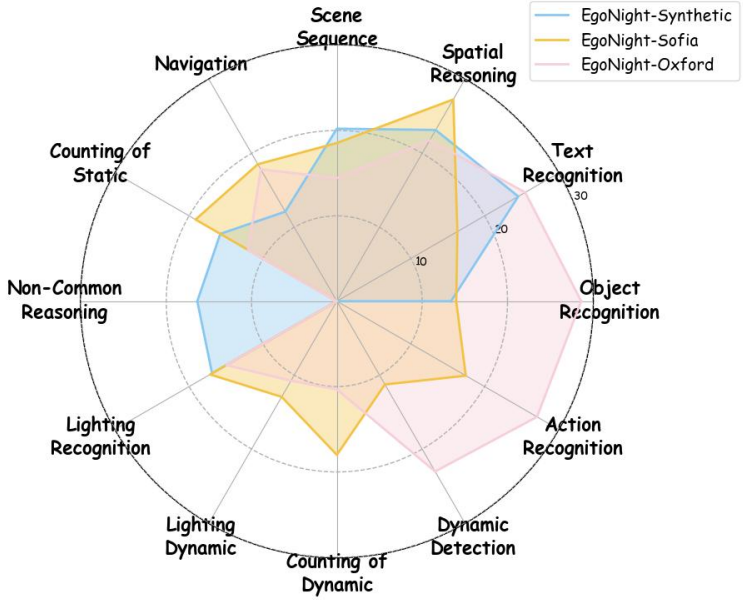
Day-Night Aligned QA Types | Night only QA Types | Using whole video to ask | Highlight for better visualization | Daytime Reference Image

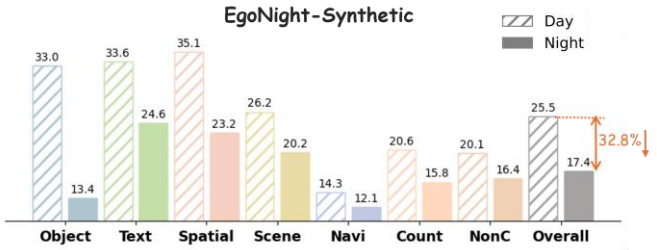
VQA Benchmark

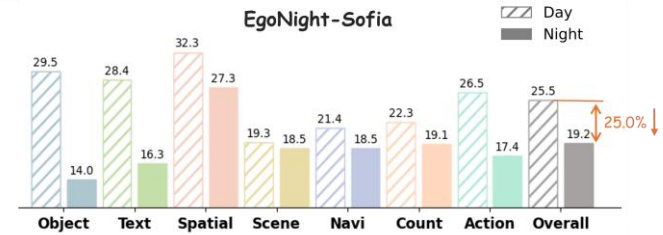
Leaderboard on EgoNight-VQA

Accuracies (%) of OpenQA results across three datasets and three difficulty levels.

Models	EGONIGHT-SYNTHETIC			EGONIGHT-SOFIA			EGONIGHT-OXFORD			AVG.
	EASY	MEDIUM	HARD	EASY	MEDIUM	HARD	EASY	MEDIUM	HARD	
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
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
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
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
LLaVA-NeXT-Video-7B	6.36	11.37	1.89	13.95	9.78	14.79	3.05	2.61	3.03	7.28
EgoGPT	15.79	13.55	12.04	12.41	12.13	10.36	12.37	13.58	13.68	14.29

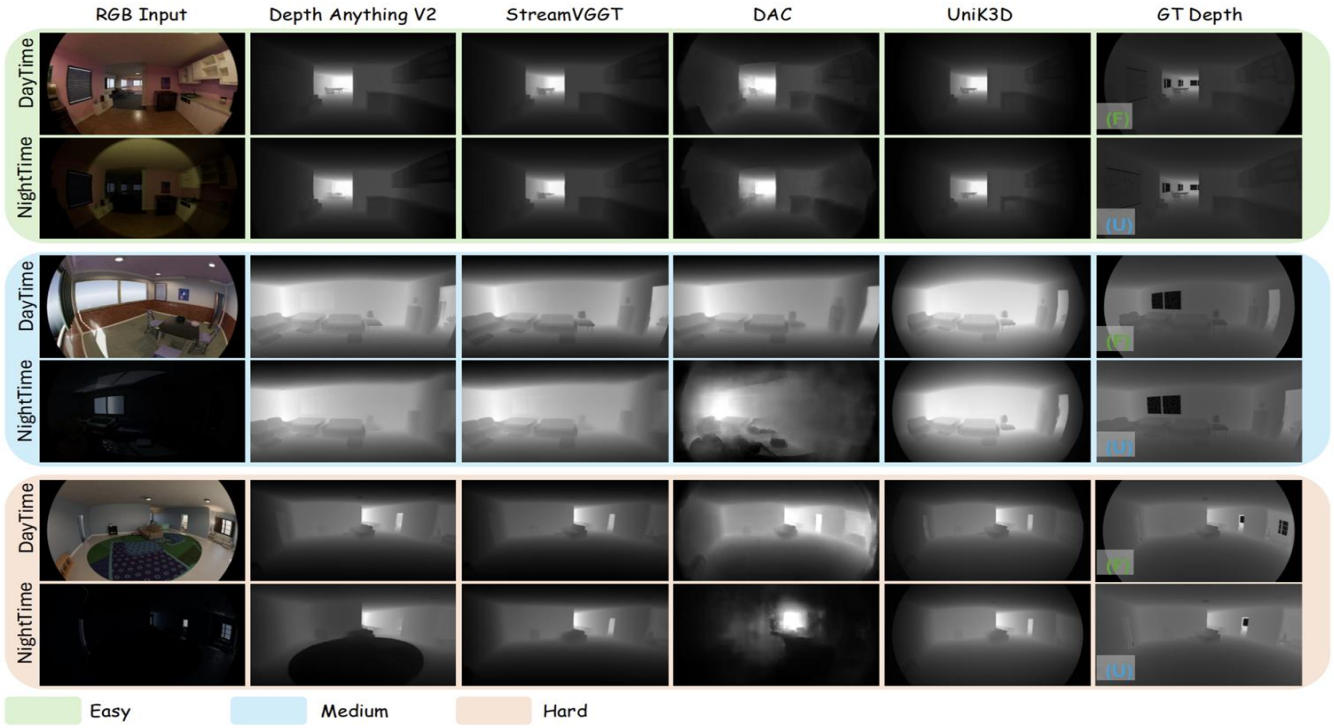






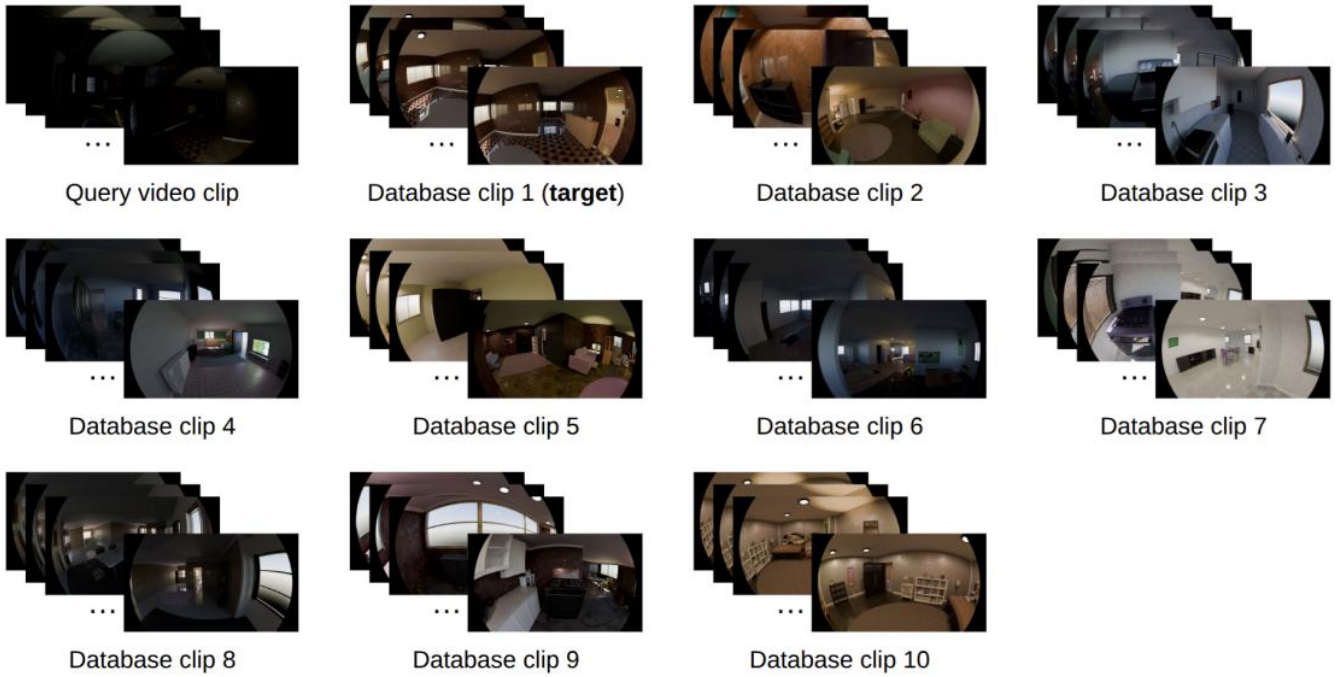
(a) Day-Night Performance Gap on Paired QA Types

Depth Estimation



Method	AbsRel D/N ↓	delta1 D/N ↑
Depth Anything	0.297 / 0.302	0.249 / 0.237
VGGTStream	0.293 / 0.298	0.234 / 0.232
DAC	0.245 / 0.292	0.255 / 0.216
UniK3D	0.224 / 0.253	0.280 / 0.254

Day-Night Correspondence Retrieval



Model	Spatial N->D (Syn/Sofia) Acc ↑	Temporal N->D (Syn/Sofia) MIoU ↑
GPT-4.1	54.1 / 84.5	10.0 / 15.5
Percep. Enc.	41.6 / 80.9	32.9 / 33.4
DINOv2	28.7 / 74.5	33.7 / 33.1
InternVL3-8B	27.7 / 56.3	9.9 / 13.3

Fine-tuning

Fine-Tuning Experiment

Model: Qwen2.5-VL-7B on EgoNight-Sofia+Oxford (7:3 train validation)

Strategy	Accuracy (%)	Improvement
Zero-shot baseline	16.40	—
Vision encoder only	20.92	Helps perception
LLM only	22.26	Helps feature alignment
Full fine-tuning	**25.61**	**+9.21%**

Model: Qwen2.5-VL-7B on EgoNight-Sofia+Oxford

Strategy	Accuracy (%)	Improvement
Zero-shot baseline	14.83	—
Synthetic finetuning	20.57	**+5.74%**

Key Insights

01 Full Fine-Tuning is Best

Achieves the best overall performance (+9.21% absolute improvement), though overall accuracy remains low.

02 Vision Encoder Tuning

Primarily improves perception-heavy tasks like object and text recognition under low light.

03 LLM Tuning

Improves both perception and reasoning, suggesting language priors matter significantly for nighttime understanding.

04 Synthetic-to-Real Transfer

Synthetic-only training transfers well to real nighttime, validating the utility of synthetic data.

Take-away Message

EgoNight is the **first comprehensive benchmark** for nighttime egocentric vision, addressing a critical gap between laboratory evaluations and real-world deployment conditions.

Three Core Takeaways

1. **The illumination gap is real and large:** State-of-the-art MLLMs degrade by 25–33% when moving from day to night, revealing a fundamental limitation.
2. **Aligned synthetic + real data is a viable strategy:** The strong correlation ($r = 0.9359$) validates using synthetic data for nighttime adaptation and evaluation.
3. **Fine-tuning helps, but the problem is far from solved:** Even after fine-tuning, models achieve only ~20% accuracy, leaving substantial room for improvement.

Future Directions

- Illumination-aware pretraining for egocentric models
- Nighttime-specific data augmentation strategies
- Cross-modal (3D + audio + vision) approaches for low-light understanding
- Larger-scale real-world nighttime egocentric datasets
- Further adversarial cases
  - Spatial invariance
  - Temporal invariance

Acknowledgement

Funders & Collaborators

vivo

SOFIA UNIVERSITY
ST. KLIMENT OHRIDSKI

Google Cloud Platform

Open-source Tools

LLaMA-Factory
Easy and Efficient LLM Fine-Tuning

blender

INFINIGEN

Contact

Thanks!

deheng.zhang@insait.ai